



UNIVERSITY *of* LIMERICK

OLLSCOIL LUIMNIGH

**Analysis of Lignocellulosic Feedstocks
for Biorefineries with a Focus on
The Development of Near Infrared
Spectroscopy as a Primary Analytical Tool**

Volume 1 of 2 (Main Text)

Thesis Presented for the award of Doctor of Philosophy (Ph.D.)

By

Daniel J. Hayes

University of Limerick

Supervisor: Dr J. J. Leahy

Submitted to the University of Limerick, July 2011

Declaration

The work presented in this Thesis is the original work of the Author, under the supervision of Dr J. J. Leahy, and due reference has been made where necessary to the work of others. No part of this Thesis has been previously submitted to this or any other University.

Daniel Hayes

Date

Acknowledgements

I would like to thank my supervisor, Dr J. J. Leahy, and my father, Prof. Michael Hayes, for their guidance and advice throughout the project.

I would also like to acknowledge the laboratory work provided by lab assistants and people on work-placements, and by the lab assistant Mark Ashworth in particular.

Outputs of the Research

Peer-Reviewed Publications

HAYES, D. J. 2008. An Examination of Biorefining Processes, Catalysts and Challenges. *Catalysis Today*, 145, 138-151.

HAYES, D. J. & HAYES, M. H. B. 2009. The role that lignocellulosic feedstocks and various biorefining technologies can play in meeting Ireland's biofuel targets. *Biofpr*, 3, 500-520.

Book Chapters

HAYES, D. J., FITZPATRICK, S. W., HAYES, M. H. B. & ROSS, J. R. H. 2005. The Biofine Process: Production of levulinic acid, furfural and formic acid from lignocellulosic feedstocks. *In: KAMM, B., GRUBER, P. R. & KAMM, M. (eds.) Biorefineries: Industrial Processes and Products*. Weinheim, Germany: Wiley.

HAYES, D. J., HAYES, M. H. B. & DALY, M. M. 2006. Operação inovadora de biorrefino para produção de óleos combustíveis e de químico-plataforma a partir de carboidratos de biomassa e de resíduos diversos. *In: DE SOUZA, F. H. D., POTT, E. B., PRIMAVESI, O., BERNARDI, A. C. C. & RODRIGUES, A. A. (eds.) Usos alternativos da palhada residual da produção de sementes para pastagens*. São Carlos, SP, Brazil: EMBRAPA.

Conference Proceedings

HAYES, D. J. 2004b. Oil substitutes utilising humic precursors: The development of a carbohydrate economy. *In: MARTIN-NETO, L. (ed.) Humic Substances and Soil and Water Environment*. Embrapa, São Carlos, Brazil. Presented at XII International Meeting of International Humic Substances Society, São Pedro, SP, Brazil, July 26-30, 2004

Other Articles

HAYES, D. J. 2004. A biomass solution to Ireland's hydrocarbon dependency. *Chemistry in Action*, Jan 2004.

The website for the Carbolea research group, www.carbolea.ul.ie, is managed by Daniel Hayes.

Successful Project Proposals Written by the Author

"Analysis of the suitability of second-generation technologies (2GTs) for Irish agriculture and the viability and costs of feedstocks: Desk-based evaluation supported by limited chemical analysis"

€120,638 project funded by the Department of Agriculture and Food Research Stimulus Fund Programme. Commenced December 2007, estimated completion July 2011.

"The laboratory analysis of Irish municipal and agricultural biomass wastes and evaluation of their utilisation in biorefining technologies"

€107,978 project funded by the EPA STRIVE Programme. Commenced December 2008, estimated completion August 2011.

"The chemical analysis and evaluation of Bord na Móna peats and AES organic wastes"

€48,071 project funded by Bord na Móna. November 2008 – December 2009.

"The Production of Sustainable Diesel-Miscible-Biofuels from the Residues and Wastes of Europe and Latin America"

€3,734,576 project (with €1,435,072 to UL) funded by the EU 7th Framework Programme. Commenced July 2009. Projected completion January 2013.

In all of the above projects the Author has had a major management role.

Abstract

“Analysis of Lignocellulosic Feedstocks for Biorefineries with a Focus on The Development of Near Infrared Spectroscopy as a Primary Analytical Tool”

Daniel J. Hayes

The processing of lignocellulosic materials in modern biorefineries will allow for the production of transport fuels and platform chemicals that could replace petroleum-derived products. However, there is a critical lack of relevant detailed compositional information regarding feedstocks relevant to Ireland and Irish conditions. This research has involved the collection, preparation, and the analysis, with a high level of precision and accuracy, of a large number of biomass samples from the waste and agricultural sectors. Not all of the waste materials analysed are considered suitable for biorefining; for example the total sugar contents of spent mushroom composts are too low. However, the waste paper/cardboard that is currently exported from Ireland has a chemical composition that could result in high biorefinery yields and so could make a significant contribution to Ireland’s biofuel demands.

Miscanthus was focussed on as a major agricultural feedstock. A large number of plants have been sampled over the course of the harvest window (October to April) from several sites. These have been separated into their anatomical fractions and analysed. This has allowed observations to be made regarding the compositional trends observed within plants, between plants, and between harvest dates. Projections are made regarding the extents to which potential chemical yields may vary. For the DIBANET hydrolysis process that is being developed at the University of Limerick, per hectare yields of levulinic acid from Miscanthus could be 20% greater when harvested early compared with a late harvest.

The wet-chemical analysis of biomass is time-consuming. Near infrared spectroscopy (NIRS) has been developed as a rapid primary analytical tool with separate quantitative models developed for the important constituents of Miscanthus, peat, and (Australian) sugarcane bagasse. The work has demonstrated that accurate models are possible, not only for dry homogenous samples, but also for wet heterogeneous samples. For glucose (cellulose) the root mean square error of prediction (RMSEP) for wet samples is 1.24% and the R^2 for the validation set (R_{val}^2) is 0.931. High accuracies are even possible for minor analytes; e.g. for the rhamnose content of wet Miscanthus samples the RMSEP is 0.03% and the R_{val}^2 is 0.845. Accurate models have also been developed for pre-treated Miscanthus samples and are discussed. In addition, qualitative models have been developed. These allow for samples to be discriminated for on the basis of plant fraction, plant variety (*giganteus*/*non-giganteus*), harvest-period (early/late), and stand-age (one-year/older).

Quantitative NIRS models have also been developed for peat, although the heterogeneity of this feedstock means that the accuracies tend to be lower than for Miscanthus. The development of models for sugarcane bagasse has been hindered, in some cases, by the limited chemical variability between the samples in the calibration set. Good models are possible for the glucose and total sugars content, but the accuracy of other models is poorer. NIRS spectra of Brazilian bagasse samples have been projected onto these models, and onto those developed for Miscanthus, and the Miscanthus models appear to provide a better fit than the Australian bagasse models.

Table of Contents

LIST OF FIGURES	V
LIST OF TABLES	IX
LIST OF ABBREVIATIONS	XI
1 INTRODUCTION	1
1.1 BACKGROUND	1
1.2 BASIS OF THE PRESENT RESEARCH	5
1.3 STRUCTURE OF THESIS	8
2 CHEMISTRY OF LIGNOCELLULOSICS	11
2.1 CARBOHYDRATES	11
2.2 LIGNIN	25
2.3 ASSOCIATIONS AND DEVELOPMENT OF LIGNOCELLULOSIC COMPONENTS	27
2.4 EXTRACTIVES	30
2.5 PROTEIN	31
2.6 ASH	31
2.7 MOISTURE CONTENT	32
2.8 HEATING VALUE	33
3 REFERENCE ANALYTICAL METHODS	35
3.1 METHODS FOR DETERMINING STRUCTURAL CARBOHYDRATES AND LIGNIN	35
3.2 ULTRAVIOLET-VISIBLE SPECTROSCOPY	43
3.3 MOISTURE CONTENT	50
3.4 EXTRACTIVES CONTENT	51
3.5 ASH CONTENT	59
3.6 ELEMENTAL ANALYSES	60
3.7 THERMO GRAVIMETRIC ANALYSIS	61
3.8 NOTES ON PARTICLE SIZE	64
3.9 SUMMARY	65
4 ION CHROMATOGRAPHY (IC)	68
4.1 BACKGROUND TO IC	68
4.2 IC FOR CARBOHYDRATES	70
4.3 DETECTION	71
4.4 THE DAVIS (1998) PAPER	73
4.5 IMPORTANT STATISTICS FOR IC	75
4.6 SET-UP OF AN ION CHROMATOGRAPHY SYSTEM AT UL	80
4.7 SUMMARY	103

5	<u>THEORY BEHIND NEAR INFRARED SPECTROSCOPY</u>	105
5.1	ELECTROMAGNETIC SPECTRUM AND SPECTROSCOPY	105
5.2	INFRARED RADIATION	107
5.3	NEAR INFRARED SPECTROSCOPY	116
5.4	SUMMARY	135
6	<u>QUANTITATIVE CALIBRATION METHODOLOGIES APPLICABLE TO NIRS</u>	137
6.1	THE BEER-LAMBERT LAW	137
6.2	UNIVARIATE CALIBRATION	138
6.3	MULTIVARIATE CALIBRATION	139
6.4	PRINCIPAL COMPONENTS	142
6.5	VALIDATIONS OF MODELS	166
6.6	PREDICTION OF UNKNOWN SAMPLES	167
6.7	IMPORTANT CALIBRATION AND REGRESSION STATISTICS	169
6.8	TECHNIQUES FOR COMPARING CALIBRATION MODELS	174
6.9	PARTIAL LEAST SQUARES (PLS) REGRESSION	177
6.10	NON-LINEARITIES IN CALIBRATION	188
6.11	IMPORTANT CONSIDERATIONS WHEN DEVELOPING CALIBRATION MODELS	191
6.12	SUMMARY	202
7	<u>QUALITATIVE ANALYSIS AND SAMPLE DISCRIMINATION TECHNIQUES</u>	203
7.1	CLUSTER ANALYSIS	203
7.2	ASSIGNING UNKNOWN SAMPLES TO PRE-EXISTING GROUPS	209
7.3	COMPARISONS BETWEEN CLASSIFICATION METHODS	218
8	<u>SPECTRAL PRE-PROCESSING TECHNIQUES</u>	220
8.1	SCATTER CORRECTION METHODS	220
8.2	DERIVATIVE SPECTRA	227
8.3	OTHER TRANSFORMATIONS	231
8.4	COMPARISONS OF PRETREATMENTS IN THE LITERATURE	233
9	<u>LITERATURE REVIEW OF QUANTITATIVE NIRS CALIBRATIONS FOR RELEVANT LIGNOCELLULOSIC COMPONENTS</u>	236
9.1	STUDIES ON DG SAMPLES – DRY AND OF A HOMOGENEOUS PARTICLE SIZE	237
9.2	STUDIES ON DU SAMPLES – DRY AND OF A HETEROGENEOUS PARTICLE SIZE	245
9.3	STUDIES ON WU SAMPLES - WET AND OF A HETEROGENEOUS PARTICLE SIZE	248
9.4	SUMMARY	251
10	<u>BACKGROUND ON MISCANTHUS</u>	253
10.1	BACKGROUND ON GRASSES	253
10.2	ESTABLISHMENT AND DEVELOPMENT CYCLE OF MISCANTHUS	260

10.3	YIELDS	264
10.4	COSTS INVOLVED IN MISCANTHUS PRODUCTION	264
10.5	LIGNOCELLULOSIC PROPERTIES OF MISCANTHUS	265
10.6	PREVIOUS NIRS RESEARCH WITH MISCANTHUS	269
10.7	MISCANTHUS IN IRELAND	271
11	<u>GENERAL ANALYSIS METHODOLOGY</u>	275
11.1	SAMPLE PREPARATION	275
11.2	NIRS CONDITIONS	281
11.3	MOISTURE AND ASH ANALYSIS	282
11.4	EXTRACTIVES CONTENT	283
11.5	HYDROLYSIS PROCEDURE	285
11.6	ELEMENTAL ANALYSIS	287
11.7	PRECISION CRITERIA FOR REFERENCE ANALYTICAL METHODS	290
12	<u>ANALYSIS OF SUGARCANE BAGASSE AND DEVELOPMENT OF QUANTITATIVE NIRS CALIBRATIONS</u>	291
12.1	BACKGROUND ON SUGARCANE BAGASSE	291
12.2	METHODOLOGY EMPLOYED	301
12.3	RESULTS AND DISCUSSION	305
12.4	SUMMARY	344
13	<u>DEVELOPMENT OF NIRS QUANTITATIVE CALIBRATIONS FOR THE LIGNOCELLULOSIC COMPONENTS OF PEAT SAMPLES</u>	345
13.1	BACKGROUND ON PEAT	345
13.2	METHODOLOGY	349
13.3	RESULTS	351
13.4	SUMMARY	369
14	<u>QUALITATIVE ANALYSIS OF MISCANTHUS SAMPLES</u>	371
14.1	METHODOLOGY	371
14.2	COMPARISON OF SPECTRA	377
14.3	PCA OF <i>MISCANTHUS</i> SPECTRA	379
14.4	DISCRIMINATION BETWEEN SAMPLES	381
14.5	SUMMARY	393
15	<u>DEVELOPMENT OF NIRS QUANTITATIVE CALIBRATIONS FOR MISCANTHUS SAMPLES</u>	396
15.1	DEVELOPMENT OF NIRS MODELS	396
15.2	RESULTS	399
15.3	SUMMARY	436

<u>16</u>	<u>LIGNOCELLULOSIC PROPERTIES OF MISCANTHUS AND EVALUATION OF THE CROP AS A FEEDSTOCK FOR BIOREFINING</u>	438
16.1	CHEMICAL COMPOSITION OF PLANT FRACTIONS	438
16.2	COMPARISON BETWEEN MISCANTHUS VARIETIES	443
16.3	COMPARISONS BETWEEN DF AND DS SAMPLES	447
16.4	CHANGES DURING THE HARVEST WINDOW	449
16.5	POTENTIAL YIELDS FROM BIOREFINING	457
16.6	SUMMARY	462
<u>17</u>	<u>ANALYSIS OF WASTE AND OTHER FEEDSTOCKS</u>	464
17.1	THERMOCHEMICAL BIOREFINING TECHNOLOGIES	464
17.2	AGRICULTURAL WASTES	465
17.3	MUNICIPAL WASTES	481
17.4	POTENTIAL YIELDS FROM BIOREFINERIES	494
17.5	NATIONAL PROJECTIONS	497
17.6	OTHER ENERGY CROPS	498
17.7	SUMMARY	499
<u>18</u>	<u>THE DIBANET PROJECT</u>	501
18.1	CONCEPT	501
18.2	NIRS WORK ON LATIN AMERICAN FEEDSTOCKS	503
18.3	NIRS SPECTRA AND CALIBRATIONS FOR PRETREATED MISCANTHUS SAMPLES	509
18.4	SUMMARY	514
<u>19</u>	<u>CONCLUSION</u>	516
19.1	QUALITY OF THE ANALYTICAL DATA	516
19.2	DIFFERENCES BETWEEN THE SPECTRAL DATASETS	518
19.3	WAVELENGTH REGIONS FOR REGRESSION	519
19.4	SPECTRAL PRETREATMENTS	521
19.5	QUALITY OF THE QUANTITATIVE NIRS MODELS	523
19.6	SUITABLE FEEDSTOCKS FOR BIOREFINING IN IRELAND	525
19.7	POSSIBLE FUTURE DEVELOPMENTS OF THE RESEARCH	526

List of Figures

FIGURE 1-1: THE VARIOUS PRE-TREATMENT AND SUBSEQUENT CONVERSION TECHNOLOGIES POSSIBLE FOR THE TREATMENT OF LIGNOCELLULOSICS.....	4
FIGURE 2-1: THE FISCHER PROJECTION OF THE ALDOSE D-GLUCOSE AND THE KETOSE D-FRUCTOSE.....	12
FIGURE 2-2: THE CHAIR AND BOAT FORMS FOR HEXOSES.....	13
FIGURE 2-3: SOME IMPORTANT MONOSACCHARIDES OF RELEVANCE TO LIGNOCELLULOSIC MATERIALS, IN THE CONFIGURATIONS AND CONFORMATIONS THAT OCCUR IN NATURE.....	15
FIGURE 2-4: SOME DISACCHARIDES.....	17
FIGURE 2-5: AMYLOSE AND AMYLOPECTIN.....	18
FIGURE 2-6: THE β -LINKED GLUCOPYRANOSIDE RESIDUES AND THE INTER- AND INTRA-MOLECULAR H-BONDING OF CELLULOSE.....	19
FIGURE 2-7: A REPRESENTATION OF A POSSIBLE MOLECULAR ARCHITECTURE OF THE CELLULOSE MOLECULE.....	20
FIGURE 2-8: A PROPOSED REPRESENTATION, BASED ON NMR STUDIES, OF THE FIBRIL STRUCTURE.....	21
FIGURE 2-9: THE PHENYLPROPANE UNITS THAT FORM THE STRUCTURAL BASIS OF THE LIGNIN POLYMER.....	25
FIGURE 2-10: A MODEL OF SPRUCE LIGNIN (ALDER, 1977).....	26
FIGURE 2-11: A SCHEMATIC REPRESENTATION OF THE PRIMARY (P) AND SECONDARY (S1, S2 AND S3) CELL WALLS OF A SOFTWOOD TRACHEID.....	27
FIGURE 2-12: A BASIC THEORETICAL REPRESENTATION OF THE LIGNOCELLULOSIC MATRIX.....	29
FIGURE 3-1: STEPS INVOLVED IN THE ACID HYDROLYSIS OF CELLULOSE.....	39
FIGURE 3-2: AN ILLUSTRATION OF SOME OF THE MOST COMMON TRANSITIONS THAT CAN OCCUR WITH UV-VIS RADIATION.....	44
FIGURE 3-3: AN EXAMPLE OF A TRANSITION IN A PRIMARY AMINE.....	44
FIGURE 3-4: THE ELECTRONIC, VIBRATIONAL, AND ROTATIONAL ENERGY STATES OF A MOLECULE.....	45
FIGURE 3-5: A DIONEX ACCELERATED SOLVENT EXTRACTOR (ASE) 200.....	57
FIGURE 3-6: A SCHEMATIC FOR THE ASE 200.....	58
FIGURE 3-7: THE ZYMARK TURBOVAP LV.....	59
FIGURE 3-8: TGA OF A SAMPLE OF SAWDUST.....	62
FIGURE 4-1: POLYMERS USED IN IC COLUMNS.....	69
FIGURE 4-2: ELECTROSTATIC BINDING IN ANION EXCHANGE RESINS.....	70
FIGURE 4-3: AN EXAMPLE OF AN OXYANION OF MANNOSE.....	70
FIGURE 4-4: TWO TYPICAL WAVEFORMS FOR PAD IN ALKALINE SOLUTIONS AT A GOLD WORKING ELECTRODE.....	72
FIGURE 4-5: SOME OF THE VARIOUS PEAK TYPES POSSIBLE IN CHROMELEON.....	79
FIGURE 4-6: DUAL NG1 COLUMNS IN THE IC SYSTEM.....	83
FIGURE 4-7: THE STANDARD "CARBOHYDRATES" WAVEFORM.....	83
FIGURE 4-8: CHROMATOGRAMS OBTAINED UPON THE ANALYSIS OF THE THREE SAMPLES SENT BY THE AUTHOR TO DIONEX.....	85
FIGURE 4-9: CHROMATOGRAM OBTAINED FROM A SUGAR STANDARD SOLUTION.....	87
FIGURE 4-10: CHROMATOGRAM OBTAINED WHEN A NON-DILUTED HYDROLYSATE OF EASTERN COTTONWOOD.....	87

FIGURE 4-11: THE RETENTION TIMES (AS A FRACTION OF THE RETENTION TIME ASSOCIATED WITH A 59.5 MM SODIUM ACETATE LOADING) FOR THE SEVEN SUGARS UNDER VARIOUS ACETATE LOADING CONCENTRATIONS.	89
FIGURE 4-12: FOUR CHROMATOGRAMS SHOWING THE RESOLUTION BETWEEN THE MAJOR SUGARS FOR SAMPLES WHERE THE PROPORTIONS OF THESE SUGARS VARY.	92
FIGURE 4-13: A PLOT OF PEAK AREA FOR SUGAR SOLUTIONS CORRESPONDING TO (UNDILUTED) HYDROLYSATES CONTAINING VARIOUS CONCENTRATIONS OF MANNOSE OR GLUCOSE	95
FIGURE 4-14: TWO GENERAL CHROMATOGRAPHY SEQUENCES THAT WERE USED IN THE RESEARCH	100
FIGURE 4-15: ZOOMED-UP CHROMATOGRAM FOCUSING ON THE PEAK DELIMITERS FOR THE SEVEN SUGARS	103
FIGURE 5-1: POTENTIAL INTERACTIONS THAT MAY OCCUR BETWEEN IR RADIATION AND A SOLID SAMPLE.....	108
FIGURE 5-2: SOME POTENTIAL VIBRATIONS AND ABSORPTIONS OF THE ALCOHOLIC HYDROXYL GROUP.	109
FIGURE 5-3: REPRESENTATION OF THE HARMONIC (A) AND ANHARMONIC (B) MODELS FOR THE POTENTIAL ENERGY OF A DIATOMIC MOLECULE.	111
FIGURE 5-4: SOME OF THE MAIN MODES OF MEASUREMENT EMPLOYED IN NIRS.....	119
FIGURE 5-5: REPRESENTATION OF THE SAMPLE PRESENTATION, LIGHT RADIATION, AND REFLECTED LIGHT DETECTION OF A DIFFUSE REFLECTANCE SYSTEM.	120
FIGURE 5-6: A SCHEMATIC OF THE REFLECTANCE DETECTOR AND THE TWO TYPES OF SENSORS USED IN THE FOSS XDS RCA.....	123
FIGURE 5-7: THE FOSS XDS MONOCHROMATOR.....	124
FIGURE 5-8: A GENERAL REPRESENTATION OF THE X-H VIBRATIONAL BANDS THAT CAN BE PRESENT IN AGRICULTURAL PRODUCTS..	130
FIGURE 5-9: THE SPECTRUM OF PURE WATER IN TRANFLECTANCE FORM ($\log(1/R)$) AND ITS FOURTH-ORDER DERIVATIVE.....	133
FIGURE 5-10: INFLUENCE OF A 3 ^o C CHANGE IN ROOM TEMPERATURE.....	135
FIGURE 6-1: A REGRESSION PLOT FOR MLR USING THE ABSORBANCES AT 2 WAVELENGTHS	141
FIGURE 6-2: THE NEW PCA LOADING VECTORS	151
FIGURE 6-3: ABSORBANCES OF GROUPS OF MATERIALS IN MULTIDIMENSIONAL SPACE.....	156
FIGURE 6-4: ILLUSTRATIONS OF HOW THE BOUNDING ELLIPSOIDS CAN CHANGE IN SIZE, SHAPE AND ORIENTATION ACCORDING TO WHAT MATRICES ARE USED TO CALCULATED THE MD	157
FIGURE 6-5: EFFECTS OF AN OUTLIER ON A PCA LOADING VECTOR.....	162
FIGURE 6-6: AN INFLUENCE PLOT WITH THEORETICAL EXAMPLES.....	163
FIGURE 6-7: A PREDICTION WITH DEVIATION PLOT FOR THE EXTRACTIVES-FREE GLUCOSE CONTENT OF 13 PEAT SAMPLES.....	167
FIGURE 6-8: A CHART PLOTTING THE PRESS STATISTIC AND SECV FOR PLS MODELS.....	176
FIGURE 6-9: AN EXPLAINED Y-VARIANCE PLOT FOR (FULL) CROSS-VALIDATION OF A PLS2 MODEL FOR THE 6 LIGNOCELLULOSIC SUGARS IN 52 WET PEAT SAMPLES.....	188
FIGURE 6-10: GLUCOSE CONTENTS OF SAMPLES IN THE CALIBRATION SET AND PREDICTED GLUCOSE CONTENTS OF SAMPLES NOT ANALYSED BY REFERENCE METHODS..	197
FIGURE 6-11: A HISTOGRAM PLOT FOR THE GLUCOSE CONTENT (EXTRACTIVES-FREE BASIS) OF 112 MISCANTHUS SAMPLES.	199
FIGURE 6-12: A HISTOGRAM BASED ON A CLUSTER ANALYSIS OF THE DATA-SET DESCRIBED IN FIGURE 6-11.	201
FIGURE 8-1: ILLUSTRATIONS OF SCATTERING EFFECTS.	223
FIGURE 8-2: MSC-TRANSFORMED SPECTRA.	224
FIGURE 10-1: AN ILLUSTRATION OF THE MAIN FRACTIONS OF MOST GRASSES	254

FIGURE 10-2: MISCANTHUS DRY MATTER YIELD, MOISTURE CONTENT, AND ASH CONCENTRATION AT THREE SITES FOR HARVESTS AT DECEMBER AND FEBRUARY (FOR THE CROP GROWN IN 1994 AND 1996) AND AT MARCH (FOR THE 1994 CROP)	262
FIGURE 10-3: DECREASE IN THE HARVESTABLE AMOUNT OF MISCANTHUS OVER THE COURSE OF SIX MONTHS	263
FIGURE 10-4: THE UV SPECTRA OF MILLED WOOD LIGNINS FROM ARUNDO DONAX AND M. SINENSIS.....	266
FIGURE 10-5: A MAP SHOWING THE MISCANTHUS PLANTATIONS ACROSS IRELAND	272
FIGURE 10-6: MISCANTHUS MODELLED PRODUCTIVITY MAPS FOR IRELAND.	273
FIGURE 11-1: AN ILLUSTRATION OF THE SEQUENTIAL METHOD ANALYTICAL SEQUENCE EMPLOYED FOR THE ANALYSIS OF STRUCTURAL CARBOHYDRATES IN PREPARED BIOMASS SAMPLES.	276
FIGURE 11-2: THE METHODOLOGY EMPLOYED IN PROCESSING BIOMASS	278
FIGURE 11-3: THE SAMPLE SEQUENCE USED IN ELEMENTAL ANALYSIS BATCHES.....	289
FIGURE 12-1: A SIMPLIFIED FLOW-DIAGRAM REPRESENTATION OF A SUGAR MILL.....	294
FIGURE 12-2: THE VISIBLE-NIR SPECTRA FOR THE BAGASSE DB SCANS.....	313
FIGURE 12-3: SELECTED BAGASSE SPECTRA.....	315
FIGURE 12-4: WU BAGASSE SPECTRA.	316
FIGURE 12-5: GLUCOSE REGRESSION PLOTS.....	325
FIGURE 13-1: SPECTRA OF PEAT SAMPLES	356
FIGURE 14-1: CHART REPRESENTING THE NUMBER OF MISCANTHUS PLANTS ACCORDING TO DATE AND SITE LOCATION, DURING THE PERIOD OCTOBER 2007 TO APRIL 2008, AS PART OF THE DEPARTMENT OF AGRICULTURE PROJECT.....	372
FIGURE 14-2: PHOTOGRAPHS OF MISCANTHUS STEMS	375
FIGURE 14-3: PHOTOGRAPHS OF SOME MISCANTHUS FRACTIONS.....	376
FIGURE 15-1: GLUCOSE REGRESSION PLOTS.....	410
FIGURE 15-2: XYLOSE REGRESSION PLOTS	411
FIGURE 15-3: ARABINOSE REGRESSION PLOTS.....	412
FIGURE 15-4: GALACTOSE REGRESSION PLOTS	413
FIGURE 15-5: RHAMNOSE REGRESSION PLOTS.	414
FIGURE 15-6: MANNOSE REGRESSION PLOTS.....	415
FIGURE 15-7: KLASON LIGNIN (KL) REGRESSION PLOTS.....	416
FIGURE 15-8: ACID SOLUBLE LIGNIN (ASL) REGRESSION PLOTS.....	417
FIGURE 15-9: URONIC ACIDS (UA) REGRESSION PLOTS.	418
FIGURE 15-10: 95% ETHANOL-SOLUBLE EXTRACTIVES REGRESSION PLOTS.....	419
FIGURE 15-11: ASH REGRESSION PLOTS.....	420
FIGURE 15-12: ACID INSOLUBLE RESIDUE (AIR) REGRESSION PLOTS	421
FIGURE 15-13: ACID INSOLUBLE ASH (AIA) REGRESSION PLOTS.....	422
FIGURE 15-14: PREDICTED CARBON VS. REFERENCE CARBON CONTENT FOR THE DT MODEL.....	423
FIGURE 15-15: NITROGEN REGRESSION PLOTS.....	424
FIGURE 15-16: PREDICTED WET-BASIS MOISTURE CONTENT (MC) VS. REFERENCE MC REGRESSION PLOTS.	432
FIGURE 15-17: PREDICTED WET-BASIS MOISTURE CONTENT (MC) VS. REFERENCE MC FOR THE MODEL BASED ON SAMPLES OF LEAVES AND STEMS.	433

FIGURE 16-1: THE INCREASED YIELD PER HECTARE OF LEVULINIC ACID, OVER THAT EXPERIENCED FROM THE LATEST POINT IN THE HARVEST WINDOW, ASSOCIATED WITH EARLIER HARVESTS.	461
FIGURE 18-1: THE DIBANET PROCESS FLOW.	502
FIGURE 18-2: LEVULINIC ACID.....	510
FIGURE 18-3: PREDICTED Y VS. REFERENCE Y FOR THE 23 PRETREATED MISCANTHUS SAMPLES AND THE ONE SAMPLE THAT WAS NOT PRETREATED.....	512

List of Tables

TABLE 3-1: SOME ASL ABSORBTIVITY CONSTANTS THAT WERE FOUND IN THE LITERATURE	48
TABLE 3-2: TGA RESULTS FOR THE MOISTURE AND OTHER CONTENTS (% WET BASIS) OF IRISH FEEDSTOCKS	63
TABLE 3-3: ASSUMED MOISTURE, LIGNOCELLULOSIC AND ASH CONTENTS (% WB) DERIVED FROM THERMOGRAMS.	64
TABLE 4-1: AVERAGE PEAK RESULTS (WITH STANDARD DEVIATIONS IN BRACKETS) FOR THE THREE INJECTONS OF SAMPLE 1	84
TABLE 4-2: THE RETENTION TIMES (RT) OF SEVEN SUGARS UNDER DIFFERENT SODIUM ACETATE LOADINGS	89
TABLE 4-3: CHROMATOGRAPHY STATISTICS (AFTER THE SYSTEM HAS EQUILIBRATED) FOR A SUGAR STANDARD SOLUTION	93
TABLE 4-4: THE HIGHEST CONCENTRATIONS OF SEVEN BIOMASS SUGARS THAT WERE USED IN THE US FPL LABORATORY'S CALIBRATION CURVE.	96
TABLE 4-5: THE LOWEST AND HIGHEST CONCENTRATIONS USED IN THE CALIBRATION CURVES FOR THE 7 BIOMASS SUGARS.	96
TABLE 4-6: COMPARISON OF THE RSDs OF THE RRFs FOR DIFFERENT PEAK CLASSIFICATION METHODS.	102
TABLE 5-1: THE DATA COLLECTION PARAMETERS THAT ARE AVAILABLE AND THE SETTINGS THAT ARE USED IN THE FOSS XDS	125
TABLE 5-2: TENTATIVE BAND ASSIGNMENTS FOR LIGNIN AND THE STRUCTURAL POLYSACCHARIDES	131
TABLE 10-1: MASS BALANCE DATA (% DM) FOR COMPONENTS OF BARLEY AND WHEAT STRAWS.....	257
TABLE 10-2: THE AVERAGE CHEMICAL COMPOSITION OF VARIOUS ANATOMICAL FRACTIONS (% WHOLE DRY MASS) OF CORN-STOVER AND SWITCHGRASS SAMPLES.	257
TABLE 10-3: CHEMICAL COMPOSITION (% DM) OF TWO GRASSES.....	259
TABLE 10-4: LIGNOCELLULOSIC CONTENTS OF MISCANTHUS, TAKEN FROM SEVERAL SOURCES.	268
TABLE 10-5: THE RELATIVE AMOUNTS (% DRY MATTER) OF LIGNOCELLULOSIC CONSTITUENTS IN THE VARIOUS ANATOMICAL FRACTIONS OF MISCANTHUS X GIGANTEUS HARVESTED IN NOVEMBER AND FEBRUARY.....	268
TABLE 10-6: CELL WALL COMPOSITION OF MISCANTHUS SPECIES AND GENOTYPES IN NOVEMBER AND FEBRUARY HARVESTS.....	269
TABLE 10-7: TOTAL HECTARES APPLIED FOR MISCANTHUS ESTABLISHMENT GRANTS OVER THE PERIOD 2007-2010	272
TABLE 11-1: A SUMMARY OF THE VARIOUS METHODS USED FOR THE REMOVAL OF EXTRACTIVES FROM SAMPLES	285
TABLE 11-2: STANDARD DEVIATION OF DUPLICATE (SDD) LIMITS FOR A RANGE OF CONSTITUENTS.	290
TABLE 12-1: COMPOSITIONAL DATA (% OF DRY MATTER) FOR SUGARCANE BAGASSE SAMPLES	296
TABLE 12-2: PROPORTIONS OF THE DIFFERENT MONOSACCHARIDES OBTAINED UPON HYDROLYSIS OF HEMICELLULOSE FRACTIONS.	297
TABLE 12-3: THE YIELD OF HEMICELLULOSES AND LIGNIN (% OF DRY MATTER) SOLUBLISED DURING THE SUCCESSIVE TREATMENTS OF DEWAXED BAGASSE	298
TABLE 12-4: THE CONTENT OF SUGARS AND URONIC ACIDS (% DRY WEIGHT) IN THE ISOLATED HEMICELLULOSE FRACTIONS	298
TABLE 12-5: THE CONTENTS OF SUGARS AND URONIC ACIDS IN THE ISOLATED HEMICELLULOSE FRACTIONS.....	298
TABLE 13-1: COMPOSITIONS OF A HIGH MOOR PEAT AND A LOW MOOR PEAT, AS WELL AS THAT OF THE VEGETATIVE MATTER THAT GROWS ON THEM.....	347
TABLE 13-2: STATISTICS FOR THE PCR FOR A RANGE OF PEAT CONSTITUENTS	355
TABLE 13-3: CLASSIFICATION OF THE RER VALUES FOR THE BEST MODEL FOR EACH CONSTITUENT AND DATASET.....	370
TABLE 16-1: RESULTS FOR THE ANOVA TESTS REGARDING WHETHER THERE IS A SIGNIFICANT DIFFERENCE IN THE CONSTITUENT VALUE MEANS BETWEEN THE PLANT FRACTIONS FOR THREE-STEM-SECTION-PLANTS.....	442

TABLE 16-2: RESULTS FROM ANOVA TESTS TO DETERMINE IF THERE IS A SIGNIFICANT DIFFERENCE IN THE CONSTITUENT VALUE MEANS OF THE “EARLY” AND “LATE” WP SAMPLES.	455
TABLE 16-3: CONVERSION FACTORS AND YIELDS FOR THE HYDROLYSIS, ETHANOL-PRODUCING TECHNOLOGIES A-E	460
TABLE 17-2: TOTAL QUANTITIES OF ANIMAL AND SILAGE WASTES (WET TONNES PER YEAR).....	470
TABLE 17-3: MAJOR MASS CONSTITUENTS (% DM) IN SOME ANIMAL WASTES.....	471
TABLE 17-4: RELEVANT MASS COMPOSITIONS (% DRY MATTER) OF CATTLE FED FOUR DIFFERENT DIETS	471
TABLE 17-5: QUANTITY OF PIG WASTE ACCORDING TO WEIGHT OF ANIMAL	472
TABLE 17-6: MEAN AND RANGE VALUES FOR IMPORTANT CONSTITUENTS (% DRY MATTER) IN THE COMPOSITION OF PIG FAECES ..	472
TABLE 17-7: COMPOSITION (% DRY MATTER) OF PIG FEED AND PIG FAECES.	472
TABLE 17-8: THE COMPOSITION OF SMC TAKEN FROM VARIOUS PRODUCERS IN THE REPUBLIC AND NORTHERN IRELAND	478
TABLE 17-9: DATA ON THE COMPOSITION OF SMC (% DRY MATTER), FROM THE ANALYSIS OF 20 SAMPLES BY MICHAEL MAHER .	478
TABLE 18-1: THE NUMBER OF SAMPLES OF SUGARCANE BAGASSE AND TRASH SCANNED BY CTC IN EACH DATASET.	504

List of Abbreviations

ADF	Acid detergent fibre
ADL	Acid detergent lignin
AF	Ash free basis
AIA	Acid Insoluble Ash
AIA_EF	Acid Insoluble Ash (extractives free)
AIR	Acid Insoluble Residue
AIR_EF	Acid Insoluble Residue (extractives free)
ARA	Arabinose content
ARA_EF	Arabinose content (extractives-free)
ARA_EF_SRS	Arabinose content (extractives-free) using the individual sugar recoveries from the batch
ARA_SRS	Arabinose content using the individual sugar recoveries from the batch
ASA	Acid soluble Ash
ASE	Accelerated solvent extraction
ASL	Acid soluble lignin
ASL_EF	Acid soluble lignin (extractives free)
AV.	Average
Av abs. diff	Average absolute difference (see Section 13.3.4)
BC	Bray Curtis distance
BMW	Biodegradable municipal waste
BSES	BSES Limited
C	Carbon
Cal:Val	Number of samples in the calibration and validation sets
CB	City block distance
CTC	Centro de Tecnologia Canavieira, a partner in the DIBANET project
CV	Cross-validation.
DB	Dry (hand) sieved fraction at BSES
DCM	Data collection method (for the FOSS XDS NIR system).
DF	A dry fine sample with a particle size less than 180 microns.
DF-E	Moisture content of DF samples prior to the removal of extractives
DF-E Dishes	Moisture content of DF samples after the removal of extractives
DG	Dry and ground sample (all particles less than 850 microns in diameter).
DH	A DG fraction that has been scanned in a different method (see Section 11.1)
DM	Dry matter
DMB	Diesel miscible biofuel
DS	Dry sieved fraction of a sample with a particle size between 180 and 850 microns.
DS-E	Moisture content of DS samples prior to the removal of extractives
DS-E Dishes	Moisture content of DS samples after the removal of extractives
DT	A DS fraction that has been scanned in a different method (see Section 11.1)
DU	Dry unground fraction scanned at BSES
DW	Weighted average scan of several DU fractions, scanned at BSES
E-H	Moisture content of DS samples prior to the analytical hydrolysis procedure
EF	Extractives-free basis
EIA	Ethanol Insoluble Ash
EIA_EF	Ethanol Insoluble Ash (extractives-free)
EMSC	Extended multiplicative scatter correction.
ESA	Ethanol Soluble Ash
ESTD	External standard
EU	Euclidean distance
EXTR_CV	Extractives content (% whole mass basis) as measured directly in collection vials
EXTR_PD	Extractives content (% whole mass basis) as measured by the mass loss in ethanol extraction

"F"	Dead leaf blade sample of Miscanthus
F	Factor (in PLSR)
F-F Test	Number of PLS factors chosen using the F-test criterion (Osten, 1988)
F-Haaland's	Number of PLS factors chosen using the Haaland and Thomas (1988) criterion
F-Min Press	Number of PLS factors associated with the minimum PRESS value
F-UNSCR	Number of PLS factors chosen by the Unscrambler X software
F-Wold's	Number of PLS factors chosen using Wold's criterion
F-Wold's 0.95	Number of PLS factors chosen using Wold's criterion and a threshold value of 0.95
F-Wold's 0.9	Number of PLS factors chosen using Wold's criterion and a threshold value of 0.90
FL	Miscanthus flower sample
GAL	Galactose content
GAL_EF	Galactose content (extractives-free)
GAL_EF_SRS	Galactose content (extractives-free) using the individual sugar recoveries from the batch
GAL_SRS	Galactose content using the individual sugar recoveries from the batch
GLU	Glucose content
GLU_EF	Glucose content (extractives-free)
GLU_EF_SRS	Glucose content (extractives-free) using the individual sugar recoveries from the batch
GLU_SRS	Glucose content using the individual sugar recoveries from the batch
"H"	Dead leaf sheath sample of Miscanthus
H	Hydrogen
ha	Hectare
HAL	Hierarchical average linkage clustering
HCL	Hierarchical complete linkage clustering
HML	Hierarchical median linkage clustering
HP	Harvested plant sample of Miscanthus
HPAEC-PAD	High performance anion exchange chromatography with pulsed amperometric detection.
HSL	Hierarchical single linkage clustering
IC	Ion chromatography
ISTD	Internal standard
K	Live leaf blade sample of Miscanthus
KDC	K-Medians clustering
KL	Klason lignin
KL_EF	Klason lignin (extractives free)
KMC	K-Means clustering
KURT/KRT	Kurtosis statistic
λ	Wavelength
LA	Levulinic acid
LDA	Linear discriminant analysis.
LvA	Levulinic acid
M	Live leaf sheath sample of Miscanthus
MAN	Mannose content
MAN_EF	Mannose content (extractives-free)
MAN_EF_SRS	Mannose content (extractives-free) using the individual sugar recoveries from the batch
MAN_SRS	Mannose content using the individual sugar recoveries from the batch
MC	Moisture content (wet basis)
MIR	Mid infrared
MLR	Multiple linear regression
MSC	Multiplicative scatter correction.
N	Nitrogen
NDF	Neutral detergent fibre
NIR	Near infrared
NIRS	Near infrared spectroscopy
odt	Oven dried tonnes
PC	Principal Component

PCA	Principal Component Analysis
PCR	Principal Component Regression
PLS	Partial Least Squares
PLSR	Partial Least Squares Regression
PLS-DA	Partial least squares discriminant analysis
PLS- λ	Wavelength region used for PLSR
Pre.	Pretreatment
PRESS	Prediction residual error sum of squares
QDA	Quadratic discriminant analysis
R^2	Multiple correlation coefficient
R_{aal}^2	Multiple correlation coefficient for calibration
R_{valid}^2, R_{pred}^2	Multiple correlation coefficient for (independent) test-set validation
R_{CV}^2	Multiple correlation coefficient for cross validation
RCA	Rapid content analyser (part of the FOSS XDS NIR unit).
RER	Range error ratio
RER_{CV}	RER in cross-validation
RER_{pred}	RER using the independent validation set
RF	Reponse factor
RRF	Relative response factor
RHA	Rhamnose content
RHA_EF	Rhamnose content (extractives-free)
RHA_EF_SRS	Rhamnose content (extractives-free) using the individual sugar recoveries from the batch
RHA_SRS	Rhamnose content using the individual sugar recoveries from the batch
RMSEC	Root mean square error of calibration
RMSECV	Root mean square error of cross validation
$RMSECV_{MP}$	Root mean square error of cross validation using the number of PLS factors associated with minimum PRESS
RMSEP	Root mean square error of prediction
RPD	Ratio of standard error of performance to standard deviation (using the validation set)
RPD_{CV}	RPD in cross-validation
RSD	Relative standard deviation = (standard deviation)/average
RT	Retention time
S	Sulphur
Sam. Excl.	Samples excluded from the PLS model
SB	Sugarcane bagasse
SD	Standard deviation
SD_{EF}	Standard deviation of the extractives-free values for a given constituent and dataset
SD_{WM}	Standard deviation of the values (whole dry mass basis) for a given constituent and dataset
SDD	Standard deviation of duplicates
SEC	Standard error of calibration
SECV	Standard error of cross validation
SEL	Standard Error of Laboratory
SEP	Standard error of prediction
SG	Savitzky Golay
SIMCA	Soft independent modelling of class analogy
SKEW/SKW	Skew statistic
SNV	Standard normal variate
SNVDT	Standard normal variate followed by detrend
SPE	Solid phase extraction.
SR	Solid residues
SRS	Sugar recovery solution.
_SRS	e.g. GLU_SRS: Sugar data corrected according to the sugar recovery of the batch
St:Lf	(Number of stem samples):(number of leaf samples)
TOT	Sum of ARA, GAL, RHA, GLU, XYL, and MAN contents.

TOT_EF	Total sugars content (extractives-free)
TOT_EF_SRS	Total sugars content (extractives-free) using the individual sugar recoveries from the batch
TOT_SRS	Total sugars content using the individual sugar recoveries from the batch
UA	Uronic acids
UL	University of Limerick
wb	Wet basis
WC	Wet and chipped – A wet unground sample that has been scanned but not further processed
WM	Whole mass basis
WP	Whole plant sample
WU	Scan of wet unground sample.
X1	1 st metre of a stem
X1N	Nodes from the 1 st metre of a stem
X1T	Internodes from the 1 st metre of a stem
X2	2 nd metre of a stem
X2N	Nodes from the 2 nd metre of a stem
X2T	Internodes from the 2 nd metre of a stem
X3	3 rd metre of a stem
X3N	Nodes from the 3 rd metre of a stem
X3T	Internodes from the 3 rd metre of a stem
XYL	Xylan content
XYL_EF	Xylose content (extractives-free)
XYL_EF_SRS	Xylose content (extractives-free) using the individual sugar recoveries from the batch
XYL_SRS	Xylose content using the individual sugar recoveries from the batch

1 Introduction

1.1 Background

The world is going through a crucial transitional phase where circumstances dictate that economic activity and growth need to be decoupled from net energy expenditure. Fossil fuels currently serve the majority of our needs, both energetically and chemically (Bozell, 2001). However, their combustion is believed to be responsible for approximately three quarters of the anthropogenic emissions of CO₂ (IPCC, 2001), a major greenhouse gas that contributes to global warming (Ramaswamy et al., 2001). The Intergovernmental Panel on Climate Change estimates that, if nothing is done, the mean global temperature may rise by a further 1 to 3.5 °C by the end of the century, and result in a rising of sea levels by between 15 and 95 cm. Of all the fossil fuels, oil has the greatest effect – its consumption is the major reason for the changes mentioned; indeed it accounts for 50% of CO₂ emissions in the EU (OECD, 1997). Furthermore, given that more than 70% of the world's proven commercial oil reserves are located in the member countries of OPEC, the "Organization of the Petroleum Exporting Countries" (OECD, 1997), the supply of oil to the western world cannot be guaranteed in the future. Economic or political factors in the OPEC countries could change very quickly leading to an insufficient supply of oil to accommodate current needs. Currently there are no effective means to respond to such a threat (Cleveland and Kaufmann, 2003). The issue of oil dependency is particularly relevant for Ireland. Due to the low abundance of proven commercial fossil fuel reserves in the Republic, Ireland depends on imports for 77% of its energy supply (OECD, 1997).

In addition to the energy dependence, the world is also reliant on fossil fuels as feedstocks for the production of numerous chemicals and consumer products. It has been estimated that, of the approximately 170 chemical compounds produced in the US in volumes exceeding 4.5 million kg per year, 98% are derived from oil and natural gas (Szmant, 1989). The vast majority of modern synthetic products are also derived from oil. This is a dependency that has been increasing over the decades. For example, the production of US petroleum-based plastics grew by more than 400% between 1970 and 1990 (Bozell, 2001).

There are therefore many reasons to search for alternatives to fossil fuels. Electricity-producing renewable technologies, such as solar and wind, may satisfy to some degree our electricity needs,

but the short-term implementation of non-fossil-fuel-based chemicals and transport fuels will require the utilisation of biomass, particularly in view of the common consensus that hydrogen and fully-electrified cars are unlikely to become viable on a mass scale for at least a decade (Solomon and Banerjee, 2006). Governments around the world have recognised this and have implemented minimum targets for the implementation of alternative fuels in the future, with the expectation that biofuels will play a major role in these targets. For example, the EU has a mandatory target for the transport sector of 10% (by energy content, based on the consumption of petrol and diesel fuels) substitution of fossil fuels by 2020 (EC, 2003). In the USA the Energy Independence and Security Act of 2007 mandates for 36 billion US gallons of renewable fuels by 2022 (US Congress, 2007).

To date, all biofuels consumed in significant quantities in the EU have been produced from first-generation feedstocks; sugar, starch, and oil-based crops and wastes. The particular chemistry of these materials allows for their relatively easy conversion to transport fuels suitable for utilisation in existing motor engines. There are drawbacks with first generation biofuels (1GB), however. For instance, their costs can be high, particularly in Europe (Hamelinck et al., 2004). Also, the utilisation of high-quality arable land for their production has said to result in competition, between biofuel and food-requirements (the so-called fuel versus food issue), for this land and for the food/oil products that are produced. The rise in recent years in the prices of various crops (such as wheat, maize, and soya beans, which have dual roles as food and biofuel feedstocks), has been attributed, by some (The Economist, 2007, House of Commons, 2008), to the greatly increased demand for biofuels by many developed nations. Furthermore, there are socio-economic concerns regarding the effects of these increased prices and reduced food production on the public health and economies of developing nations, particularly where these nations divert indigenous food production towards satisfying the biofuel needs of the western-world (Ferrett, 2007, Diouf, 2007).

There are also concerns over the net benefits in terms of greenhouse gas abatement regarding many first generation feedstocks. Some studies have indicated that the net effect on greenhouse gases levels and global warming of a displacement of fossil fuels by 1GB could be only a small reduction (Hill et al., 2006a) or even an increase (Crutzen et al., 2007). For example, it has been estimated that the well-to-wheels greenhouse gas emission change, relative to petroleum, for ethanol from an average maize to ethanol biofuel plant in the USA is -19% (Wang et al., 2007). The requirement for significant quantities of fertiliser and agricultural work may also result in low net-energy balances when a full life cycle analysis of the biofuel is considered (Hill et al., 2006b). These many concerns regarding biofuels have led to calls for sustainability indexes to be developed that consider these myriad issues (BBC, 2008). Governmental incentive schemes now need to consider these when

deciding the level of support to give to various biofuels. Given the low net carbon savings associated with many first generation feedstocks, and their need for large excise relief in order to be competitive with fossil fuels, such indexing may result in the EU 2020 alternative fuels target being unattainable if these feedstocks are to be produced within Europe. It has been suggested (BBC, 2007) that, in the light of these concerns, the mandatory target should be abolished rather than letting Ireland and the EU be reliant on, potentially unethically sourced, imports.

Second generation biofuels (2GB) may offer a potentially cheaper and more sustainable means of producing transport fuels than first generation biofuels. 2GBs are sourced from lignocellulosic biomass feedstocks – plants and wastes that are principally composed of the biopolymers cellulose, hemicellulose and lignin. Dedicated lignocellulosic energy crops, such as *Miscanthus x giganteus*, are typically high-yielding, low cost, and low maintenance (Venturi and Venturi, 2003, Hill et al., 2006a). This results in superior energy and greenhouse gas balances compared with their first generation analogues. For example, switchgrass-derived ethanol has been estimated to produce 540% more renewable energy than the non-renewable energy consumed; with estimated greenhouse gas emissions being 94% lower than those from gasoline (Schmer et al., 2008). Also, while the production of annual first-generation feedstocks, such as maize, can be associated with soil degradation and the loss of organic carbon from the soil (Wang et al., 2007), certain perennial energy crops, such as *Miscanthus*, avoid the need for frequent tillage, while the extensive roots that they develop over their lifespan are said to result in an increase in soil organic carbon (Hansen et al., 2004, Grogan and Matthew, 2001). Considering this dynamic, Adler et al. (2007) calculated that ethanol from switchgrass and hybrid poplars reduced greenhouse gas emissions (over petrol-derived transport fuels) by approximately 115%. It has also been claimed that low-input high diversity mixtures of native grassland perennials can allow carbon-negative biofuels (Tilman et al., 2006).

Crucially, many biological waste materials (such as agricultural residues and municipal wastes) contain significant amounts of lignocellulose meaning that there is a potentially vast resource of material available of low (or negative) value. Utilising these wastes will provide some biofuel production without the need for dedicated agricultural land. Also, land that may be unsuitable for the production of food can be utilised for some lignocellulosic energy crops.

Unfortunately, the particular chemistry of the lignocellulose polymers and their inter-associations mean that their conversion to fuels has historically been difficult, expensive, and low-yielding (Hayes, 2008). However, there have been significant advances in the art in recent years (Solomon et al., 2007). The modern technologies that have resulted from research in this area can produce a

variety of end products according to the chemical components of the feedstock. This leads to the concept of a **biorefinery**, a biomass analogue to an oil refinery.

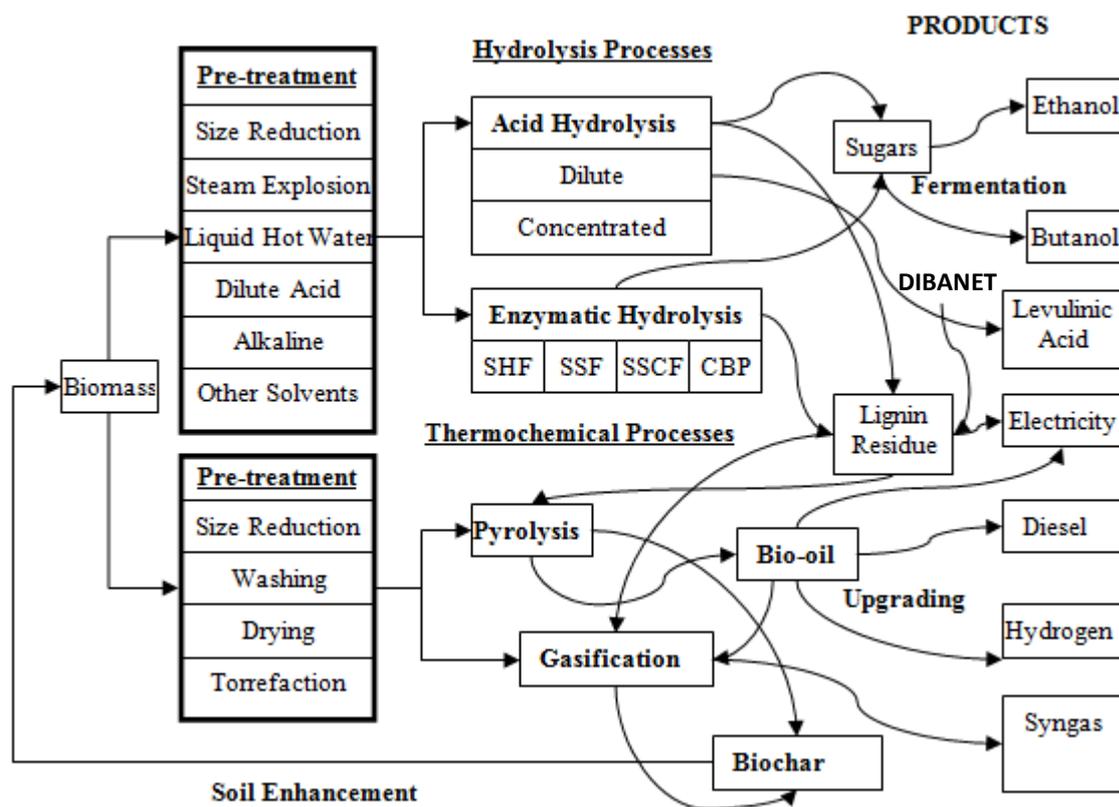


Figure 1-1: The various pre-treatment and subsequent conversion technologies possible for the treatment of lignocellulosics. The production of a biochar through pyrolysis can offer a feedback mechanism to the cycle by increasing biomass yields when it is applied to land. SHF = sequential hydrolysis and fermentation, SSF = simultaneous saccharification and fermentation, SSCF = simultaneous saccharification and co-fermentation; CBP = consolidated bio-processing. For details on the DIBANET process see Chapter 18.

A paper by the Author (Hayes, 2008) provides significant detail regarding the types of biorefining technologies that are being developed in order to provide biofuels and platform chemicals from lignocellulosic materials. The reader is referred to that article for a more thorough understanding of the mechanisms employed. In brief, there are two major pathways by which biorefineries operate: through hydrolytic mechanisms that aim to liberate free monosaccharides (sugars) from the lignocellulosic polysaccharides (cellulose and hemicellulose), and through thermochemical processes that degrade more extensively the components of both polysaccharides and lignin. The various technologies available for the hydrolytic or thermochemical processing of biomass, along with some of their potential products and possible pre-treatment steps, are illustrated in Figure 1-1. Hydrolysis processes generally use enzymes or acids to liberate the sugars which can then be fermented to

ethanol or other products, or alternatively chemically degraded to platform chemicals such as levulinic acid and furfural (as is the case with the DIBANET process, Chapter 18). The thermochemical processes that are typically used are either gasification or pyrolysis. The former produces a syngas (CO and H₂) from the biomass and this can then be catalytically reformed to an array of potential chemicals and transport fuels (Hayes, 2008). Pyrolysis involves the thermal degradation of biomass in the absence of oxygen and produces: a bio-oil from the condensable vapours generated; a biogas from the non-condensable vapours; and a biochar from the solid residue. Various parameters (temperature, residence time etc.) can be changed in order to modify the relative mass proportions of these three products as well as their properties.

At the Carbolea Biomass Research Group (www.carbolea.ul.ie) in the University of Limerick (UL) several biorefining processes are being developed and these all focus on either the chemical or thermochemical degradation of biomass for the production of chemicals and value-added products without the requirement for biotic activity (hydrolytic enzymes, fermentation, etc.). In addition to the DIBANET hydrolysis process, intermediate and slow pyrolysis studies are being undertaken and there is a pilot scale (10 kg of feedstock per hour) gasifier operational.

1.2 Basis of the Present Research

The research initially started out as a desk-based MSc examining the suitability of a range of Irish lignocellulosic feedstocks for potential utilisation in biorefining technologies. During the course of that research it became clear that secondary data regarding the lignocellulosic compositions of feedstocks relevant to Ireland and to Irish conditions were inadequate in many instances. For some feedstocks there were either no data available, or the analytical techniques used to characterise these materials (e.g. the detergent fibre methods, see Section 3.1.1) were not sufficiently accurate. Without a detailed and certain knowledge of the compositions of these materials it would not be possible to determine the potential yields that may result from their processing in biorefineries.

The Author therefore identified a critical need for a program of research to categorise to a high level of accuracy and precision the chemical parameters of lignocellulosic materials of relevance to biorefining technologies. The focus would primarily be on those constituents most relevant to the hydrolysis technologies (in particular the constituent monosaccharides of the structural polysaccharides), although some analyses for constituents relevant to the processing of these feedstocks in thermochemical processes would also take place.

It was recognised that a substantial amount of time would be required to: collect the biomass samples; process these to a form suitable for analysis; and to undertake the wet-chemical analyses of these materials. It was considered that rapid analytical techniques could therefore offer advantages in terms of sample throughput and the scope of the research. In a study of the literature, many different rapid analytical techniques were examined (e.g. thermogravimetric analysis, see Section 3.7) and it was concluded that near infrared spectroscopy (NIRS, Chapter 5) offered many advantages. These included the ability to develop quantitative and qualitative analytical methods for a wide variety of parameters for biomass samples that required minimal or no processing prior to their NIRS analysis.

NIRS is not a direct analytical method, however. The spectra that are obtained from samples cannot be quantitatively interpreted without the development of models that relate the variations seen in a number of spectra with the chemical variations seen in the samples. Chemometrics is the science of extracting information from chemical systems using statistical/mathematical techniques on data. In the case of NIRS it allows for the models linking spectra and chemical data to be developed. Hence, the wet chemical data would still need to be collected for samples in order to develop these models. However, the ultimate target was that sufficiently robust models would result so that these could be applied for the direct prediction of the properties of interest for unknown samples. In such a way NIRS could be developed as a primary analytical tool. This was a major target of the research in this Thesis.

The development of these NIRS models can be a lengthy process that requires a significant amount of laboratory work. The most accurate models tend to be focussed on a particular type of feedstock, which means that several models may be necessary for a diverse array of materials. Hence, while a large number of different sample types have been analysed by wet-chemical methods in this Thesis, only a few feedstocks were chosen for NIRS model development (although NIRS spectra have been collected for all of the samples analysed). The most important of these is *Miscanthus* (Chapters 10 and 14 to 16). This predominately lignocellulosic feedstock is a perennial C₄ grass that, when grown as an energy crop, can achieve high yields in Ireland with much lower monetary and energy costs in its supply cycle than many first generation feedstocks. It grows particularly well in the west of Ireland and, through Dr JJ Leahy, there is a great degree of knowledge and experience of it in UL.

The second feedstock for which NIRS models were developed was peat. This feedstock was analysed as a result of a project written by the Author and funded by Bord na Móna (see Chapter 13). The third feedstock was sugarcane bagasse (Chapter 12), the solid residue that results after sugars have been extracted from sugarcane in a sugarmill. This is not a feedstock of relevance to Ireland but is a

hugely abundant resource in Brazil and is also produced in significant quantities in Australia. Sugarcane bagasse is also a key feedstock for the DIBANET process (Chapter 18).

Most importantly, to date there have been no publications regarding the development of accurate NIRS quantitative models for these three feedstocks covering the wide range of constituents that the Author has developed models for in this Thesis. Furthermore, while there are articles in the literature regarding the development of NIRS models for many of these constituents for other feedstocks (e.g. straws), the vast majority of these are based on predicting the compositions of samples that have been dried and reduced to a homogeneous particle size. As will be described in Chapter 11, the work in this Thesis also develops models for samples in this state, which is considered to be the easiest state for model development, but in addition also develops models for samples in various other states of preparation. In particular, models have been developed for wet biomass that has either not been processed in any way (bagasse and peat samples) or has only been chopped by a sufficient amount to allow for presentation to the sample cell of the NIRS unit (*Miscanthus*).

Appendix B presents the results, in tabulated form, of a literature review conducted by the Author regarding the regression statistics for quantitative models developed for the important constituents of lignocellulosic materials. These can be compared with the results obtained by the Author. It can be seen that the levels of accuracy and precision for models developed in this Thesis, even for wet unground samples, are superior to most of those in Appendix B and also allow, for the first time, the chemical composition of the highly-promising energy crop *Miscanthus* to be accurately predicted in a matter of minutes.

It can therefore be said that the research presented in this Thesis is justified because it has provided valuable information regarding the compositional data of numerous lignocellulosic feedstocks of relevance to Ireland and because it has demonstrated improvements in the NIRS analysis of wet materials.

1.3 Structure of Thesis

- **Chapter 2 - “Chemistry of Lignocellulosics”:** This provides a background to the chemical constituents of lignocellulosic biomass that are relevant to their utilisation in biorefining technologies. Many of these constituents have NIRS models developed for their quantitative prediction in subsequent Chapters. For details regarding the various biorefining technologies that utilise these constituents the reader is referred to an earlier paper by the Author (Hayes, 2008).
- **Chapter 3 – “Reference Analytical Methods”:** This Chapter details the various wet-chemical methods that can be employed to quantify and characterise the important constituents detailed in Chapter 2. It demonstrates the literature review that the Author undertook in order to determine the most appropriate analytical methods to employ in this research. It concludes with justifications regarding why the final analytical methods used were chosen.
- **Chapter 4 – “Ion Chromatography (IC)”:** IC was selected as the most appropriate device for characterising the monosaccharides that constitute the lignocellulosic polysaccharides. Chapter 4 details the background to IC. The Author was also responsible for the purchase of a suitable IC instrument and in the development of accurate and reproducible chromatographic methods. These are also detailed.
- **Chapter 5 – “Theory Behind Near Infrared Spectroscopy”:** This provides the background to NIRS and also discusses its relevance to lignocellulosic materials. As with the IC system, the Author was responsible for the purchase and development of a suitable NIRS device for the University of Limerick laboratory and that system is detailed here.
- **Chapter 6 – “Quantitative Calibration Methodologies Applicable to NIRS”:** This chapter goes into detail regarding the chemometric techniques for developing quantitative NIRS models. It also provides explanations for many of the important statistics used.
- **Chapter 7 - “Qualitative Analysis and Sample Discrimination Techniques”:** Here techniques for the qualitative analysis and discrimination of samples are discussed.
- **Chapter 8 – “Spectral Pre-Processing Techniques”:** It is often the case that the best performing NIRS models do not utilise the raw NIRS spectra but instead work on spectra that have been transformed. Chapter 8 discusses the transformation methods that have been utilised in this Thesis.
- **Chapter 9 – “Literature Review of Quantitative NIRS Calibrations for Relevant Lignocellulosic Components”:** A wide literature review was carried out by the Author concerning the use of quantitative NIRS models for predicting the important lignocellulosic

constituents of biomass feedstocks. This Chapter summarises the important observations that were made in the course of this review. These informed much of the work that was done in subsequent sections of the Thesis. Numerous articles are cited throughout the Chapter and the important regression statistics from these are provided in Table form, according to constituent type and sample-state, in Appendix B. This allows for easy comparison with the results obtained in this Thesis.

- **Chapter 10 – “Background on Miscanthus”:** This section details the background to grasses, and Miscanthus in particular.
- **Chapter 11 – “General Analysis Methodology”:** Here the standard laboratory practices that were employed in the processing, wet-chemical analysis, and NIRS analysis of biomass samples are outlined. In some specific cases the analytical techniques may have differed somewhat from the general methods outlined here, in which case these are detailed in subsequent Chapters.
- **Chapter 12 – “Analysis of Sugarcane Bagasse and Development of Quantitative NIRS Calibrations”:** This Chapter summarises the work carried out on the sugarcane bagasse samples provided by the Bureau of Sugar Experimental Stations (BSES) in Brisbane, Australia. Since this is the first Chapter in which quantitative NIRS models are developed it goes into some detail regarding the steps that were involved (e.g. spectral transformations used, wavelength regions for model development, etc.). Subsequent chapters mostly present the final results of model development, i.e. the best models according to constituent and sample-state. However, these final models have been reached through a similar means of development to those outlined in Chapter 12. Many of the Figures and Tables for this Chapter are presented in Appendix C.
- **Chapter 13 – “Development of NIRS Quantitative Calibrations for the Lignocellulosic Components of Peat Samples”:** Here the work carried out on peat samples is presented and discussed. Many of the Figures and Tables for this Chapter are presented in Appendix D.
- **Chapter 14 – “Qualitative Analysis of Miscanthus Samples”:** This Chapter first details the sample collection and processing methodologies that were employed for the Miscanthus samples studied in this Chapter and Chapters 15 and 16. It then discusses the qualitative analytical methods that were developed to discriminate between samples on the basis of: plant fraction, stand age, sample collection period, and variety type. Many of the Figures and Tables for this Chapter are presented in Appendix E.
- **Chapter 15 – “Development of NIRS Quantitative Calibrations for Miscanthus Samples”:** The quantitative NIRS models for the important constituents of Miscanthus samples are

presented and discussed. Many of the Figures and Tables for this Chapter are presented in Appendix F.

- **Chapter 16 – “Lignocellulosic Properties of Miscanthus and Evaluation of the Crop as a Feedstock for Biorefining”:** Here ways in which the chemical properties of Miscanthus samples vary within and between plants are discussed. The compositions of a range of Miscanthus varieties are presented and compared against Miscanthus x *giganteus*, the only variety that is currently grown commercially in Ireland. This Chapter also examines the changes that take place in several biomass stands over the course of the harvest window, and discusses how these may relate to the potential yields of chemicals when the biomass is processed in a range of biorefining technologies. Many of the Figures and Tables for this Chapter are presented in Appendix G.
- **Chapter 17 – “Analysis of Waste and Other Feedstocks”:** A large number of waste materials and samples other than Miscanthus, peat, and bagasse were also collected, processed, and analysed. This Chapter discusses the results obtained and also presents the projected yields associated with processing these in biorefining technologies. Feasible national scenarios based on processing the most suitable feedstocks are also presented. Many of the Figures and Tables for this Chapter are presented in Appendix H. Chapter 17 is an update to many parts of an earlier paper written by the Author (Hayes and Hayes, 2009). That paper used secondary analytical data to determine biorefinery yields from waste materials.
- **Chapter 18 – “The DIBANET Project”:** DIBANET is a large on-going research project funded by the EU 7th Framework Programme and co-ordinated by the University of Limerick. The Author was the main person responsible for the development of this project. This Chapter summarises the objectives of the project and how the NIRS work detailed in previous Chapters relates to these. In particular, the use of UL NIRS models on the spectra of Brazilian sugarcane bagasse and sugarcane “trash” (field harvesting residues) is discussed as is the development of quantitative NIRS models for the prediction of the lignocellulosic constituents of pretreated Miscanthus samples. Many of the Figures and Tables for this Chapter are presented in Appendix I.
- **Chapter 19 – “Conclusion”**

2 Chemistry of Lignocellulosics

Second generation biorefining technologies (Hayes, 2008) predominately operate on lignocellulosic feedstocks since these tend to be the cheapest, most abundant, and most productive resources available. The main constituents of these materials are the polysaccharides cellulose and hemicellulose and the biopolymer lignin. There will also be lesser amounts of ash and “extractive” components. This chapter will describe these and some other relevant chemical constituents. It will also mention in places their relevance to biorefining technologies, although the main dynamics of biorefining technologies will be outlined in Appendix B.

2.1 Carbohydrates

Carbohydrates are the primary constituents of most lignocellulosic materials. These are polyhydroxy compounds with a general elemental composition of $(\text{CH}_2\text{O})_n$, which gives a generally uniform carbon content of approximately 40% (Roberts, 1996). This is less than that of hydrocarbons, but the oxygenated nature of carbohydrates affords superior chemical properties for conversion and utilisation, and explains why they are so dominant in biota (Fan et al., 1987).

Carbohydrates may be classified into three groups:

- Monosaccharides – simple sugars such as glucose and xylose.
- Disaccharides and Oligosaccharides – two (disaccharides) or up to ten (oligosaccharides) monosaccharide residues joined together by glycosidic linkages.
- Polysaccharides – generally considered as polymers of more than ten monosaccharide residues.

The monosaccharides, formed as early products of photosynthesis from CO_2 and water, are the building blocks of all other plant carbohydrates. They are rarely present as free entities in plants; instead they occur as units in oligosaccharides and polysaccharides - which store reserve food for the plant (e.g. starch) or provide support for the cell wall (e.g. cellulose and hemicellulose).

2.1.1 Monosaccharides

Monosaccharides are either polyhydroxyaldehydes, which are also known as aldoses and are monosaccharides with an aldehyde group, or polyhydroxyketones, which are also known as ketoses and are monosaccharides with a ketone group. D-glucose is a common aldose and D-fructose is a common ketose. These are drawn in the Fischer projection in Figure 2-1:

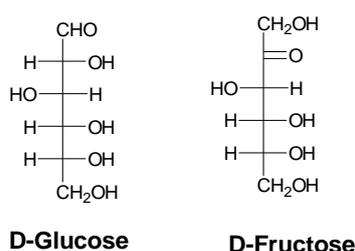


Figure 2-1: The Fischer projection of the aldose D-glucose and the ketose D-fructose

Sugars are classified according to the number of carbon atoms in the monosaccharide: for example trioses (3 atoms), tetroses (4), pentoses (5), hexoses (6), and heptoses (7). The major carbohydrates in nature are aldopentoses and aldohexoses.

The prefix deoxy- is used when one of the hydroxyl groups is replaced by a hydrogen atom. Ether groups tend to have the methyl functionality ($\text{C-OH} \rightarrow \text{C-OMe}$, denoted by the prefix O-methyl). The prefix for an ester group such as acetate ($\text{C-OH} \rightarrow \text{C-OCOCH}_3$) is O-acetyl (Sjostrom, 1981).

2.1.1.1 Configurations of Monosaccharides

A mixture containing an aldehyde and alcohol will be in equilibrium with the corresponding hemiacetal (Davis and Fairbanks, 2002). Carbohydrates, having both carbonyl and free hydroxyl functions are capable of forming intramolecular hemiacetals, ring-type structures that are termed lactols. The equilibrium is highly skewed towards the lactols with over 90% of the monosaccharides being in the cyclic form at any one time (Bungay, 1981). Formation of polysaccharides destroys the equilibrium and locks the sugars into rings.

Given that most naturally occurring monosaccharides have more than one hydroxyl group, rings of various sizes are possible. Monosaccharides with five membered rings are given the suffix furanose while those with six membered rings are termed pyranose.

With hexoses, the pyranose structure is always preferred. Pentoses also tend towards the pyranose form, although ribose has approximately 20% of the furanose form at equilibrium (Davis and Fairbanks, 2002). While the quantities of the ring types in solution are based on least energies, in plant polysaccharides the conformation of the monosaccharide substituents depends on the nature of the polymer they constitute.

Five-membered rings tend to be planar, hence these have significant strain. The 6-membered rings are puckered and assume different conformations. Two strainless forms are possible, the rigid chair or the flexible form. There are several possible shapes for the flexible form although only the boat and skew boat (or twist) forms are regular. The boat and chair forms are illustrated in Figure 2-2. In the chair form, C-2, C-3, C-5 and the oxygen of the lactol ring are on the same plane, with C-4 above and C-1 below. In the boat form, C-1, C-2, C-4 and C-5 are on the same plane, with the lactol oxygen and C-3 both above. All bonds in the chair conformation are staggered whereas, in the boat form, the bonds attached to C-2 and C-3 are eclipsed resulting in a higher potential energy through the repulsion between groups attached to C-2 and C-4 (Rees, 1967).

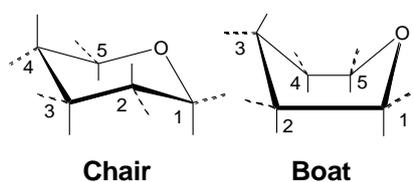


Figure 2-2: The chair and boat forms for hexoses. The vertical (full) lines represent axial bonds, the horizontal (dotted) lines represent equatorial bonds.

In the chair and boat conformations, those bonds that progress vertically from the ring are called axial and those that progress to the side (outwards) from the ring are termed equatorial (or radial). Hydrogen atoms are so small that they do not sterically interfere with each other in axial positions, but larger atoms and groups repel each other in all axial positions on the same side of the ring. Hence, with regard to hydroxyl orientations and mono- and multiple-substituted molecules, axial bonds tend to be more sterically cramped than equatorial bonds, which tend to be energetically favoured. There are certain exceptions, for example axial hydroxyl groups tend to favour intramolecular hydrogen bonds with the ring oxygen.

When sugars twist to form a ring, the C-1 goes from being planar in the aldehyde to tetrahedral and, hence, a new stereogenic centre is formed. This gives rise to two anomers termed α or β depending on whether the C-1 substituent is cis (α) or trans (β) to the oxygen atom at the highest numbered stereogenic centre. The general rule for sugars is that they are designated with an α - prefix if the hydroxy substituent group is axial and a β -prefix if it is equatorial. The properties of monosaccharides (and particularly the polysaccharides they constitute) are vastly different according to the orientation of the hydroxyl group (often termed the glycosidic hydroxyl) at the anomeric centre.

2.1.1.2 Important Aldoses

Glucose is the most abundant of the monosaccharides in nature, and it occurs predominately as a constituent of the homopolysaccharides cellulose (Section 2.1.3.2) and starch (Section 2.1.3.1). It is also an important precursor of other monosaccharides and carbohydrates. For example, the CH_2OH group of hexoses can be replaced with COOH (e.g. glucuronic acid), CH_3 (e.g. fucose and rhamnose), or H (e.g. xylose).

Other important aldohexoses in biomass are mannose and galactose – both present in certain hemicellulosic polysaccharides (Section 2.1.3.3).

The most common aldopentoses are xylose and arabinose.

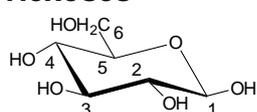
2.1.1.3 Important Ketoses and Other Monosaccharides

Ketoses in the open chain form have a primary alcohol group at both ends and a ketone (carbonyl) functionality within the chain. These are much less common than aldose groups (Davis and Fairbanks, 2002). The most common ketose is D-fructose.

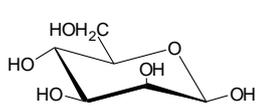
Other monosaccharides of relevance tend to be derivatives of the aldoses and ketoses. These include the deoxysugars. Deoxy D- sugars generally tend to occupy a very small fraction of total plant biomass (Davis and Fairbanks, 2002). There are some deoxy L-monosaccharides of relevance. L-rhamnose (6-deoxy-L-mannose) occurs in some hemicelluloses (Sjostrom, 1981) and in peats

(Section 13) and it is also a major constituent of the pectins (Section 2.1.3.4). The deoxysugar L-fucose is derived from D-galactose and can also occur in the D- form but it is only present in very minor amounts in most plants.

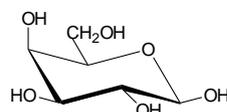
Hexoses



β -D-glucopyranose

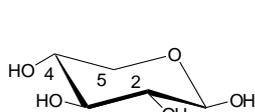


β -D-mannopyranose

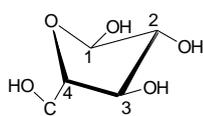


β -D-galactopyranose

Pentoses

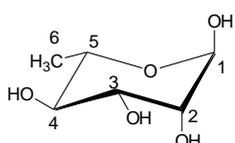


β -D-xylopyranose

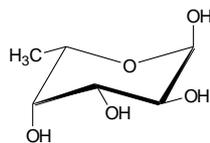


α -L-arabinofuranose

6-Deoxy-hexoses

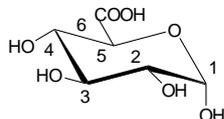


α -L-rhamnopyranose

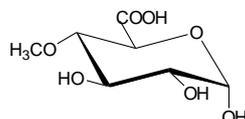


α -L-fucopyranose

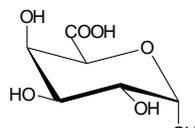
Uronic Acids



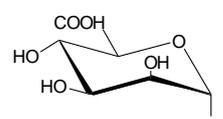
α -D-glucopyranosuronic acid



4-O-methyl- α -D-glucopyranosuronic acid



α -D-galactopyranosuronic acid



α -D-mannopyranosuronic acid

Figure 2-3: Some important monosaccharides of relevance to lignocellulosic materials, in the configurations and conformations that occur in nature.

Uronic acids are obtained from sugars when the primary alcohol group is oxidized to a carboxylic acid (Bungay, 1981). Galacturonic, mannuronic, and particularly glucuronic, acids (Figure 2-3) are important components of acidic polysaccharides such as glucuronoxylan and the pectins.

The quantity of uronic acids can be important in some biorefining technologies since they can act as fermentation inhibitors when some yeasts and other microorganisms try to ferment the non-acidic sugars in hydrolysates (Ranatunga et al., 2000). Uronic acids are also believed to reduce the accessibility of hemicelluloses for enzymatic hydrolysis (Carvalho et al., 2008). Work has been undertaken, however, in order to try and utilise uronic acids; it has been shown that a strain of *E. coli* can ferment the uronic acids in the hydrolysate to produce acetic acid and ethanol (Lawford and

Rousseau, 1997). In a modelling study, Mao *et al.* (2008) assumed an efficiency of 90% of the theoretical maximum for this conversion. It has also been suggested that the uronic acids may contribute to the formation of hydronium ions in the autohydrolysis of biomass (using compressed hot water) but their role here is still not completely understood (Gírio *et al.*, 2010).

2.1.1.4 Monosaccharide Derivatives

Glycosides:

The anomeric carbon at C-1 has different properties from the other carbon atoms in the ring and tends to be more reactive; a property attributable to its unique involvement in two (as opposed to one) carbon-oxygen bonds. A special property of this anomeric carbon is the way in which the hydroxyl can often preferentially be replaced by groups of the type OR', where R' can have a variety of alternative structures. The resulting product is termed a glycoside. The group derived from the hydroxyl (R') is termed the aglycone (Sjostrom, 1981). Pyranosides and furanosides are both possible, each possessing two anomeric forms (the α and β glycosides). In general the axial form has lower energy and is more stable than the equatorial form (Rees, 1967).

Glycosides involving an -O- bridge between the C-1 of one sugar unit and any non-anomeric carbon of another sugar, with the loss of one water molecule, are common in nature and the iterative procedure represents the formation of many polysaccharides. The formation of glycosides results in a preservation of the ring structure since glycosides are not easily isomerised in water. Glycosides are relatively easily hydrolysed by acids to sugars and alcohols but are usually fairly stable toward alkali.

Other Ethers and Esters:

A non-glycosidic ether can be formed through the replacement of any non-anomeric hydroxyl group(s). They often have different properties from glycosides and are usually retained during acid hydrolysis of glycosidic linkages.

2.1.2 Disaccharides and Oligosaccharides

Disaccharides contain two monosaccharide units bound by a glycosidic linkage. Sucrose, a disaccharide of α -D-glucopyranose and β -D-fructofuranose, and shown in Figure 2-4 (iii), is the most important disaccharide in plants, occupying much of the mass balance in species such as sugarcane, sugar beets, and sweet sorghum (Bungay, 1981).

Other important disaccharides include the dimers cellobiose and maltose. These are the repeating units of cellulose (Section 2.1.3.2) and starch (Section 2.1.3.1), respectively, and can be obtained upon their partial hydrolysis. In maltose, Figure 2-4 (ii), the monosaccharides are linked through an α -(1 \rightarrow 4) linkage, while in cellobiose, Figure 2-4 (i), the linkage is β -(1 \rightarrow 4).

Oligosaccharides have between three and ten monosaccharides (Sjostrom, 1981). Due to their low chain length, oligosaccharides are usually easily hydrolysed to their monomeric components and are often soluble in water. They are therefore important constituents of the extractives fraction (Section 2.4).

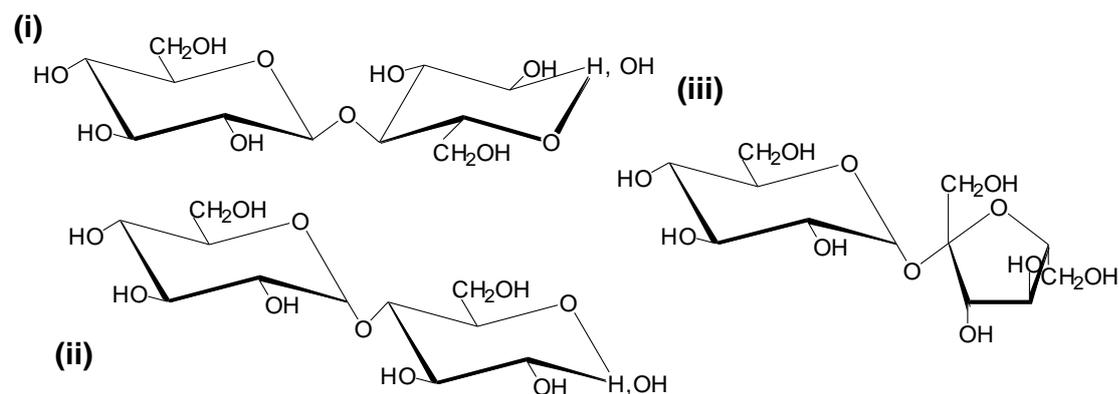


Figure 2-4: Some disaccharides. (i) Cellobiose [4-O-(β -D-glucopyranosyl)-D-glucopyranose]; (ii) Maltose [4-O-(α -D-glucopyranosyl)-D-glucopyranose]; (iii) Sucrose [α -D-glucopyranosyl β -D-fructofuranoside]. Note, H, OH means that the carbon substituents can be either axial or equatorial.

2.1.3 Polysaccharides

The vast majority of sugars in most biomass systems exist as constituents of polysaccharides. Homopolysaccharides are those that contain a single repeating sugar and are generally named according to the respective monosaccharide; for example glucans, mannans, xylans, galactans and arabinans. Heteropolysaccharides contain more than one sugar type in the chain.

2.1.3.1 Starch

Starch is the simplest of the glucans. It has an important sugar-storage function, as opposed to structural support, in some plants (Bungay, 1981). It is a major constituent of potatoes, maize, wheat, and oats, among other plants. Starch is less prevalent in the cases of the biomass feedstocks generally considered economically feasible for lignocellulosic biorefining technologies. However, it can be a significant mass constituent in wastes from the industrial sector and in some grass and food wastes.

Starch is a mixture of two polysaccharides: amylose and amylopectin. Amylose, Figure 2-5 (a), is the polysaccharide equivalent of maltose (Section 2.1.2) and consists of α -(1 \rightarrow 4)-linked D-glucopyranose units (Sjostrom, 1981). The axial nature of the glycosidic linkage reduces the strength and abundance of intermolecular hydrogen bonds between amylose molecules when compared with the β -linked cellulose homopolysaccharide (Section 2.1.3.2). Amylose has an approximate degree of polymerisation (DP - number of monosaccharide units) of 2000 and the polysaccharide forms a helix with six glucose units in each turn (Rees, 1967).

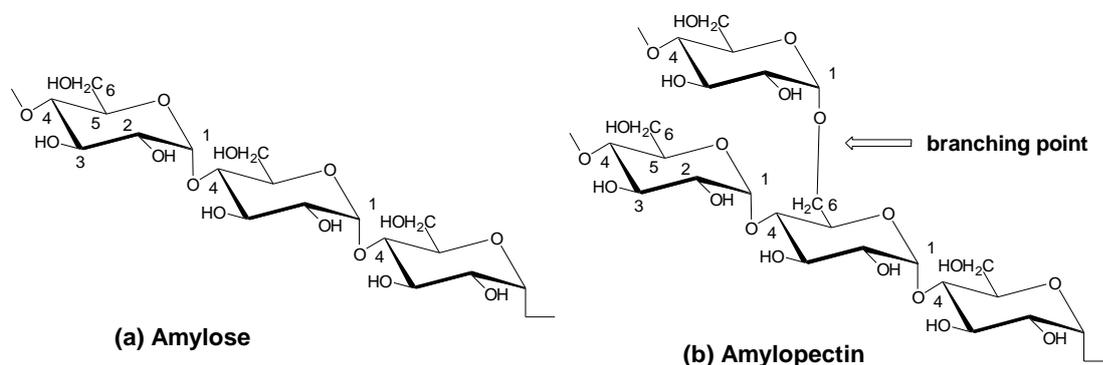


Figure 2-5: Amylose and amylopectin.

Amylose is generally a minor component of most starches with amylopectin, Figure 2-5 (b), being the major constituent. Amylopectin also contains glucose units linked via α -(1 \rightarrow 4) bonds; however, there are also α -(1 \rightarrow 6) branches that occur every 24 to 30 glucose units in plants. Amylopectin has a much higher DP than amylose. For example, the amylopectin starch of potatoes has been found to have a DP of approximately 200,000 residues (Marchal et al., 2001). It cannot coil into a long helix, however, due to the extensive branching and the fact that the α -(1 \rightarrow 4) chains that do exist before branching are too short. This means that no compact intermolecular alignment, and hence no

significant hydrogen bonding, can take place. This accounts for its extensive solubility, especially when compared with cellulose.

2.1.3.2 Cellulose

While starch is an important mass constituent of some plants, for most it is present only in small amounts as a short-lived intermediate in the vegetative part of plants. Cellulose, on the other hand, is the major mass constituent of most naturally occurring biomass. It is in fact the most abundant biogenic polymer with estimates of $3.24 \times 10^{11} \text{ m}^3$ available globally (Fan et al., 1987). Any biomass strategy that focuses on the chemical utilisation of a wide variety of feedstocks will therefore need to be effective in its handling of this polysaccharide.

As with amylose, D-anhydro-glucopyranose units are linked through (1 \rightarrow 4)-glycosidic bonds. Unlike amylose, the bonds are β -linked, as with the disaccharide cellobiose (Section 2.1.2). This structure, and the associated intra- and inter-molecular bonds, is illustrated in Figure 2-6. The chemical and physical behaviour of cellulose differs completely from that of starch.

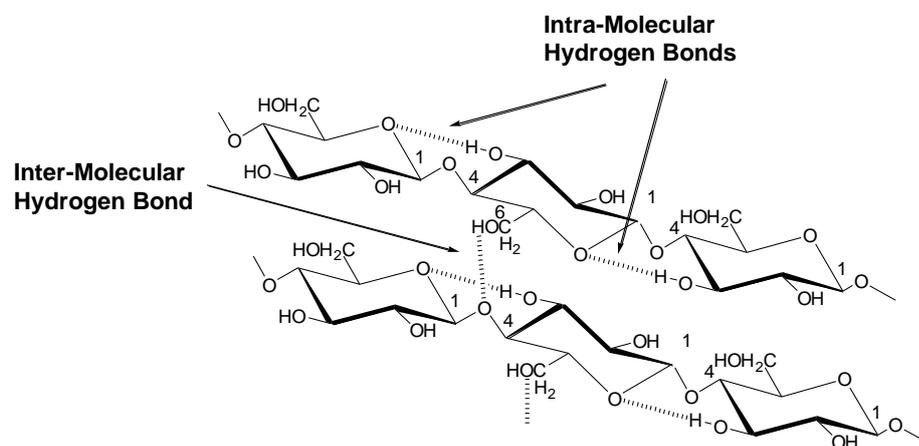


Figure 2-6: The β -linked glucopyranoside residues and the inter- and intra-molecular hydrogen bonding of cellulose.

As shown in Figure 2-6 the intermolecular hydrogen bonds in cellulose exist between the O-C-6 of one glucose residue in one chain and the O of the glycosidic linkage of another chain (Figure 2.12). As with cellobiose, cellulose also has the intramolecular hydrogen bond between C-3-H and the ring-oxygen. The chain of β -linked cellulose is very differently shaped compared with that of the α -linked

amylose. Amylose is a wide hollow tube, rather like a wire spring, whereas cellulose is a flat narrow ribbon. This narrow nature therefore allows more intimate intermolecular associations (Rees, 1967).

Cellulose has a broad molecular weight distribution, the number of β -D-glucopyranose residues in a cellulose molecule varies according to the feedstock and environmental conditions but, for example, can range from 1000 for newsprint to 10,000 for cotton (Bungay, 1981).

Aggregate Structure and Crystallinity of Cellulose:

The grouping of cellulose molecules into tightly bound aggregates through hydrogen-bonds is considered to result in the formation of microfibrils – long thin threads of cellulose aggregates that act in the same way as reinforcing rods in pre-stressed concrete. Microfibril size can vary from 'elementary fibrils', with approximately 36 chains, to the large microfibrils of cellulosic algae, with more than 1,200 chains (Sugiyama, 1985). In forages, it is considered that 60 to 70 cellulose chains constitute a microfibril (Hatfield, 1993). Microfibrils build up fibrils and ultimately cellulose fibres, Figure 2-7.

Microfibrils are said to contain two different regions. The crystalline region consists of highly ordered cellulose molecules while the molecules in the amorphous (or para-crystalline) region are less highly ordered. There are uncertainties concerning the exact nature and configuration of such regions, however. The “fringe micellar model” (Sjostrom, 1981) describes crystalline regions, which, without any distinctive boundary, change into amorphous regions.

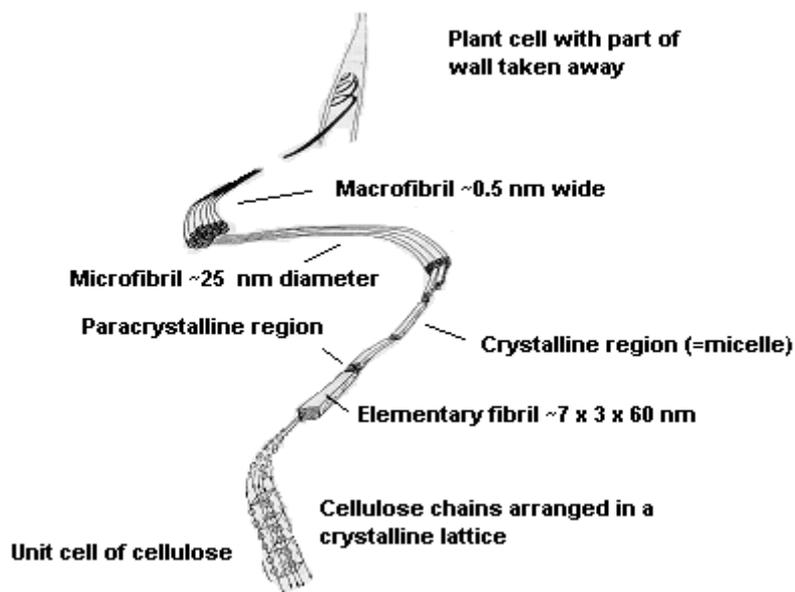


Figure 2-7: A representation of a possible molecular architecture of the cellulose molecule, showing its relationship to the microfibrils and to the total cell wall. Adapted from Roberts (1996).

Other models, such as that proposed by Wickholm (2001) after NMR studies, and illustrated in Figure 2-8, describe the crystalline region as being present only in the inner core of the microfibril and with a non-crystalline ordered structure (paracrystalline cellulose) between the core and the surface. Accessible and inaccessible surfaces (due to microfibril aggregation) also exist in the structure.

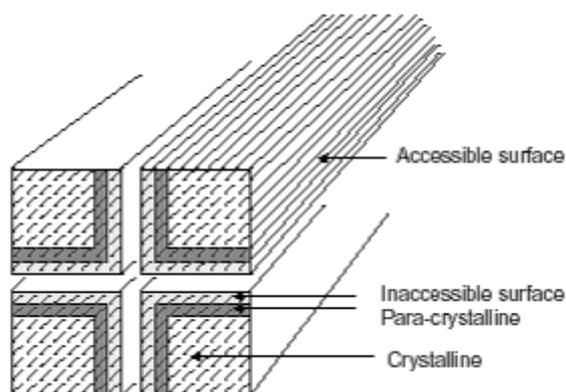


Figure 2-8: A proposed representation, based on NMR studies, of the fibril structure. Taken from (Wickholm, 2001).

The crystallinity of cellulose is known to vary depending on the origin of the sample - cotton cellulose is more crystalline than cellulose in wood, for example (Sjostrom, 1981). It is also likely to vary between the primary and secondary wall.

The extensive hydrogen bonding and compact, tightly-bound aggregate structure of cellulose are factors that contribute to its strength, fibrous character, and insolubility. Its resistance to hydrolysis is high compared to most biogenic polysaccharides, and starch in particular. While starch is an important source of energy in the human diet, cellulose cannot be broken down and passes through the system. Ruminants are able to use cellulose but only after it has been broken down to sugars by micro-organisms that live in the alimentary tract (Rees, 1967). Hydrolysing cellulose without excessive yield losses in glucose and without significant energy input is a crucial requirement for lignocellulosic biorefining technologies. The amorphous region tends to be easily hydrolysable by acids while the crystalline cellulose is more resistant (Bungay, 1981). The result is that hydrolysis with a dilute acid tends to remove the amorphous regions from cellulose leaving a polysaccharide of increased crystallinity that is resistant to further hydrolysis. Concentrated acid hydrolysis (as

employed in the analytical methodology outlined in Section 11.5) should hydrolyse cellulose completely, however.

2.1.3.3 Hemicelluloses

The term hemicellulose covers a variety of complex carbohydrate polymers that are mostly not extractable by hot water but, unlike cellulose, are extractable in aqueous alkali (Dehority, 1993). They constitute the cell wall polysaccharides of land plants that are not cellulose or pectins (Casey, 1980). Hemicelluloses tend to be branched heteropolysaccharides that are mostly built up of: the pentoses D-xylose and L-arabinose; the hexoses D-glucose, D-mannose and D-galactose; with smaller amounts of L-rhamnose, in addition to D-glucuronic acid, 4-O-methyl-D-glucuronic acid, and D-galacturonic acid. While the proportions of these substituents vary between hemicellulose and feedstock types, the majority tend to be pentoses (Bungay, 1981).

There are key differences between cellulose and hemicelluloses. These include the absence of a highly ordered state in hemicelluloses, and that their molecular weight is much lower – their DP is only 100 to 200 units (Goldstein, 1991). Such factors help explain their comparatively easy hydrolysis.

Based on composition, plant cell wall hemicelluloses can be divided into 3 groups (Aspinall, 1981):

- xylans
- mannans
- glucans - the non-homopolysaccharide types of relevance to hemicellulose being the xyloglucans

There are also the arabinogalactans which some classify as extractives rather than hemicelluloses. The general structures of these types will be discussed below.

Xylans:

These tend to be the most abundant hemicelluloses in nature and involve a backbone of β -(1 \rightarrow 4)-linked D-xylopyranose residues (Bungay, 1981). In its absolute form, a xylan would be a homopolysaccharide able to adopt a conformation similar to cellulose. The pure, unsubstituted, conformation would differ somewhat, however, due to the lack of the large CH₂OH unit on C5. This would mean that the Van der Waals repulsion involving this group, which prevents certain rotations

of the glycosidic linkage in cellulose and cellobiose, would be absent. The resulting conformation would, like cellulose, allow intra- and inter-molecular hydrogen bonding. In most plants, however, there are few occurrences of xylans that are solely made up of anhydro-xylose units.

Most xylans found in land plants contain short side-chains consisting of one or two sugar units linked to the main backbone of the anhydro-xylose units. These additional units vary between species and can include arabinose, glucuronic acid, mannose as well as other pentoses and hexoses.

The existence of side-chains, and possible occasional branching, has a similar effect on the properties of xylans as it does on that of amylopectin – the linear ribbon-like conformations required for tight intramolecular bonding are precluded resulting in a disruption of the interaction of backbone residues and aggregation (Hatfield, 1993). The net effect of this is that hemicellulose is up to 1000 times easier to hydrolyse than cellulose (DiPardo, 2000).

A significant portion of the xylose residues in higher plants are acetylated, mainly on the hydroxyl groups on C-2, but also on C-3 (Timell, 1965). The ratio of acetyl residues to xylose units depends on the species, with no acetyl groups in softwood xylans for example (Sjostrom, 1981).

Glucuronoxylans, acetyl-4-methyl-glucopyranosylurono-xylan [(4-*O*-methylglucurono)xylan], are abundant, particularly in hardwoods and grasses (Bungay, 1981). These are sometimes simply referred to as xylans. Branch points consisting of the pyranose forms of 4-*O*-methyl-*D*-glucuronic acid are attached by an alpha linkage mainly to the C-2 position of the xylose unit with the ratio of xylose to uronic acid varying among species (Casey, 1980). The xylosidic bonds between the xylose units are easily hydrolysed by acids, whereas the linkages between the uronic acid groups and xylose are very resistant. Acetyl groups are easily cleaved by alkali.

Short sidechains of arabinofuranosyl and glucuronosyl residues (arabinoglucuronoxylans) are also possible in xylans (Wilkie, 1979).

Mannans:

An unsubstituted mannan usually constitutes a linear chain of β -(1 \rightarrow 4)-linked *D*-mannopyranose residues. Such pure mannans are relatively uncommon in higher plants, however, being more important in microorganisms such as yeasts (McMurrough and Rose, 1967).

The mannans of most relevance are galactoglucomannans. These possess a slightly branched backbone of (1 \rightarrow 4)-linked β -*D*-glucopyranose and β -*D*-mannopyranose units. These are

complemented, to varying degrees, by α -D-galactopyranose residues linked as a single-unit to the framework by (1 \rightarrow 6)-bonds.

There are two distinct types of galactoglucomannans. In one, often termed the glucomannans, the ratio galactose:glucose:mannose is about 0.1:1:4. This is generally a water-soluble polysaccharide (Casey, 1980). The other galactoglucomannan type is insoluble and much more galactose-rich, with a corresponding ratio of 1:1:3 (Sjostrom, 1981).

Galactoglucomannans are easily depolymerized by acids, and especially so the bond between galactose and the main chain. The acetyl groups are much more easily cleaved by alkali than by acid (Aman, 1993).

Xyloglucans:

These are a type of heteropolysaccharide glucans. They are hemicellulosic polysaccharides that are present in the primary cell walls of all higher plants (Doco et al., 2003). These are characterised by the β -(1 \rightarrow 4)-glucosyl main chain of cellulose (Aspinall, 1981) with sidechains usually substituted at C-6 (Hayashi, 1989).

Arabinogalactan:

Arabinogalactan is only present in significant amounts in a few species, such as the *Larix* (larch) genus and it is a minor component in most biomass (Hakkila, 1989). It is based on a backbone of (1 \rightarrow 3)-linked- β -D-galactopyranose units. Almost every unit carries a branch attached to position C-6. The exact nature of all side chains is not known but most are composed of (1 \rightarrow 6)-linked β -D-galactopyranose residues and have an average length of two such units (Casey, 1980).

2.1.3.4 Pectins

Acidic structural polysaccharides that are extracted with hot water are often referred to as pectins. Pectic polysaccharides are present in the primary cell walls of all seed-bearing plants (Bacic et al., 1988). These are most often found in significant quantities in fruits and vegetables and are less common in most feedstocks that are considered to be viable for commercial lignocellulosic

biorefining. However, some wastes from the food industry, apple pomace for example, can have significant pectin concentrations and may be viable feedstocks.

Pectins are polyuronides and are considered to be the most highly branched polysaccharides (Stombaugh et al., 2000). The backbone of the pectin structure tends to consist of partially methylated α -(1 \rightarrow 4)-D-galacturonic acid residues. However, there are areas of alternating α -(1 \rightarrow 2)-L-rhamnosyl- α (1 \rightarrow 4)-D-galacturonosyl disaccharide sections (Aman, 1993). While pectins are acidic polysaccharides, they do not contain glucuronic acid and 4-methyl-glucuronic acid, which only exist in hemicellulose (Van Soest, 1993).

2.2 Lignin

Lignin is a structurally important polymer in biomass. Its formation is unique to vascular plants; primitive plants such as fungi, algae, and mosses, do not contain lignin (Casey, 1980). It is said to function primarily as a supporting agent in cell structure and it also assists in the resistance of biomass against microbial attack and decay (Roberts, 1996).

Structurally, lignin can be described as a complex three-dimensional polymer of phenylpropane units. It is a much smaller molecule than cellulose, with only approximately 25 aromatic rings per polymer. The phenylpropane units are mostly either 4-hydroxycinnamyl alcohol (para-coumaryl alcohol, H) or its 3- and/or 3,5-methoxylated derivatives - coniferyl (guaiacyl, G) alcohol, and sinapyl alcohol (or syringyl, S), respectively (Figure 2-9). The ratio of these units varies between plants; for example in hardwoods S and G forms dominate, whereas softwood lignins contain only G units (Theander and Westerlund, 1993).

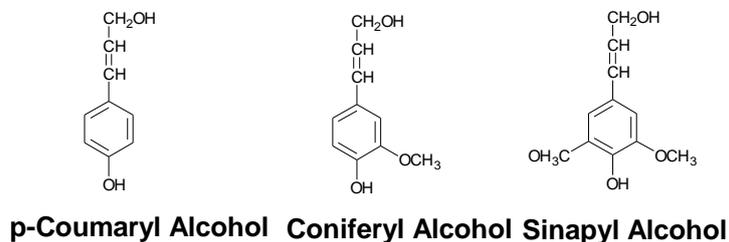


Figure 2-9: The phenylpropane units that form the structural basis of the lignin polymer

The phenylpropane units are relatives of carbohydrates, coming from the dehydration and cyclisation of sugars (Bungay, 1981). Phenylpropane units in lignin are said to be linked in various

ways - directly between the rings, between the propane units, and through ether linkages via the hydroxyl groups (Klass, 1998). Ether linkages between aromatic rings are possible at several positions; thus a three-dimensional structure results (Roberts, 1996). These ether linkages are very resistant to cleavage, a factor in explaining the low lignin degradation rates by most biota (Bungay, 1981). Lignin is also relatively hydrophobic. Figure 2-10 shows the complexity of the lignin macromolecule by illustrating the model of spruce lignin.

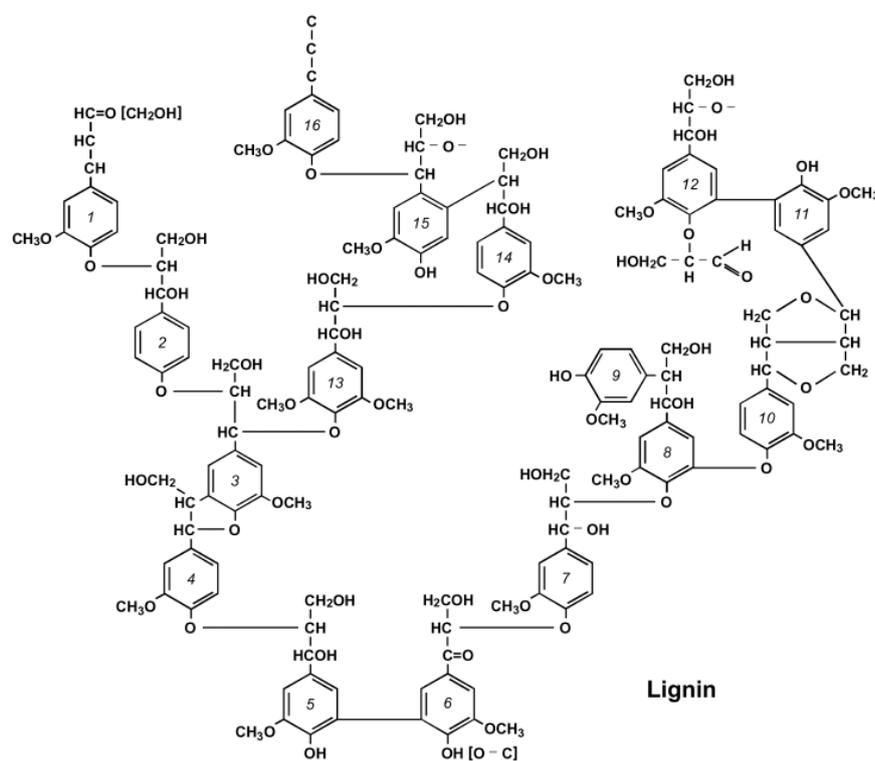


Figure 2-10: A model of spruce lignin (Alder, 1977)

It has been shown that a high lignin content is correlated with the recalcitrance of polysaccharides to enzymatic hydrolysis and that increasing the relative proportion of the S monomer may aid in biofuel conversion mechanisms (Li et al., 2010, Chen and Dixon, 2007). An earlier paper by the Author (Hayes, 2008) mentions more on the influence of lignin on biorefining technologies

2.3 Associations and Development of Lignocellulosic Components

This section details the configuration of the structural polysaccharides and lignin in the cells of plants. Cells can be categorised according to their shapes and functions. Figure 2-11 shows the cell structure of a softwood tracheid cell. Tracheids are cells that exist in the xylem (a type of transport tissue) of vascular plants (plants that have lignified tissues for conducting water, among other things, through the plant). The general cell structure shown in Figure 2-11 is also of relevance to hardwoods and grasses.

The middle lamella (ML in Figure 2-11) is the amorphous region between the cells and serves the function of binding the cells together. In the early stages of growth it tends to be mainly composed of pectic substances, but eventually it becomes highly lignified. It is about 1-2 microns thick, amorphous, and generally porous (Theander and Westerlund, 1993).

The primary wall is a thin layer (0.1-0.2 μm thick) that consists of cellulose, hemicelluloses, pectin and protein and is completely embedded in lignin. Its lignin content is high, but because this layer is thin, only 20-25% of the total lignin in wood is located in this layer (Roberts, 1996). Sometimes the term compound-middle-lamella is used for the primary wall combined with the middle lamella.

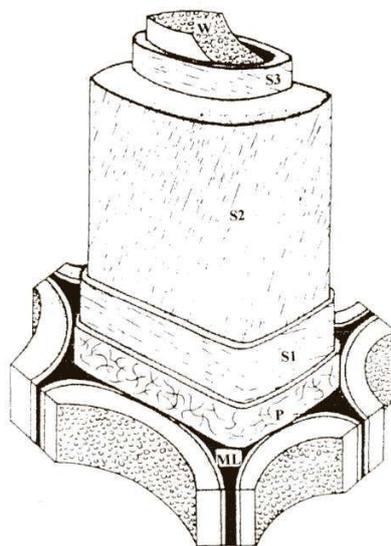


Figure 2-11: A schematic representation of the primary (P) and secondary (S1, S2 and S3) cell walls of a softwood tracheid. ML = middle lamella. Taken from (Roberts, 1996)

The secondary wall is immediately below the primary wall, it comprises nearly all of the cell wall. It is divided into three layers that are distinguished ultrastructurally by their different orientation of

cellulose microfibrils (Harris, 1990). The outer layer (S1) and inner layer (S3) tend to be relatively small (approximately 0.1-0.3 μm and 0.1 μm thick, respectively). The middle layer of the secondary wall is much thicker and constitutes the main portion of the cell wall (Roberts, 1996).

The tertiary wall is a very thin membranous layer after the secondary wall. A special modification of this is the warty layer, located in the inner surface of the cell wall in all conifers and some hardwoods; this amorphous layer contains watery deposits together with lignin precursors that form globules (Wilson, 1993). Each species has its own characteristic warty layer.

Cell Development:

The structure of a cell can be related to the way in which it develops (Sjostrom, 1981). Upon cell division, a pectin-rich cell-plate is developed. Each of the two new cells subsequently encloses itself with the thin, extensible primary wall consisting of cellulose, hemicelluloses, pectin and protein. During the following phase of differentiation, the cell first expands to its full final size, after which formation of the thick, secondary wall is initiated. At this stage, the wall consists of cellulose and hemicelluloses. Lignification begins when the secondary wall is still being formed.

Cellulose-Hemicellulose Associations:

The association of hemicelluloses with cellulose is a complex issue which is still not fully documented. It is generally considered that there are no chemical bonds between the polysaccharides, with interactions instead based on intermolecular hydrogen bonds and Van der Waal's forces (Sjostrom, 1981). In order for such attractions to occur, the conformations of both molecules would need to be similar so that close, quasi-parallel, alignment could be attained.

With respect to the xyloglucans, tight associations with cellulose that require strong alkali for dissociation are also often found (Hayashi, 1989). Given that the backbone of this polysaccharide is very similar to that of cellulose, such associations are likely to occur where this backbone is open and unsubstituted. The association between the polysaccharides is said to form a cross-linked network in the primary wall that may be important for the structural integrity of the walls (Carpita and Gibeaut, 1993, Pauly et al., 1999).

Lignin-Carbohydrate Bonds:

The complex associations of lignin with carbohydrate polysaccharides is of particular importance for many lignocellulosic fractionating technologies. Researchers have postulated the effects of cross-linking on the properties of plant cell walls, such as accessibility and degradability (Ralph and Helm,

1993). The resistance of polysaccharide-lignin bonds to hydrolysis not only depends on the type of linkage involved but also on the types of sugars associated with the linkage, and the chemical structure of the lignin unit attached to the sugar (Joniak and Kosikova, 1976).

Experimental observations suggest that there are at least two major bonds between lignin and carbohydrates. One type is that based on adsorption and chemisorption (predominately hydrogen bonds and van der Waals forces), and the other based on covalent bonds (Casey, 1980). Ester-, ether-, phenyl-glycoside and acetyl bonds are all possible covalent linkages (Grushnikov and Shorygina, 1972).

Lignocellulose Macrostructure:

Cellulose is said to form a skeleton which is surrounded by other substances functioning as matrix (hemicelluloses) and encrusting (lignin) materials (Sjostrom, 1981). More specifically, disordered cellulose molecules, as well as hemicelluloses and lignin, are located in the spaces between microfibrils. The hemicelluloses are considered to be amorphous even though they are apparently orientated in the same direction as the cellulose microfibrils.

It is a combination of the crystalline nature of cellulose and the existence of a physical barrier of lignin surrounding the cellulose fibres that is said to be the principal reason why the cellulose monomer linkages are harder to break than those in amylose, rather than the equatorial versus axial nature of the glycosidic linkages (Bungay, 1981). Therefore, the primary linkage of the cellulose polymeric chain may not be as important in causing slow and incomplete hydrolysis as are secondary and tertiary structures of cellulosic materials.

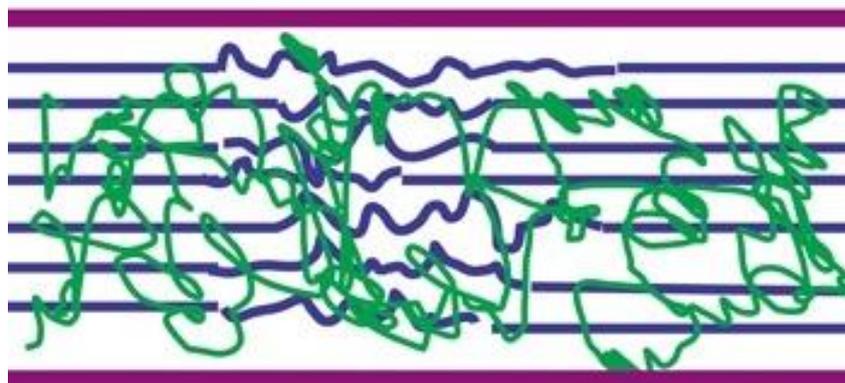


Figure 2-12: A basic theoretical representation of the lignocellulosic matrix, with crystalline cellulose regions (straight blue line), amorphous cellulose regions (wavy blue lines), hemicelluloses (green lines), and lignin (purple lines). Adapted from (accessed 25-3-11): http://genomics.energy.gov/gallery/biomass/view_np/view-09.html

Figure 2-12 shows a basic representation of the lignocellulosic matrix. There are crystalline regions (straight blue lines) and amorphous regions (wavy blue lines) of cellulose, cross linked through these are hemicelluloses (green lines), and lignin (purple lines) acting as a reinforcing structure.

2.4 Extractives

Extractives are defined as extraneous components that may be separated from the insoluble cell wall material by their solubility in water or neutral organic solvents. Solvents of different polarities are required to remove different types of extractives. Hence the extractives are often classified according to which solvent can extract them, for instance ethanol-soluble-extractives. These are of relevance to lignocellulose fractionation due to their effects on the dynamics of stored biomass piles and the accuracy of analysis procedures (see Section 3.4). Also, the water-soluble carbohydrate extractives represent an additional potential source of sugars. It has also been shown that extractives exert an influence on the mechanism of biomass pyrolysis, with biomass that has had its extractives removed releasing main products earlier in the thermogram (Guo et al., 2010).

There are a large number of different extractives, many of which are species-specific. These are often classified into categories according to the similarity of their chemical structures. Major categories include monosaccharides, polysaccharides, volatile oils, terpenes, fatty acids and their esters, waxes, polyhydric alcohols, alkaloids, and aromatic compounds (Goldstein, 1991).

Many extractives have roles in the metabolic processes of a plant. The primary metabolites are interconvertible biogenic intermediates and include monosaccharides, amino acids, simple fats and various carboxylic acids. The more complex secondary metabolites tend to be irreversibly formed. These include starch, simple terpenoids, chlorophyll, phenylpropanoids, the common flavonoids, and simple tannins (Goldstein, 1991).

Given the transient nature of many extractives, quantities vary greatly depending on the characteristics of the producing tissue and the influence of the environment. Given that photosynthesis takes place in the leaves, chemical synthesis and carbohydrate storage tends to be highest in these and the extractive concentrations tend to be greater in the foliage of grassy and woody biomass (Hakkila, 1989).

2.5 Protein

Proteins are polypeptides that are composed of units called amino acids. Amino acids are molecules that contain an amino group (NH_2), a carboxylic acid group, and a side chain. Two amino acids join by a peptide bond - a covalent chemical bond between the carboxyl group of one amino acid and the amino group of the other, resulting in a loss of a molecule of water in the process. This joining of two amino acids forms a dipeptide and the joining of many units in such a way will form the polypeptide protein. There are about 20 amino acids that are found regularly in naturally occurring proteins (Bungay, 1981). Given this number of different amino acids, and the large protein polypeptide chain length (up to ten thousand residues), there are a huge number of different potential protein configurations. Furthermore, other bonds are possible, in addition to peptide bonds, increasing the complexity of proteins (Bungay, 1981). Proteins can also be covalently linked to polysaccharides (Theander and Westerlund, 1993).

Protein is not a major component of most energy crops or agricultural residues, particularly at the stage of their development cycle when harvesting will occur. For that reason, most biorefining technologies that are focused on the conversion of lignocellulose do not attempt to extract proteins from the feedstock. Instead it will be likely that some of what protein there is may be incorporated into the solid residue of hydrolysis technologies. Regarding thermochemical technologies, nitrogen is volatile and will be mostly lost to the gas phase.

The content of protein in plants is often approximated from the nitrogen content by multiplying the mass of nitrogen by a factor (6.25 as described by Stormbaugha *et al.* (2000)). Nitrogen is an important nutrient for plants, and sustainable agricultural practices recommend that residues from the harvesting of biomass be retained, at least partially, on the land in order for these nutrients to be re-assimilated into the soil so that productivity levels on the stand can be maintained (Kohel-Knabner, 2002). Biorefining practices that may change this dynamic (see Section 16.5) should therefore be carefully examined in order to ascertain how much nitrogen could potentially be lost from the land by any new practice involving the removal of more biomass.

2.6 Ash

Ash in lignocellulosics is generally considered as the residue remaining after the material has been incinerated (Fengel and Wegener, 1984). Ash therefore has no energy value and, being made up of the inorganic elements in the biomass, is of no value in lignocellulosic fractionation technologies. High ash-contents can cause problems in pyrolysis reactions and in combustion schemes of biomass

or the residual char (Wiseloge et al., 1996). As far as acid hydrolysis procedures are concerned, an increased ash content may imply a higher consumption of acid due to the alkaline nature of some ash.

The ash content can vary greatly between plant species, and is generally higher in agricultural residues. The ash present in plants will depend on their stage of growth, the time of year and their location. The leaching of stored biomass may reduce the level of inorganics in some instances (Wiseloge et al., 1996).

The major cations present in ashes from lignocellulosic materials are Ca, K and Mg. Other elements such as Mn, Na and P are present in minor amounts. Trace constituents, such as Al, Fe, Zn, Cu, Ti, Pb, Ni, V, Co, Ag and Mo are also found in most substrates (Pohlandt et al., 1993, Naidenov et al., 1982). The anions that are usually present are chloride, carbonate, sulphate and silicate (Osman and Goss, 1983). With waste feedstocks (municipal solid wastes in particular), ashes are often more abundant and more diverse.

2.7 Moisture Content

The moisture content of biomass is a crucial efficiency parameter when biomass combustion is the main consideration, but less important in lignocellulose fractionation technologies. Its determination is also necessary in most carbohydrate analytical procedures (Section 3.3) and it has a huge influence on the NIR spectra of lignocellulosic materials (Section 5.3.5).

Water is generally held in biomass in two ways - either as a free liquid and vapour that is contained in the cell cavities, or as a molecule that is bound within the cell walls. The Fibre Saturation Point is the point in the drying process at which the first type of water is removed from the biomass but the second remains.

Moisture content tends to vary widely with biomass species, age, geographic locations and genetic differences. It also varies between different anatomical fractions of the same plant and throughout the year (Klass, 1981).

The moisture content of biomass can either be measured on a wet or dry basis. The wet-basis (M_{wb}) expresses the ratio of moisture mass to the total mass of the substance:

$$M_{wb} = \frac{M_{H_2O}}{M_{H_2O} + M_{dm}} \quad (2.1)$$

Where: M_{H_2O} = mass of moisture, M_{dm} = mass of dry matter

The dry-basis (M_{db}) expresses the ratio of the moisture mass to the mass of dry matter:

$$M_{db} = \frac{M_{H_2O}}{M_{dm}} \quad (2.2)$$

Unless stated otherwise, moisture contents in this Thesis will be presented on a wet-basis.

2.8 Heating Value

The heating value is currently one of the most important qualities of biomass given the predominance of combustion facilities over second-generation biorefining hydrolysis facilities.

There are several units of measurement. The caloric, or higher heating value (HHV), is independent of moisture content and reliant on the chemical composition of the material. A linear relationship exists between the heat of combustion and the carbon content of the substrate while oxygen, nitrogen and inorganic elements tend to reduce the value (Hakkila, 1989). Elemental analysis (see Section 3.6) can be used to show the proportions of C, H, O, N in a fuel. The HHV can be calculated from the ash, C, and H content by (Sheng and Azevedo, 2005):

$$\text{HHV (MJ/kg)} = -1.3675 + 0.3137C + 0.7009H + 0.0318 O^* \quad (2.3)$$

Where O^* = the sum of the contents of oxygen and other elements (including S, N, Cl, etc.) in the organic matter, i.e. $O^* = 100\% - C - H - \text{Ash}$.

Carbohydrates, having an elemental composition of $(CH_2O)_n$, will have a generally uniform carbon content of 40%, which is lower than many other biomass mass constituents such as lignin which has an average carbon content of approximately 60-65% (Roberts, 1996). This results in a lower HHV: Susott *et al.* (1975) recorded a mean caloric value for polysaccharides of 3853 cal/g; lower than the less-oxygenated lignin (5884 cal/g) and extractives (8124 cal/g for terpenoid hydrocarbons and 9027 cal/g for resins).

These data indicate that, for efficient utilisation of biomass, selective combustion of biomass constituents, if possible, could be profitable. This is particularly true given that, unlike cellulose and

hemicellulose, the complex nature of lignin generally precludes effective fractionation into value added products to date.

The Lower Heating Value (LHV), or effective heating value, is perhaps more relevant than the HHV in practical operations. It considers the energy required to vaporise the water generated when the hydrogen and oxygen elements of the biomass combine. Hydrogen content then becomes a reducing factor in the heating value. The LHV can be calculated on a dry basis from the equation below (Hakkila, 1989):

$$LHV = HHV - 0.22 * H \quad (2.4)$$

Where H = Hydrogen content of dry biomass (%)

For non-dry biomass a reduction in the LHV is necessary. The following formula is therefore used (Hakkila, 1989):

$$W_{em} = LHV - 2.45 \left(\frac{MC}{100 - MC} \right) \quad (2.5)$$

Where: W_{em} = Effective heating value of biomass with moisture, MJ/kg dry mass; and MC = Moisture content in the biomass on a wet mass basis (%).

3 Reference Analytical Methods

This section will provide a discussion of the reference analytical methods available for characterising the important chemical characteristics of lignocellulosic materials outlined in Section 2. These are the reference methods that the Author compared in order to determine the best analytical protocol to use. These methods need to be highly reproducible, accurate, and precise if they are to be suitable for the development of effective quantitative NIR calibration equations (Section 6). These should also be focussed towards the most important constituents regarding lignocellulosic second-generation biofuel technologies – principally the constituents of the structural polysaccharides. Hence, the following methods were evaluated with these criteria in mind.

3.1 Methods for Determining Structural Carbohydrates and Lignin

There are two main techniques that, over the last hundred years, have been employed in order to analyse the structural cell wall components of biomass. One of these are represented by variants of the Weende method (Section 3.1.1), a “detergent fibre” method that approximates for the polysaccharides and lignin gravimetrically. The second technique is represented by variants of the Klason lignin method (Section 3.1.2) which in most cases involve a two stage hydrolysis with, typically, sulphuric acid. This approach allows for the chromatographic separation of the constituent monosaccharide units of the hydrolysed polysaccharides although the acid insoluble lignin component is only measured gravimetrically.

3.1.1 Detergent Fibre Analysis

The crude fibre or Weende method was an early method for fibre determination (Hennenberg and Stohmann, 1859). This involved treating the biomass at an elevated temperature, first with 0.128 M sulphuric acid and then with 0.344 M sodium hydroxide after defatting with petroleum ether. There have been severe limitations with this method and many modifications were made (Thomas, 1972, Southgate, 1976). The major problems involve the loss of most of the non-cellulosic polysaccharides and part of the lignin (Theander and Westerlund, 1993).

The use of detergents in animal feed analysis was introduced by Van Soest (1963a). It is now the most commonly used procedure for the determination of ruminant-feed forage fibre quality

(Hussain et al., 1996). However, there are problems associated with its use for the accurate quantitative determination of polysaccharides.

The acid detergent fibre (ADF) method is one such detergent procedure (Goering and Van Soest, 1970, Van Soest, 1963b). Here the sample is heat-treated with 0.5 M sulphuric acid containing cetyl-trimethyl-ammonium bromide. The residue that remains is called the ADF residue and is said to be comprised mostly of cellulose and lignin.

The neutral detergent fibre (NDF) method involves extraction using a hot neutral solution of sodium lauryl sulphate (Van Soest, 1963a, Van Soest and Wine, 1967). This is said to be an approximation for the cell wall content and, hence, the difference between NDF and ADF is said to give the content of non-cellulosic polysaccharides (which is typically assumed to be hemicellulose).

The reliability of detergent methods is likely to be dependent on the ash and bound protein contents of the feedstock, since these are present in the ADF and NDF fractions and could result in an overestimation of cellulose.

While, in some instances, fibre analysis can give results that are somewhat similar to chromatographic methods for the cellulose content (Wiselogle et al., 1996, Theander and Westerlund, 1993), the hemicellulose content tends to be overestimated. This is due to the variable responses of the hemicellulose polysaccharides of differing structure to the extraction conditions (Wiselogle et al., 1996).

Examples of inaccuracies in the detergent method include that of the experiments by Theander and Aman (1980) where interferences were found with straw, alkali-treated straw, grass, and alfalfa samples. Examples of the interferences include the ADF fraction containing 7-14% hemicellulose and 1-4% crude protein, besides cellulose and lignin. The cellulose residue also contained 8-13% hemicellulose and 2-7% lignin. The NDF fraction contained 1 to 6% crude protein, as well as the expected hemicellulose, cellulose and lignin.

Lignin can be determined in detergent fibre procedures via the permanganate method (Van Soest and Wine, 1968). This involves oxidation of the ADF fraction with potassium permanganate. These permanganate lignin values tend to be significantly lower than the Klason lignin (Section 3.1.2) values for various types of lignocellulosic material, particularly grasses (Lindgren et al., 1980, Sundstol et al., 1978). Indeed, the 2-stage acid-hydrolysis methods may sometimes give lignin contents that are as much as 300% greater (Theander and Westerlund, 1993).

An alternative to the permanganate lignin method is to solubilise the ADF residue with 72% sulphuric acid and assume that the remaining solid residue is the lignin (Moller, 2009). This fraction is labelled the acid detergent lignin (ADL). In using 72% sulphuric acid it is somewhat similar to the KL method (Section 3.1.2). However, that process does not require the separation of the ADF fraction beforehand.

It is also important to note that these gravimetric methods calculate the cellulose, hemicellulose and lignin contents indirectly (either through mass loss or mass retained). They do not include the characterisation of these polymers which, in the case of the polysaccharides, means that the relative proportions of the individual monosaccharides which constitute them will not be known. This is not of importance in the homopolysaccharide cellulose but it will be in the hemicelluloses which can have many different sugar configurations (Section 2.1.3.3).

Despite the limitations of detergent fibre analysis, the ADF, NDF, permanganate lignin and ADL methods are still used extensively, particularly in forage analysis. There are a large number of NIR calibration equations that have been developed for these fractions for a range of biomass feedstocks (see Appendix B for some). Indeed, as outlined in Section 5.3.1, the initial development of NIR as a tool for the quantitative analysis of biomass was a direct result of its application in the forage and grain sectors. However, the inability of the method to differentiate between the component monosaccharides of the structural polysaccharides, while of less importance in forage analysis, is a great limitation in second generation biorefining applications where these different sugars may have very different fates, and so result in significantly different yields and product streams. A method that could differentiate between these different sugars accurately and reliably would therefore have a distinct advantage.

3.1.2 Acid-Hydrolysis Methods

Acid hydrolysis methods, for the most part involving sulphuric acid and a two stage hydrolysis, have been used and developed extensively for lignocellulosic samples over the last hundred years. The term Klason lignin comes from such a two-stage procedure developed by Klason for the isolation of lignin from wood (Klason, 1922). In the first few decades of the 20th century the focus was on determining the lignin content, but the hydrolysis methodologies from the 1930s onwards started to give more consideration to the yields of sugars as well as lignin (Ritter et al., 1933). Papers also started to examine the use of such methodologies for grasses as well as woods.

3.1.2.1 The Uppsala Method

The most significant more recent development of lignocellulosic analytical methodologies, specifically relating to the field of biofuels, came from a paper by Theander at the Swedish University of Agricultural Sciences in Uppsala (Theander, 1985). The procedure, known as the Uppsala Method, allows for the determination of uronic acid residues, extractives, water-soluble carbohydrates, acid-soluble lignin (ASL) and Klason lignin (KL) as well as neutral sugar residues. It incorporates the removal of 80% ethanol-soluble extractives (and also, if necessary, the removal of starch with enzymes) prior to the two-stage hydrolysis. The procedure is important because a modified method of it was thoroughly tested in a round-robin involving multiple laboratories in which several whole biomass feedstocks were analysed, including a hardwood, a softwood, wheat straw, and sugarcane bagasse (Milne et al., 1992, Agblevor et al., 1993). It was concluded that this procedure could be used to obtain feedstock compositions for both woody and herbaceous feedstocks (Chum et al., 1993). A round robin was also conducting using the method on food samples (Theander et al., 1995).

The results of the biomass round robin gave rise to the certified compositions (with associated error levels) of glucan, xylan, arabinan, mannan, galactan, KL, acid soluble lignin, ash, and ethanol soluble extractives for four feedstocks – sugarcane bagasse, Eastern Cottonwood, Monterey pine, and wheat straw. These materials are available for purchase from the National Institute of Standards and Technology (NIST) and allow analysts to compare the results of their analytical procedures with those of the round robin (NIST, 2011).

The Uppsala method is primarily designed for the determination of the cell-wall content of monosaccharides (which exist in the cell wall in their polymeric forms). Hence, the removal of low molecular-weight sugars (water soluble carbohydrates and non-structural carbohydrates) and extractives is necessary before polymer hydrolysis.

After removal of extractives (and starch) the next stage of the Uppsala Method involves the hydrolysis itself. This involves adding 3 ml of 12 M (72%) sulphuric acid to 300 mg of the biomass sample. The sample should then be incubated at 30°C for one hour in a water bath, stirring occasionally. This is the primary hydrolysis stage. Following this, 84 ml of water is added so that the sulphuric acid concentration of the solution is reduced to 4%. The secondary hydrolysis stage then takes place with the solution being hydrolysed in an autoclave for one hour at 121 °C. Following this, the hydrolysed solution is filtered through a filter-crucible and the hydrolysate retained for subsequent analysis. This solution will contain all of the hydrolysed sugars, any sugars that may have been degraded during the acid hydrolysis process, the acid soluble lignin (ASL), the uronic acids, acid soluble ash, and any other acid-soluble constituents of the biomass. The solid residue that does not

pass through the filter crucible contains all the acid insoluble material. This can then be ashed and the ash content subtracted from the weight of the acid insoluble residue (AIR) weight in order to determine the Klason lignin content.

3.1.2.2 Principles of Acid Hydrolysis

Cellulose, in pure water, is hydrolysed through attack by the electrophilic hydrogen atoms of the H₂O molecule on the glycosidic oxygen. This is a very slow reaction because of the resistance of cellulose to hydrolysis. Acids can be defined as non-specific catalysts that speed up this reaction and push the equilibrium towards hydrolysed units. Only strong acids, such as sulphuric acid, are suitable for efficient hydrolysis without the degradation of the liberated sugars.

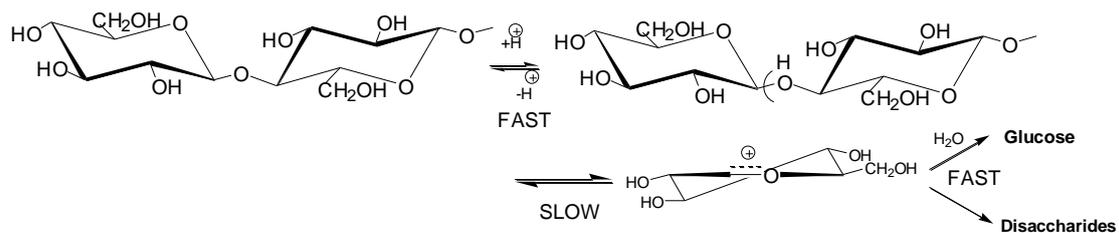


Figure 3-1: Steps involved in the acid hydrolysis of cellulose. Adapted from (Sjostrom, 1981)

Figure 3-1 illustrates the steps involved in the acid-catalysed hydrolysis of the glycosidic bonds of cellulose. It is presented as though the first glycosidic link in the chain is being attacked. The reaction starts with a rapid proton addition to the aglycone oxygen atom. The H⁺ ions equilibrate between the O atoms in the system, including those of water and the glycoside. That means that there is an equilibrium concentration of protonated glycoside. This equilibrium concentration tends towards protonation of the glycoside with increasing temperature and/or increased acid concentration. The protonated conjugate acid then slowly breaks down to the cyclic carbonium ion (while the other residue retains the OH at C-4). After a rapid addition of water, free sugar is liberated. Because the sugar competes with the water small amounts of disaccharides are formed as *reversion* products.

The point of highest energy in this sequence is the transition state for the formation of the carbonium ion; the structure at this point is somewhere between the protonated glycoside and the carbonium ion. The reason for the high energy in the carbonium ion is that a positively charged

carbon atom is an unstable arrangement. The instability of this form is the reason for the adaptation of the half-chair configuration. In this state, along with C-2, C-1, the ring oxygen and C-5 are coplanar, the positive charge is “shared” somewhat with the ring oxygen thereby conferring some stability.

Given the slow rate of formation of the carbonium ion, the lower energy required for its formation should result in a faster reaction. This can explain why most other glycosidic linkages are easier to break than those in glucopyranosides. For example, xylopyranosides hydrolyse faster due to the exchange of CH₂OH on C-5 on the ring for H (smaller and with a lower electronegativity). The replacement of H by CH₃ (to form an analogue of fucose and rhamnose) leads to a further small increase in rate which can be explained by the fact that, although CH₃ is larger than H, it donates electrons and hence stabilises the charge of the carbonium ion (Rees, 1967).

The rate of formation of the carbonium ion can also be used to explain why the differences in the hydrolysis rates of diastereomeric glycosides are significant. For example, the relative hydrolysis rates of methyl- α -D-gluco-, manno- and galactopyranosides are 1.0:2.9:5.0 (Sjostrom, 1981). This can be related to the stabilities of their respective conjugate acids, which are transformed into the half-chair carbonium ions at different rates. Also, substituents bound to the C-2 position obviously prevent the formation of the half-chair conformation.

There are several considerations with regard to the relative rates of hydrolysis of polysaccharides. These include:

- The formation of the intermediate carbonium ion takes place more rapidly at the end than in the middle of the polysaccharide chain. Hence many of the glycosidic linkages in polysaccharides with lower degrees of polymerisation may be hydrolysed relatively quicker than those in longer chains;
- Furanosides are hydrolysed much more rapidly than the pyranoside analogues; and
- Carboxyl groups bound to the polysaccharide chains have a considerable influence on the rate of acid hydrolysis (for example glucuronides are hydrolysed more slowly than glycosides – see Section 3.2.4). The influence is mainly due to steric interaction, even when inductive (stabilising) effects are also considered.

An important process in Figure 3-1 is that, as well as the glucose monomers, reversion products can be formed. This reversion process involves these monomers re-dimerising to several disaccharides,

these can include maltose and cellobiose (Section 2.1.2 and Figure 2-4). There also exists the potential for trimer formations. These reversion products exist in an equilibrium with the monomers and in dilute acid solutions the position of the equilibrium is shifted towards favouring monomers. However, the rate of change is slow. Therefore, part of the reason for the secondary hydrolysis stage in the Uppsala Method is to ensure that as much as possible of the dimers are rapidly broken up and that the sugars exist as free monosaccharides in solution, so enabling standard chromatographic methods for monosaccharide analysis to be employed. The secondary hydrolysis stage is also partially necessary to hydrolyse any oligosaccharides that may not have been fully degraded by the acid.

Therefore, it is considered that the primary acid hydrolysis stage will result in little to no degradation of the sugars but that these will exist partially in forms (dimers, trimers, oligosaccharides etc.) that require further treatment, or secondary hydrolysis, before accurate quantitative analysis methods can be employed. However this secondary stage occurs at elevated temperatures and is considered to result in the partial degradation of the sugars (Dinus, 2000). Possible products of this degradation include hydroxymethyl furfural, levulinic acid, and formic acid from the hexoses and furfural from the pentoses. These products can potentially polymerise to produce solid residues in some cases (Hayes et al., 2005).

In order to correct for the degradation of proportions of the monosaccharides, the Uppsala Method involves the use of sugar recovery solutions (SRS) that contain known quantities of the sugars of interest in a 4% sulphuric acid solution with the concentrations of the sugars being equivalent to those that would be expected of the biomass samples being analysed. The SRS are then placed in the autoclave with the biomass hydrolysates so that they are also put through the secondary hydrolysis procedure. Chromatographic determination of the sugar concentrations of these SRS solutions after the secondary hydrolysis can then be compared with the known concentrations before and sugar loss correction factors determined for each sugar. These can then be applied to the concentrations determined for the sample hydrolysates in order to approximate for the sugar compositions before sugar-loss was experienced.

3.1.2.3 NREL Hydrolysis Method

Probably the most well-known and widely used methodologies for the preparation and analysis of lignocellulosic feedstocks are those that are available for download on the National Renewable Energy Laboratory (NREL) website and described in Sluiter *et al.* (2010) and Temleton *et al.* (2010). (The URL on 26/3/2011 was http://www.nrel.gov/biomass/analytical_procedures.html). The

procedure for the determination of KL, acid-soluble lignin (see Section 3.2.3) and structural carbohydrates is similar to the Uppsala method (Theander et al., 1995) except sealed hydrolysis tubes are used in the secondary stage and HPLC is used as opposed to the GC analysis of the alditol acetate derivatives. In 1995 the NREL methodologies were issued as American Society for Testing and Materials (ASTM) standard 1721 “Standard Test Method for Determination of Acid-Insoluble Residue in Biomass” and ASTM standard 1758 “Determination of Carbohydrates in Biomass by High Performance Liquid Chromatography”.

3.1.3 Other Methods and Problems with the Uppsala/NREL Procedures

While the Uppsala/NREL methods, in contrast to the detergent fibre methods, allow for the direct analysis of the constituents of the polysaccharides, the technique used for lignin estimation is, like the fibre methods, an indirect gravimetric approach. It is possible that non-ligneous elements - such as condensed proteins, suberins and cutins - may be present in the residue, resulting in an overestimation of the lignin content. It is also possible for sugar derivatives to be degraded by the acid to insoluble materials, and these will be incorrectly assigned to the Klason lignin fraction (Theander, 1983). For this and other reasons it is considered that lignin content may be particularly vulnerable to overestimation when analysing weathered feedstocks (Agblevor et al., 1994). Furthermore, incomplete hydrolysis of the polysaccharides (due, perhaps, to inefficient stirring of the biomass sample in the primary hydrolysis stage) may add to the acid-insoluble component that will be assumed to be lignin (Templeton and Ehrman, 1995). On the other hand, if the hydrolysis procedure is too severe then HMF, furfural, and other potential sugar degradation products could condense/polymerise to solid products that would again be incorrectly assigned as part of KL.

It should also be considered that the acid treatment results in a partial change in the lignin structure and properties, due to condensation reactions (Lai and Sarkanen, 1971). Hence the structure of the KL will be different to that of lignin in the original biomass.

Another potential problem with the Uppsala/NREL methodology is that the hydrolysis procedure does not differentiate between cellulose and the hemicelluloses – all polysaccharides are hydrolysed. As detailed in Sections 2.1.3.3 and 10.1.3, the hemicelluloses, as well as cellulose, can contain glucose. Furthermore, there may also be oligosaccharides, or other polysaccharides, that were not removed in the extractives step that will be liberated upon acid hydrolysis. The results of

chromatographic analysis of the hydrolysed sugars will not be able to differentiate between the origins of any of the sugars.

In some instances formulae have been used to approximate the cellulose and hemicellulose contents, based on the glucose content and the quantities of other sugars in the hydrolysate, for a range of feedstocks. For example Jones *et al.* (2006) estimated the glucose content of pine wood by taking the glucan content and subtracting from this one-third of the mannan content of the sample. The hemicellulose content was then determined as the sum of arabinan, galactan, glucan, mannan and xylan minus the calculated cellulose content. These formulae are only approximations, however, since the relative proportions of monosaccharides in hemicelluloses can change (section 10.1.5) and since most of these equations do not usually consider the influence of other polysaccharides in the matrix.

3.2 Ultraviolet-Visible Spectroscopy

3.2.1 Principles of UV-Vis

Section 5.1 details the principles behind spectroscopy. Visible light covers a small part of the electromagnetic (EM) spectrum and has shorter wavelengths (from around 400 nm for violet light to approximately 780 nm for red light), and thus higher frequencies and energy, than infrared radiation. Ultraviolet (UV) is the section of the EM spectrum whose frequencies are above those of visible light but less than that of X-rays, with wavelengths ranging from 10 nm to 400 nm. UV radiation can be subdivided into Near-UV (200-400 nm) and Vacuum UV (10-200 nm). Vacuum UV is so-called because oxygen is highly absorbant of light at these wavelengths; hence either air or oxygen must be removed in order to utilise this region of the spectrum for analytical purposes.

While near infrared radiation (Section 5.2) is of sufficient energy to cause atoms within a molecule to vibrate, it does not have the energy required to cause electrons to change their orbital locations. Radiation in the UV-Vis region is, however, capable of achieving this for some of the bonds of molecules, as shown in Figure 3-2.

In order for electrons in the stronger σ (sigma) bonds to be elevated to an excited state, light of wavelengths less than 200 nm, e.g. vacuum UV radiation, is required. However, Near-UV-Visible radiation can result in the excitation of π bond electrons and in the excitation of non-bonding (n)

electrons to the excited antibonding π state (i.e. π^*) or in their promotion to the non-bonding σ^* orbital (Denney and Sinclair, 1993). Figure 3-3 shows the $n \rightarrow \sigma^*$ transition in primary amines, this transition occurs at around 220 nm (Rouessac and Rouessac, 1998). An $n \rightarrow \pi^*$ transition usually occurs in molecules that contain a heteroatom (i.e. not carbon or hydrogen) as part of an unsaturated system (e.g. the carbonyl band at around 270 to 295 nm (Rouessac and Rouessac, 1998)).

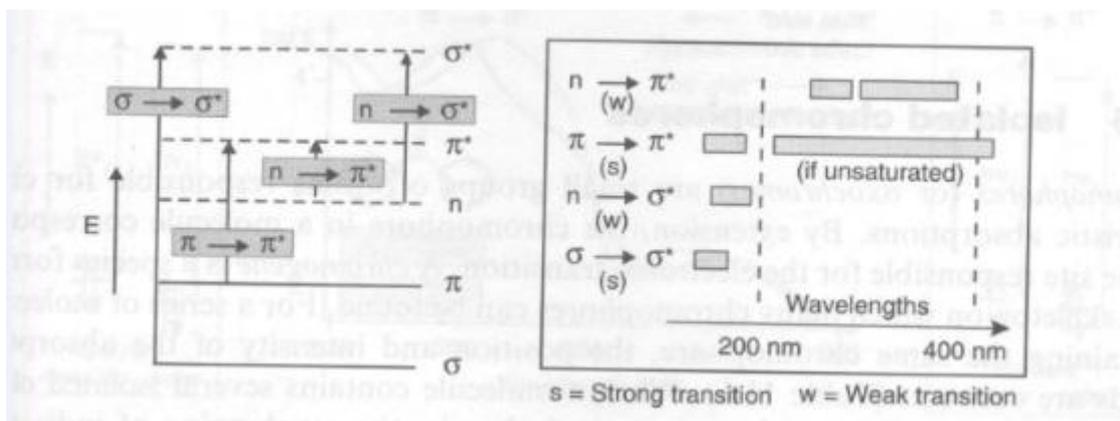


Figure 3-2: An illustration of some of the most common transitions that can occur with UV-Vis radiation along with the associated wavelengths for these transitions. Taken from (Rouessac and Rouessac, 1998).

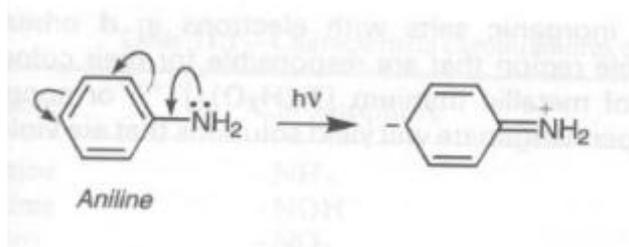


Figure 3-3: An example of a $n \rightarrow \sigma^*$ transition in a primary amine. Taken from (Rouessac and Rouessac, 1998).

The absorbance bands that are seen for the excitation of electrons from a lower to higher energy level are narrow for atoms but in molecules the change in electronic energy is accompanied with a corresponding change in the vibrational (see Section 5.2) and rotational energy levels and, since there are a large number of possibilities for these energy changes, the absorption spectra become much broader (Denney and Sinclair, 1993). This is illustrated in Figure 3-4.

A chromophore is defined as a group in a molecule that has an absorbance in the wavelength region 185-1000 nm. The wavelengths of the absorptions of these chromophores, and their extinction coefficients, can be greatly influenced by the presence of other chemical groups in the molecule, with such influencing groups being known as auxochromes. For instance, groups that possess unshared electrons (e.g. amines, hydroxyl) have the ability to donate electrons to the conjugated

system and therefore tend to have a bathochromic effect (shifting to longer wavelength) since this delocalisation will mean that less energy will be required to promote one of these electrons to the excited state. Auxochromes also tend to increase the molar absorptivity of chromophores (Denney and Sinclair, 1993).

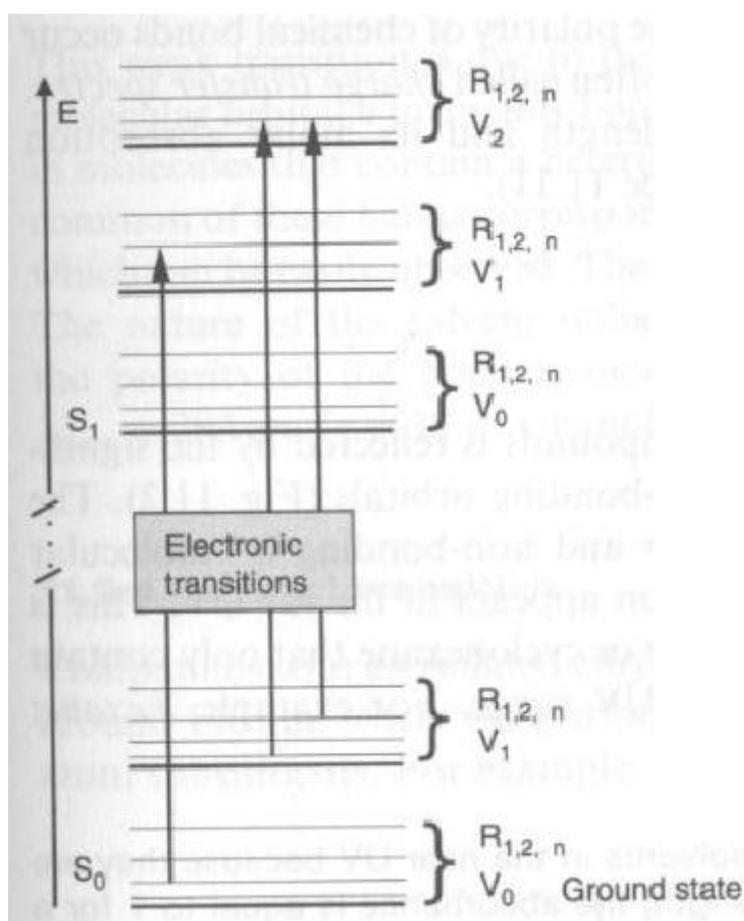


Figure 3-4: The electronic, vibrational, and rotational energy states of a molecule. Taken from (Rouessac and Rouessac, 1998).

If a compound is transparent in the UV-Vis region when isolated it can potentially be made to absorb this radiation when it is present with a species with which it can interact through a donor-acceptor relationship. In this situation an electron from a bonding orbital of the donor is transferred to an unoccupied orbital of this compound resulting in the donor species becoming a cation. In order for this transition to take place the unoccupied level of the accepting species needs to be close in energy to the bonding orbital of the donor (Rouessac and Rouessac, 1998). Transparent molecules can also be analysed in the UV-Vis region if they are derivatised to non-transparent forms. Alternatively derivatisation can be employed to resolve between two species that have similar spectral profiles.

3.2.2 UV-Vis Instrument Used

In October 2009 a dedicated UV-VIS instrument for the analytical needs of the Carbolea group was installed by the Author after a period of time searching for the most appropriate device. The unit is an Agilent HP 8452A diode array spectrophotometer. It is a single beam unit that can analyse liquids, via transmission spectroscopy, over the wavelength region 190-510 nm. It has a bandwidth of 2 nm, a wavelength accuracy of ± 2 nm and a photometric accuracy at 1 Absorbance Unit of ± 0.005 nm.

This instrument uses a deuterium lamp which provides good intensity over the UV and over part of the visible spectrum. It has a (silicon-based) photodiode array as its detector system. In a photodiode the light that falls on the semiconductor material results in electrons flowing through it which depletes the charge in a capacitor connected across it. The amount of charge that is needed to recharge the capacity will be a function of the light intensity. When a series of these photodiodes are arranged together they form a photodiode array (PDA).

The HP8452A does not disperse the light prior to its interaction with the sample. Instead, after passing through the sample and an entrance slit to the detection module, the transmitted light is dispersed with a holographic grating so that the different wavelengths will fall upon different photodiodes and a full spectrum of the sample can be obtained for every wavelength simultaneously. The diode array configuration allows for a spectrum to be collected in 0.1 seconds. The standard set-up employed in analysis by the Author involved a 0.5 second measurement time over which the average responses of each diode over this period were taken to form the spectrum.

In contrast to the FOSS XDS NIR instrument (see Section 5.3.2.3) there are no moving parts in the HP8452A (except for the shutter and cooling fan) and, since only light incident with the source-emitter can pass through the entrance slit to the polychromator, interference from ambient light is not an issue.

A 1 cm path-length UV-grade quartz cuvette (3 ml capacity of sample) is used for obtaining the spectra of samples. Since the device is a single-beam instrument a blank measurement will need to be taken at a different time to the sample collection. The blank is required to determine the value for I , the intensity of incident radiation, so that, following determination of I_0 , the intensity of transmitted radiation, the transmittance and absorbance can be determined (see Section 6.1).

3.2.3 Acid Soluble Lignin Analysis

A small portion of the lignin of many materials becomes soluble in the hydrolysis procedure (Section 3.1.2). This means that Klason lignin (KL) is not a true reflection of the total lignin content of the sample. Methods of determining the acid soluble lignin (ASL) are therefore required to calculate total lignin content (Lignin = KL + ASL).

There have been some investigations of the structures of ASL. Swan (1965) suggested that it is primarily composed of two types of material with differing polarities – low molecular weight lignin products that are soluble in chloroform, and various water-soluble substances. It has also been found (Yasuda et al., 2001) that some ASL materials that are soluble in 72% sulphuric acid (the primary stage of hydrolysis) become insoluble again, and part of the KL, when boiled in 3% sulphuric acid (the secondary stage of hydrolysis).

Typically the ASL is measured using ultraviolet spectrophotometric transmission methods on the diluted hydrolysate. Various wavelengths can be used for this, including 200-205 nm, 240 nm, 280 nm or 320 nm. The following basic formula is used for determining ASL content from an absorbance value (Maekawa et al., 1989):

$$ASL (\%) = \frac{D \times V (A_{Sample} - A_{Solvent})}{a_L \times S_w} \times 100 \quad (3.1)$$

Where D is the dilution factor; V is the hydrolysate volume (ml); A_{Sample} is the absorbance of the sample; $A_{Solvent}$ is the absorbance of the solvent (i.e. the blank); a_L = absorbtivity of ASL (in $M^{-1} cm^{-1}$); S_w = sample weight (in mg).

The UV methodology is relatively simple to execute (provided extinction coefficients are available). However, the method is indirect since there can be interferences from other UV-absorbing non-lignin components such as HMF and furfural, and possibly acid degradation products of extractives that may not have been fully removed in the extraction step (if one were used). It can also be difficult to determine an accurate extinction coefficient; for example, the 205 nm absorptivity for a species of eucalyptus was calculated to be between 88-103 $M^{-1} cm^{-1}$ ($a_L = Absorbance / (molarity \times pathlength)$) depending on which model lignin compound was used (Bland and Menshun, 1971). Also Beer's law (Section 6.1) may not hold for differing concentrations of ASL, due to non linearities at high concentrations. To avoid this non-linearity effect the solution should be diluted to target reasonably consistent absorbance values.

The presence of furfural and hydroxymethylfurfural in biomass hydrolyates typically precludes the use of 280 nm for ASL analysis since these molecules absorb at this wavelength.

Table 3-1: Some ASL absorbtivity constants that were found in the literature

Feedstock	λ (nm)	Absorp. $M^{-1} cm^{-1}$	ASL (%)	Reference
Hardwoods	205	110		(Maekawa et al., 1989)
Hardwoods	240	30		(Maekawa et al., 1989)
Hardwoods	280	23.6		(Maekawa et al., 1989)
Softwoods	280	23.3		(Maekawa et al., 1989)
Birch	205	113		(Musha and Goring, 1974)
Eucalyptus	205	106		(Musha and Goring, 1974)
Norway Spruce *a	203	128	-	(Raiskila et al., 2007)
Wood and pulp *b	205	110	-	(TAPPI, 1991)
Sugarcane bagasse *c	240	25	6.6	(Yu and Stahl, 2008)
Sugarcane bagasse *d	280	23.7	-	(Jackson de Moraes Rocha et al., 2011)
Pinus radiate	240	12	-	(Sluiter et al., 2006a)
Sugarcane bagasse	240	25	-	(Sluiter et al., 2006a)
Corn stover	320	55	-	(Sluiter et al., 2006a)
Populous deltoides	240	60	-	(Sluiter et al., 2006a)

*a = Figure corrected for absorbance by carbohydrates; *b = This is a TAPPI standard that is often used in research papers; *c = The authors used the NREL absorptivity constant for bagasse (Sluiter et al., 2010); *d = The absorbances of furfural and HMF were removed

There are a very wide range of absorbtivity values for different biomass materials at different UV wavelengths. Much of the data in the literature cover the values for woods at 205 nm with the values typically falling in the range of 88 to 113 $M^{-1} cm^{-1}$ (Sluiter et al., 2010). Some of the asbsorbtivity constants and chosen wavelengths that the Author came across are listed in Table 3-1. Clearly, the UV method for ASL determination has its problems and uncertainties. However, since researchers are still unsure of the composition of ASL, and given that there are no ideal methods currently available (chromatographic methods may not recover all ASL components and may involve solubility problems in the mobile phase; NMR methods have not yet been fully developed (Sluiter et al., 2010)), the UV method appears to be the best option for determining ASL content without placing over-demanding work-loads on the researcher for the analysis of a lignocellulosic component which is relatively minor for most feedstocks and of no direct known value in most biorefining schemes focussed on the hydrolysis of polysaccharides.

3.2.4 Uronic Acid Analysis

There are numerous methods by which uronic acids (UA) can be analysed. However, these methods cannot usually be employed on the hydrolysate from the standard hydrolysis methods outlined in Section 3.1.2 since the conditions employed are insufficient to result in the production of UA as free

monomer units that can be readily analysed. The hydrolysis of UA is restricted due to the high stability of the glycosyl uronic acid linkages and may only result in the formation of dimeric aldobiuronic acids rather than monomers (Theander and Westerlund, 1993). Therefore, most methods involve a further acid hydrolysis, with more concentrated acid, in order to release the UA as monomers.

Colorimetry (visible spectroscopy) has been used as a means of analysing for the UA once they have been hydrolysed. The Uppsala procedure (Section 3.1.2.1) has a protocol for the hydrolysis and analysis of uronic acids. It involves taking a 250 μL aliquot of the 4% sulphuric acid hydrolysate that was produced in the standard hydrolysis method and mixing it with 250 μL boric acid-sodium chloride solution in a glass tube. The subsequent hydrolysis step involves the addition and mixing of 4.0 mL 18 M H_2SO_4 (12M H_2SO_4 is used in the standard hydrolysis method) in the tubes which are then placed in a 70°C water bath for 40 minutes. Following this the tubes are cooled to room temperature in the water bath and then 200 μL of 3,5-dimethylphenol solution is added and the solution thoroughly mixed. The sample absorbance is measured at two wavelengths, 400 and 450 nm, between 10 and 25 minutes after the addition of 3,5-dimethylphenol. It is considered that hexoses will absorb over these wavelengths, and that, by subtracting the absorbance at 400 nm from that at 450 nm, this can be corrected for.

A calibration needs to be developed to determine UA concentration from the absorbance. This is done using galacturonic acid monohydrate as a standard, made up at various concentrations in 4% sulphuric acid solutions. These are put through the second stage (i.e. autoclaving) of the standard hydrolysis method. The absorbances of these known solutions can then be determined using the same method as mentioned above.

The colorimetric method is far from ideal, and many of the problems associated with the spectroscopic analysis of ASL content are also present here – principally that it is an indirect method and other components may also absorb at these wavelengths. Such components, apart from the neutral sugars, include proteins and phenols (Ahmed and Labavitch, 1977). It is also considered that the method is somewhat sensitive to the conditions employed and results will not be reproducible unless the temperature and time in the water bath, as well as the time between the addition of the dimethylphenol solution and spectroscopic analysis are consistent. Also, the method only gives an estimate of total UA content and does not differentiate between the different types of UA that may be present in the hydrolysate. Furthermore, these different UA may have different reaction rates than the standard (galacturonic acid). That means that the calibration curve may not be entirely

accurate for all UA, leading to inaccuracies in the estimation of the total UA content. The method was involved in a round-robin study involving many different laboratories and it was found that there was between 20 and 30% variability for the four NIST standard biomass feedstocks that were tested (Aglevor et al., 1993).

There are methods in the literature for the analysis of UA by: gas chromatography (Ha and Thomas, 1988, Lehrfeld, 1981), although this requires derivatisation; and cation-exchange chromatography (Kaar et al., 1991). However, there is a problem for all chromatographic methods in that many UA standards are not commercially available, meaning that individual response factors and quantifications for all UA is not currently possible unless these UA are isolated by the researcher from biomass feedstocks – a lengthy and complicated process.

Given that UA make up a relatively small proportion (between 1 and 3% - see Sections 13.3 and 15.2.2.1) of the feedstocks that are most relevant to this study, and that their value (if any) to biorefining is far from clear (Section 2.1.1), it was decided by the Author that UA analysis would only be carried out to a limited degree and that a simple, yet reasonably reliable, protocol that would not be too time-consuming should be employed. It was therefore decided that, based on the information available in the literature, the colorimetric method would be used for a selection of biomass samples.

3.3 Moisture Content

Most analytical procedures determine the moisture content and dry matter (solids content) by heating a sample at 105 °C to constant weight (Ehrman, 1994c). The moisture content from this method will be a measure of the amount of water (and other components volatilized at 105 °C) present in a sample while the solids content is the amount of sample remaining. These are usually put forward on a wet-mass basis. It should be noted, however, that some extractives may become volatile at this temperature (see Section 3.4) and so the term “moisture content” is a bit of a misnomer, a term such as “mass loss on heating to 105°C” may be more appropriate.

Water content can also be determined with lyophilisation (freeze-drying). While drying at 105°C results in damage to the sample and a decrease in extractive content, lyophilisation generally has less of an effect. It therefore may be of particular use in the storage of samples that would otherwise be subject to biotic degradation. It is, however, the most complex, time-consuming and expensive

form of drying and its use is usually restricted to delicate, heat-sensitive materials of high value (Snowman, 1988). This procedure will result in a higher estimation of the dry matter content since most extractives remain in the biomass; Jirjis (1995) found that it was approximately 2% greater for wood than when measured by the standard 105 °C procedure.

3.4 Extractives Content

The removal and determination of the extractives content is important for four main reasons:

1. It allows the total mass balance for the analysed constituents to get closer to 100%, indicating a complete and reliable analytical methodology.
2. Extractives may have an influence on biorefining technologies and storage dynamics of biomass.
3. There may be valuable constituents in the extractives themselves.
4. If they are not removed, the extractives may interfere with the acid hydrolysis procedure (Section 3.1.2) and give inaccurate results for KL and structural polysaccharides.

3.4.1 Methods of Extraction

A variety of solvents are possible for the removal of the extractive components. For example, water-soluble carbohydrates, tannins, and inorganic salts can be liberated from the biomass with hot or cold water (Sjostrom, 1981). Organic solvents such as ethanol, acetone or dichloromethane are needed for the extraction of other extractives, such as resin acids, fat, and terpenes, for example. Therefore, as the composition of extractives varies between species, and within different anatomical fractions of the same species over time, no solvent is equally applicable to all biomass and all biomass fractions (Hakkila, 1989).

Standard methodologies exist for the removal of extractives from wood and these typically require a sequence of extractions, using different solvents. For example, an ASTM method puts forward a sequence involving ethanol-benzene, then ethanol, and then hot water (ASTM, 1993). Clearly such a process will require a significant amount of work and it would be preferable to utilise a single treatment that will remove the majority of the extractives and, most importantly, allow accurate determinations of the KL content and of the constituent monosaccharide units of cellulose and hemicellulose.

The most relevant one-solvent procedures are the 80% ethanol extraction that is part of the Uppsala Method (Theander et al., 1995), and a modified version of this using 95% ethanol (Ehrman, 1994b). Both of these procedures are said to be able to remove hydrophilic and lipophilic extractives without affecting the structural lignocellulosic components of the biomass (Theander and Westerlund, 1993).

The 80% ethanol method (as initially used for the sugarcane bagasse samples, Section 12.2) involves approximately 500 mg of the sample being placed into a centrifuge tube and 25 ml of 80% ethanol added. The sample is placed in an ultrasonic water bath for 15 minutes, and the tube is then placed in a centrifuge and the supernatant liquid (containing the extractives) removed. This is repeated four times and then the extracted residue is dried at 40°C. Since these extractions are made in centrifuge tubes, the extract can be retained during washing and centrifugation steps. This method was evaluated by Theander and Westerlund (1993) for the removal of extractives from wheat straw and it was found that four 80% ethanol extractions removed the majority of the extractives and that, with the subsequent extraction with acetone followed by petroleum only trace amounts of extractives were removed.

Since the original Uppsala Method was put forward, there have been revisions proposed to the extractives removal methodology. For example, the procedure by Ehrman (1994b) at NREL proposes using 95% ethanol. This procedure was accepted by the ASTM as a Standard Test Method for the determination of extractives in biomass feedstocks and it is said to be suitable for the determination of ethanol soluble extractives of woods, herbaceous materials, agricultural residues and wastepaper. In this procedure a sample of the biomass is added to a soxhlet extraction thimble and the thimble is then inserted into a conventional Soxhlet apparatus and heated at reflux for 24 hours. The procedure says that approximately 100 to 200 solvent exchanges are required during the period and that at least 160 ml of 95% ethanol will be required per sample. After the extraction, any residual solvent is removed from the residue using vacuum filtration and the sample is then washed thoroughly with 95% ethanol, collecting all of the filtrate. The filtrate should then be combined in a flask with any solvent from the upper section of the Soxhlet apparatus and the flask then placed on a rotary evaporator and the solvent removed under vacuum. When all of the visible solvent is removed by the rotary evaporator, the flask should be placed in a vacuum oven at $40 \pm 1^\circ\text{C}$ for 24 ± 1 hour. When cooled the flask should be weighed. The weight of the extractives can then be determined directly, or indirectly based on the mass loss of the original sample (once moisture content corrections have taken place). A more recent version of the Ehrman procedure, as referenced in Sluiter *et al.* (2010) and available for access at the NREL website mentioned previously, recommends that the same methodology be employed for water extraction prior to the ethanol

extraction if there are likely to be large quantities of water-soluble extractives in the feedstock (as is the case with corn stover, for example).

3.4.2 Effect of Extractives on Hydrolysis

There are papers from more than 70 years ago indicating that extracting wood with solvents prior to the analysis of lignin will improve the accuracy of the KL determination (Ritter et al., 1932). The reasoning behind this is that if these extractives are not removed they will cross-react with the acid and condense to acid insoluble components that will be associated with and classified as Klason lignin (Browning, 1967). The presence of these extractives during the acid hydrolysis process may also impact on the hydrolysis dynamics and ultimate fate of the monosaccharides that are liberated on hydrolysis of the polysaccharides. Furthermore, carbohydrates that were part of the extractives may either be degraded by the relatively harsh acid conditions (that are designed for the hydrolysis of structural polysaccharides) or survive as free monosaccharides in the hydrolysate meaning these could be incorrectly classified as constituents of cellulose or hemicellulose.

A paper by Thammassouk *et al.* (1997) compared the analytical results for KL, glucan, xylan, arabinan, mannan, galactan, acid soluble lignin, protein, ash, and uronic acids after extraction with either (i) nothing, (ii) hot water, (iii) 95% ethanol, for three biomass feedstocks (switchgrass, fescue straw, and corn stover). It was found that for switchgrass the extractives content differed significantly between the two extraction methods (9.7% of the dry matter with ethanol extraction compared with 16.4% with water extraction), and that hydrolysis with no extraction significantly increased the reported value for KL and ASL. There was a similar situation for fescue straw, and both it and switchgrass had lower glucan contents after extraction. In the case of corn stover, there was also a significant increase in the KL and ASL content of the non-extracted sample, but the glucan content was not significantly different. The authors concluded that water and 95% ethanol were both effective in removing extractives, although different components were removed by each solvent in some cases, and in allowing a more accurate determination of the true KL content.

3.4.3 Issues with Extractives Calculations

As mentioned in Section 3.3, the determination of moisture content through heating at 105 °C to constant weight may result in some volatile extractive compounds being lost to the atmosphere. That could lead to an overestimation of moisture content and an underestimation of extractives content.

This error has been particularly influential in storage studies, where determinations are made regarding the dry-mass-loss of a material as it is stored. In some cases negative losses have been reported (Hudson et al., 1988, Heding et al., 1993).

The presence of volatile extractives as a source of error has long been accepted in the evaluation of silages, with dry matter often being determined via toluene distillation rather than heating (Jirjis, 1995). Schneider (1995) presented correction formulae and graphs to correct dry matter loss calculations in wood fuel containing volatile extractives other than water. To use these calculations, however, a knowledge of the concentrations of the extractives in the biomass is required.

Of course, if the presence of extractives brings error into (oven-based) moisture content calculations then, by the same logic, the use of oven-based moisture content calculations will bring error into extractives content calculations. Toluene distillation or lyophilisation are alternative methods for determining the moisture content but these will be time-consuming, and also may be somewhat difficult to use for small samples sizes. Therefore, despite the limitations of the oven-based moisture content determinations, these are still utilised in most extractives content methodologies (Ehrman, 1994b, Theander et al., 1995).

3.4.4 Accelerated Solvent Extraction (ASE)

Following a review of the available literature it was determined that, given that the focus of the project is on the biomass constituents most relevant to second generation biorefining technologies (principally the cellulosic and hemicellulosic sugars, as well as the lignin content), the methodology employed for extractives removal should remove as much as possible of the extractives (so that their influence upon the acid hydrolysis is minimal). This should be carried out in a relatively quick and effective way allowing for more samples to be prepared and analysed in the timeframe. It was decided that extraction would occur with only one solvent. This was 80% ethanol (Theander et al., 1995) in the early study on sugarcane bagasse, and 95% ethanol (Ehrman, 1994b, Sluiter et al., 2010) in subsequent experiments.

An estimation of the time involved employing these methods, however, revealed that the extraction step, using soxhlet apparatuses, would still be a limiting factor in deciding how many samples could be analysed (considering the whole analytical methodology; sample preparation, hydrolysis, etc.) in a given time period. According to the method of Ehrman (1994b), one soxhlet apparatus would be required per sample and the extraction would take place over the course of 24 hours. Following this,

significant time would also need to be spent on solvent evaporation (using a rotary evaporator). There would need to be at least two replicates in order to check for errors in the analysis. The acid hydrolysis procedure (as outlined in Section 11.5) was able to hydrolyse up to 5 samples in duplicate per day, while the sample preparation batches (as outlined in Section 11.1) could potentially operate on an even greater number of samples. Thus, for the extractives analysis not to be a sample-limiting component in the analysis-sequence it would be important to be able to prepare 5 samples a day (in duplicate, at least). This would require a minimum of 10 Soxhlet apparatus to be run for 24 hours. The difficulty and cost involved in setting up, maintaining, and safely monitoring such a set-up was considered to be prohibitive.

The use of an Accelerated Solvent Extraction (ASE) device was put forward as an alternative to soxhlet extraction in the NREL 95% ethanol procedure (Sluiter et al., 2010). An ASE 200 is shown in Figure 3-5. It is an automated system that can extract organic compounds from solid samples. The upper level of the device is a wheel that has 24 positions. In each of these a cell, containing a sample, can be placed. There are various sample cell sizes that can be used in the ASE 200 system, ranging from 1 ml to 33 ml capacity. The lower wheel houses the collection vials that are used to collect the extract. There are a total of 30 positions on this lower wheel, 26 of these can be used to collect extract from the cells while the remaining four positions can be used to house rinse vials that can collect any solvent that is used to clean the system between samples. The ASE is “accelerated” because the extraction takes place at elevated temperatures, which allows for the extraction process to be completed more quickly. In order for the solvent to stay in a liquid state during the extraction pressure is applied to the cell containing the sample.

Figure 3-6 shows a schematic for the ASE 200. The extraction process consists of the following steps (Dionex, 1999):

1. **The cell is loaded into the oven** - Once the oven is at the required temperature the upper wheel moves until the selected cell is in-front of the oven and then an arm moves out to collect the cell and bring it into the oven, at which point the oven applies pressure to seal the cell. The lower wheel also moves so that the associated collection vial is in the correct position at which point the needle assembly moves out and inserts into the collection vial.
2. **The cell is filled with the solvent** – the static valve (Figure 3-6) opens and the pump begins pumping solvent into the cell. When the cell is full and the collection vial contains about 1 ml of solvent the static valve closes and the flow stops.
3. **Heating stage** – The cell can be heated for a fixed period of time so that the sample reaches thermal equilibrium.

4. **Static cycle** – This user-defined period of time is the time that is designated for the extraction cycle. During the heating and static steps the static valve may open periodically to maintain the set pressure point in the cell.
5. **Flushing** – the static valve opens and the extract flows into the collection vial and a (user defined amount) of fresh solvent is pumped through the cell.
6. If desired the cell can be filled again and steps 4 and 5 can be repeated, with each repetition being one “static cycle”.
7. **Purging** – The purge and static valves open and the remaining solvent is displaced from the cell by the purge gas (nitrogen) so that it ends in the collection vial.
8. **End relief** – The pressure relief valve is opened and residual pressure is released from the cell.
9. **Unloading of cell** – The cell is unloaded from the oven and returned to the tray and the needle mechanism is removed from the vial.

User-defined parameters for the extraction are the length of the static cycle, the number of static cycles, the flush volume (in terms of a percentage of the capacity of the cell), the temperature, the pressure and the length of time for purging. These parameters define a **method** and a **schedule** defines a sequence of methods and the associated sample cell and collection vial that these will work on. Once a schedule has started and all the cells and vials are in place, the device can be left on its own to function without any input from the operator.

The efficiency of the ASE in solvent extraction has been compared against conventional soxhlet extraction for studies in many areas, such as in the extraction of oils from oilseeds (Dionex, 2004a) and in the extraction of lipids from soils (Wiesenberg et al., 2004). These studies have demonstrated that the ASE extraction efficiency can, in some cases, be superior (depending on the extraction pressure used in the system (Jansen et al., 2006)). Specifically regarding biomass extractives, NREL carried out a series of tests on various corn stover samples and found that the ASE method gave a similar precision and accuracy to the soxhlet method; approximately $\pm 1.5\%$ of the biomass weight over an extractives content range from approximately 4% to 25% (Sluiter et al., 2006b). Chen et al. (2010) compared ASE with soxhlet extraction for the removal of water soluble extractives from switchgrass and found that the results were comparable but that the ASE technique resulted in a seven-fold reduction in the amount of time required to prepare triplicate samples as compared to the soxhlet approach.

It was therefore decided that funding would be sought for a Dionex ASE 200 so that the lab analysts' times could be spent more effectively, and they would not be overburdened with management of a

series of soxhlet systems. The search for funding was eventually successful and a second-hand ASE 200 unit was installed in the laboratory in May 2008. It was decided that the 11 ml sample cells would be used for extraction. This size was chosen so that, in most cases: there would be sufficient quantities of sample to allow for two or three replicate extractions per sample; there would be a sufficient amount of material after extraction for the subsequent acid hydrolysis procedure (Section 11.5); that a manageable amount of ethanol would be consumed by the operation; and that no more than one collection vial would be needed to collect the extract from one sample.



Figure 3-5: A Dionex Accelerated Solvent Extractor (ASE) 200.

Initially, the determination of the extractives content of a sample was made according to the mass loss experienced after it had gone through the 95% ethanol extraction method of the ASE 200. However, at a later point a Zymark LV Turbovap evaporator was installed. This device, shown in Figure 3-7 (a), contains a heated water bath and several small gas nozzles that are designed to point, at an angle, inwards into the collection vial. The gas travels in a helical flow, as shown in Figure 3-7 (b). This sets up a vortexing action that allows for efficient extraction of the solvent and continuous rinsing of the sides of the vial. The gas, carrying with it the vapours of the solvent, exits the vial via an unobstructed path in the centre of the vial and is removed by an exhaust fan to the exhaust pipe of the unit.

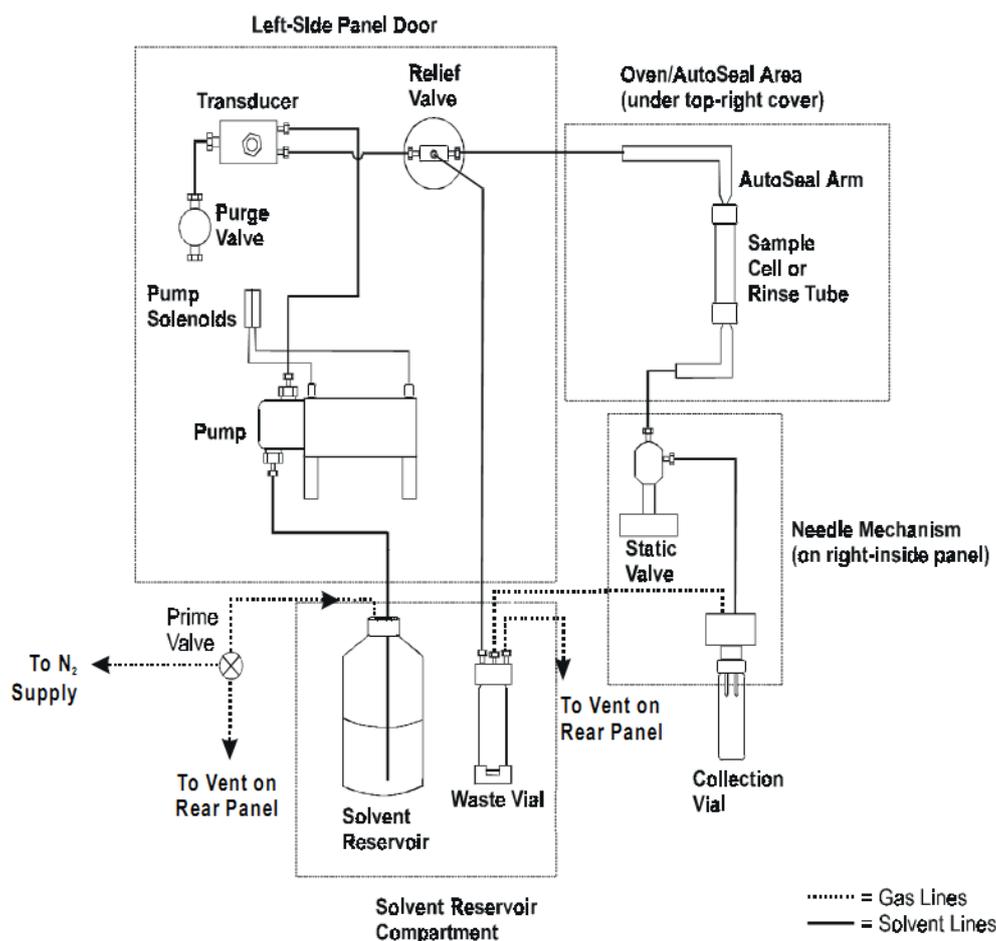


Figure 3-6: A schematic for the ASE 200. Taken from (Dionex, 1999)

Towards the end of the project a second-hand Dionex Solvent Controller add-on was purchased. This device allows up to four different solvent bottles to be connected to the ASE 200 meaning that either:

1. The same sample can be extracted with a different solvent; or
2. Different samples can be extracted with different solvents in the same sequence; or
3. Up to 4 solvents can be mixed in various proportions in a single extraction method.

The ASE/extraction methodologies employed in the research are outlined in Section 11.4.

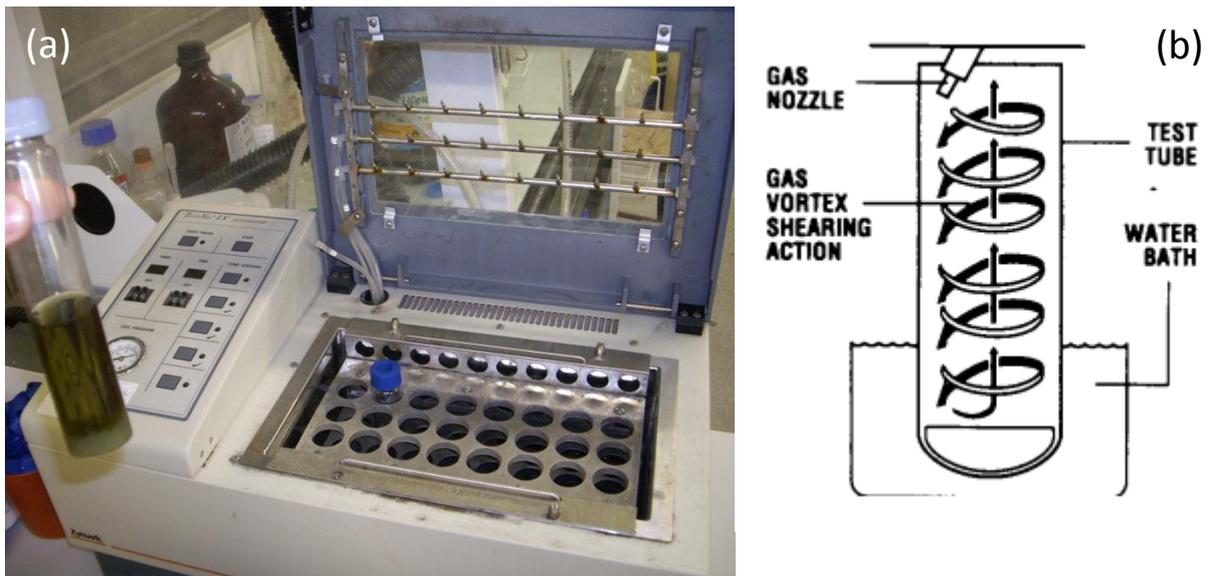


Figure 3-7: The Zymark Turbovap LV - (a) A picture of the Zymark Tubovap LV; (b) A diagram of the gas vortex shearing action within the collection vial – taken from (Zymark, 1999).

3.5 Ash Content

Most determinations of ash involve the removal of organic constituents through high temperature (usually approximately 600 °C) treatment of the biomass. That which remains is said to constitute the inorganic fraction of the biomass and is termed ash.

The procedure of Ehrman (1994a) has been adopted by ASTM as an ASTM Standard Test Method for the determination of ash in biomass. It is considered to cover the determination of ash for woods, herbaceous materials, wastepaper, acid and alkali pretreated biomass, and the solid fraction of fermentation residues. The results are reported relative to the 105 °C oven-dried weight of the sample.

Ash content of lignocellulosics may also be determined via a summative analysis where each fraction is analysed for ash (Anglès et al., 1997). There is little information about the distribution of the ashes in the extractive, holocellulosic and lignin fractions in the literature. Anglès *et al.* (1997) attempted a summative ash analysis for lignocellulosic materials with ash contents greater than 4%. They then compared this total ash with the direct ash determination and found that the summative analysis could overestimate total ash contents by up to 10% for high-ash feedstocks.

3.6 Elemental Analyses

In March 2009 an Elementar Vario EL cube elemental analyser was installed in the Carbolea laboratories. It is capable of analysing for the carbon (C), hydrogen (H), nitrogen (N), and sulphur (S) contents of a sample, as well as analysing for the oxygen content (in a separate mode).

The unit has a thermal conductivity detector (TCD). TCDs operate on the principle that the thermal conductivity properties of a gas mixture (containing the volatile gas derived from the sample, and the carrier gas) differ from that of the carrier gas alone.

The coefficient of thermal conductivity, λ (measured in $\text{cal cm}^{-1} \text{s}^{-1} \text{ } ^\circ\text{C}^{-1}$), will determine the heat flux, q_z , through a gaseous stream that spans a temperature gradient, dT/dz , as outlined in the following equation (Hinshaw, 2006). The negative sign is present because heat flows from higher to lower temperatures:

$$q_z = -\lambda \frac{dT}{dz} \quad (3.2)$$

Daily Factors:

Ultimately a peak for each of the analytes will result, and this is correlated with the absolute quantity of the element via calibrations. These calibrations relate variations in the integrated detector signal with absolute element content for samples of known weight and composition. Upon installation of the system the Author was instructed that the device had already been calibrated and would not need recalibration until the daily factors, see below, for the important constituents (C, H, N) deviated significantly from 1 and were unstable during an analytical sequence.

However, quantification does require the determination of **blank values** and **daily factor** values for each of the elements. Blank values involve running an analysis without CHNS containing material present and measuring the determined areas for each element. This can be done over several "sample" injections and the blank value taken as being the average of these areas for each element. The corresponding blanks are then subtracted from the elemental areas of the samples of interest.

Daily factor values are required to fine tune the instrument calibration to the room conditions at the time of analysis and to observe any drifts or trends in the performance of the device over time. These are calculated (after the determination of the blank values) by analysing sample(s) of known elemental composition via the following equation:

$$f = \frac{c_{theo}}{c_{act}} \quad (3.3)$$

Where f is the daily factor, c_{theo} is the reference value for the elemental concentration of the sample, and c_{act} is the elemental concentration as determined by the device upon analysis of the sample. The known sample should be analysed more than once to reduce the effects that possible human/device errors may have on the weighing or analysis of that sample and, hence, incorrect determination of the daily factor. If the factor values provided by these replicates are in close agreement then their averages can be taken and used as the factor for samples that can subsequently be analysed. For long analytical sequences, and to check for drifts in device performance over the course of a day, more than one daily-factor calculating batch can be used in a sequence. In this case the samples will use the daily factor values determined from the most recently injected batch of standard samples.

It is important that the sample that is used for daily factor calculation is stable over time and has a similar elemental composition to the samples of interest. Furthermore, as is also the case for the analysis of actual samples, the areas measured for each of the elements should be within the calibration range of the calibrations available to the device or else the accuracy and reliability of analysis may be poor.

3.7 Thermo Gravimetric Analysis

Thermo Gravimetric Analysis (TGA), where weight loss of a feedstock is monitored over increasing temperatures either in the absence or presence of air, has been shown to lead to the derivation of approximations for moisture and extractives, hemicellulose, cellulose, lignin and ash contents (Stipanovic, 2003 – personal communication).

There has been contact with the pioneers of this use of the technology. They are located at SUNY - “State of New York University, College of Environmental Science and Forestry,” Syracuse University. They carried out TGA analyses on four Irish biomass samples that were sent to them from the University of Limerick. The process of deriving carbohydrate and other mass data from the results will be explained with reference to their thermogram of a sawdust sample, Figure 3-8.

The thermogram in Figure 3-8 is the line that decreases with increasing temperature; it indicates the weight (in percent) of the sample remaining. The other line represents its first derivative (weight loss with respect to time). The losses in mass associated with certain temperature ranges are said to be indicative of the amount of various chemical components in the biomass.

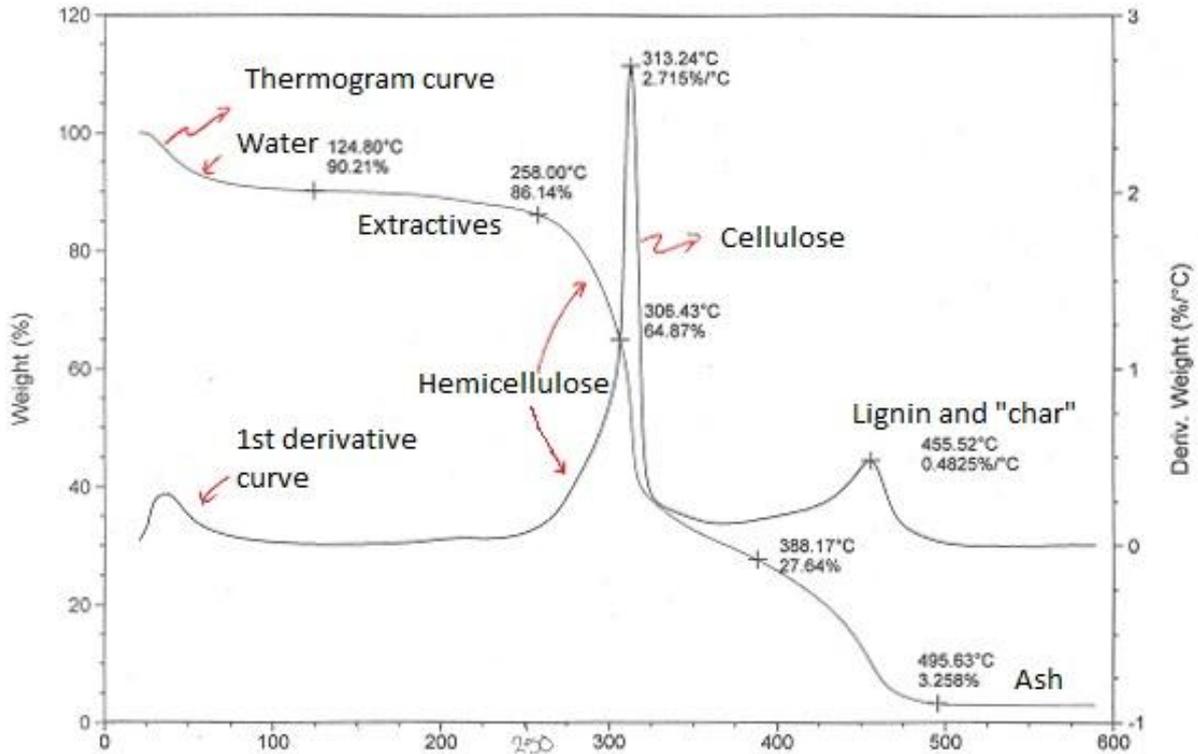


Figure 3-8: TGA of a sample of sawdust. The decreasing line represents the weight of the sample with increasing temperature. The other line is its first derivative. This TGA was taken at SUNY in a nitrogen (oxygen-free) environment.

For example, water is lost at temperatures between 50-120°C, low molecular weight compounds and other volatile organics (many extractives) evaporate from 120-250°C, hemicelluloses degrade from 250-300°C, and cellulose exhibits a very sharp degradation near 300-325°C in air (higher in nitrogen). The lignin and residual chars from cellulose and hemicellulose show a broad range of degradation temperatures leading up to 450°C, and "ash" is considered to be what remains above 500°C. A sharp drop in weight is usually indicative of an abundance of a particular component, while a slow decline indicates the presence of multiple volatile components in this region. Using the weight losses associated with the five regions, the researchers at SUNY obtained the data in Table 3-2 for the four Irish feedstocks. Drying the feedstock was sometimes necessary for analysis.

Predicting the chemical composition of composted green waste, green waste, and dried sewage sludge pellets is difficult because these are likely to contain a wide range of chemical components of various degrees of degradation. The sawdust, from the softwood Sitka spruce, should be expected to have a mass constitution of approximately 42.7% cellulose, 28.2% lignin, 25.2% hemicellulose, and 2.1% ash (FEC Consultants, 1990). In this instance, therefore, these data would seem to be in the appropriate range.

Table 3-2: TGA results for the moisture and other contents (% wet basis) of Irish feedstocks. The sludge was dried sewage pellets from Ringsend sewage treatment plant, Dublin. The (composted) green waste was from St. Anne’s Park, Limerick.

Sample	Moisture	Hemicellulose	Cellulose	Lignin	Ash
Saw dust *a	10	21	37	25	3
Composted green waste (as received)	17	-	-	-	-
Composted green waste (dried) *b	4	-	36 *c	22	36
Green waste (as received)	14	-	-	-	-
Green waste (dried) *d	4	17	34	18	26
Dried sewage sludge *e	6	33	26	16	10

*a – plus 4% extractives; *b – plus 2% extractives; *c – could contain hemicellulose as well (the peak was very broad for pure cellulose); *d – 1-2% extractives; *e – unknown peak near 200°C (7-8% of total).

The SUNY researchers provided procedural details for a work associate, Shane McArdle, to conduct the TGA analysis of several additional Irish feedstocks. It was found that the equipment at UL gave a slight broadening of the thermal degradation region and to compensate for this the cellulose “band” was extended to cover the 300-340°C range. Utilising an N₂ rather than an air atmosphere also helped to improve resolution. The data are given in Table 3-3. Several other samples, including some peats, were analysed via TGA but their thermograms had poor resolutions and these results are not presented.

It can be seen that, for the two samples that are shared between the UL and SUNY experiments, sewage sludge and sawdust, the results are comparable. However, based on the results of this Author’s work (Section 16.1) the analytical results for the Miscanthus sample appears to be somewhat inaccurate, with a higher than normal lignin content and a lower than normal hemicellulose content.

There is always the problem with TGA analysis that the mass lost between a given temperature range may not necessarily be from the perceived component of that “band”, and hence that perceived component could be overestimated. Also, the same chemical components in different feedstocks may not always be lost under the same temperature ranges. For these reasons TGA could only ever be a qualitative estimate of the amount of sugars and lignin present in a feedstock. Its main use is likely to be in screening procedures where feedstocks are firstly analysed under TGA to see if their sugar contents are sufficient to warrant more detailed (primarily wet-chemical) analysis. However, the development of NIR as an even more rapid, and potentially more accurate, analytical tool would be preferred for feedstocks which have analogues existing in the calibration.

Table 3-3: Assumed moisture, lignocellulosic and ash contents (% wb) derived from thermograms.

Sample	Moisture	Hemicellulose	Cellulose	Lignin	Ash
Miscanthus	5-6	17	36	26	2.83
Sawdust	4	11	35	37	0.4
Sewage sludge	4	21	19	15	14.7

3.8 Notes on Particle Size

Particle size can have an influence on the accuracy, precision, representability, and efficiency of analytical techniques. Methods for particle size reduction (such as milling, grinding and chipping) may have differing effects on the various anatomical fractions of a plant with the result that there will be differences in the particle size distributions of these fractions. If a subsequent analytical method (such as extractives removal) does not take a representative sample of the comminuted biomass, but instead is biased towards a certain particle size, then the relative contributions of each of the anatomical fractions to that sample will also not be representative of the sample as a whole. This discrepancy in sampling will lead to an inaccurate (but not necessarily imprecise) result. The problem will increase as the amount of sample taken for the analysis decreases. Therefore care needs to be taken to ensure that sample preparation and subsample collection (for analysis) is representative and homogeneous.

The effect of particle size on the efficiency of analytical methods has also been demonstrated in the case of solvent extraction (Cacace and Mazza, 2003) and acid hydrolysis (Hoebler et al., 1989).

For these reasons most analytical methodologies outline a maximum (and sometimes a minimum) particle size. For example, the Uppsala Method (Theander et al., 1995) suggests that samples should be ground in a cyclone mill until the particle size is less than 0.5 mm, and the NREL protocols (Sluiter et al., 2010) say that only samples with a particle size greater than 180 microns and less than 850 microns can be used for acid hydrolysis experiments. The reasoning behind this particle size range is that, for particle sizes over the critical limit, the volume to surface area ratio would be too high, meaning that the acid would not penetrate through the whole particle and so would not hydrolyse it effectively in the hour timeframe of the primary hydrolysis stage. That would result in less of the polysaccharides being hydrolysed. If polysaccharides were still solid after the hydrolysis procedure then these would be incorrectly classified as part of the KL. Hence, KL would be overestimated and cellulose/hemicellulose underestimated. For particle sizes below the limit the reasoning was that the volume to surface area ratio would be too low meaning that the particle might be degraded too severely in the hydrolysis procedure. That could result in the liberated sugars being converted to sugar degradation products such as hydroxymethylfurfural and furfural and, hence, the

cellulose/hemicellulose contents could again be underestimated. Should the sugar degradation products condense and polymerise solid residues or tars could potentially be formed and these may be incorrectly classified as part of the KL. Hence, KL may be overestimated.

3.9 Summary

This chapter has summarised the literature review that the Author carried out in determining what analytical protocols to use for the analysis of potential biorefining feedstocks. It has also put forward the reasons why the analysis methods outlined in Section 11 have been chosen.

A vast amount of time could potentially be spent characterising in fine detail all of the major and minor components of biomass samples. However, an intelligent use of time would be to focus only on those properties of biomass that are of greatest importance to the subject matter. Biorefining technologies function on the basis that saleable products can be obtained from the majority of the feedstock. Hence, the principal components of relevance in most lignocellulosic feedstocks will be cellulose, hemicellulose, and lignin.

This chapter has outlined the various methods that are used for the characterisation of these three polymers, and the differences between the gravimetric and hydrolysis methods have been compared. It was determined that, while the gravimetric methods may be easier and quicker to undertake (thereby allowing more samples to be analysed), the accuracies and levels of detail they provide are poor. In particular, the inability to resolve between the different hemicellulosic sugars is a major flaw. Furthermore, the uncertainties about what is being assumed as a certain component are considered too great when gravimetric methods are used. That will apply when comparisons are being made between the compositions of different feedstocks, between different anatomical fractions of the plant, or between samples from different seasons. Hence, the confidence in NIR calibration equations that may cover several feedstocks, different anatomical fractions, etc. would be poor.

The decision was therefore made that hydrolysis methods that would allow direct analysis of the polysaccharide sugars would be employed. A review of the literature indicated that the Uppsala method, and its variants, such as the NREL protocols, were the most prevalent and well accepted by the scientific community. The ion chromatography equipment that was ultimately chosen and validated for monosaccharides analysis is outlined in Section 4.

The solid residue that remains after the hydrolysis method is applied is termed the Klason lignin. It can be summed with the ASL content to determine the total lignin content. However, both of these methods are indirect analyses – the KL analysis is gravimetric and the ASL analysis is spectroscopic. There are many potential interfering UV-absorbing compounds that may also be present in the hydrolysate. There are methods, some very labour intensive, for the more accurate determinations of these components. However, employing these is not considered to be a sensible use of time given that the main role of lignin in the biorefining technologies is to act as a fuel for the process. Furthermore, given that KL components are those that are resistant to acid hydrolysis, its content may provide the most relevant lignin data for lignocellulose technologies, such as DIBANET, that involve the acid treatment of biomass.

The relatively low content of ASL in most feedstocks and its uncertain fate/value in biorefining processes caused the Author to decide that excessive time should not be spent on the analysis of this component and to use the standard spectroscopic method. The situation is similar regarding the analysis of UA, except that in the case of these a separate wet chemical procedure is required to liberate the monomers prior to their spectroscopic analysis. That would mean that much more time would be required to incorporate their analysis as standard for all samples. It was therefore decided that only a limited number of samples would be analysed for their UA contents.

Extractives are much more important than ASL and UA in the analysis of lignocellulosic samples. However, this importance is not primarily based on their value to biorefining technologies but instead it is based on the potential errors that these components could impose, if not removed, on analysis of the components of the lignocellulosic matrix. Again, there are a plethora of methods that can be applied for removal of extractives. These are mostly based on the solvent, or sequence of solvents, used to extract the sample. Once again, finding a reliable but efficient method was key, and a review of the literature suggested that the one-stage extraction of samples using 95% ethanol was one of the most effective methods for removing potentially problematic components prior to the acid hydrolysis step. The requirement for a series of soxhlet apparatuses to run for 24 hours was a potentially critically limiting factor (that could result in most of the analyst's time being spent on this step), reducing the number of samples for which lignocellulosic data could be obtained. Fortunately, the Author managed to obtain funding for an accelerated extraction system that allowed a massive increase in throughput.

As described in a previous paper (Hayes, 2008), biorefining technologies may not involve the hydrolysis of biomass but instead operate via thermochemical mechanisms (in the cases of gasification and pyrolysis technologies). In the thermochemical processes the constitution of the

lignocellulosic matrix may be less important than the elemental composition of the sample. Therefore, an elemental analyser is used for determining this composition for selected samples. From these data estimates of the various heating value statistics of the samples can be calculated.

Since accuracy and precision are both important in the analysis of samples, particle size needs to be considered. The Author decided, based on the literature, that two particle size fractions would be collected from each sample for potential analysis. The most important fraction would be between 180 and 850 microns since this presents the most suitable size for accurate hydrolysis (as suggested by NREL). The other fraction would include particles with a diameter less than 180 microns. Since the sample preparation methods (Section 11.1) would be targeted to maximise the larger fraction, this smaller fraction would not be analysed as standard. However analyses were undertaken for some samples in order to determine any trends that existed between the two fractions.

In summary, the primary focus for obtaining the most accurate analytical results is on the component monosaccharides of the structural polysaccharides. Ideally the polysaccharides could be analysed and hydrolysed separately from each other so that more detail could be learned concerning their structures. Conclusions could be drawn from these results concerning the most efficient means to obtain high value platform chemicals from them in biorefining processes. Unfortunately, the work-load involved in doing this was considered to be too great to be undertaken in this study. In late 2009 the Author was visited by colleagues from the University in Cali in Columbia. They spent two weeks at the Carbolea laboratories during which time they attempted to replicate, for a *Miscanthus* sample, their sugarcane-bagasse procedure for the separation and characterisation of hemicellulose. This task could not be completed during the visiting period and the Author did not have the time needed to repeat the experiments. The Author would like to attempt to repeat this study at a later date and it could be a research area of high impact because there are currently no publications concerning the analysis of the hemicelluloses in *Miscanthus* samples.

The extraction and hydrolysis methods described in this section, and the ion chromatography method described in Section 4, are of most relevance for the development of quantitative NIR calibrations, the focus of this Thesis. The Author considers that the methods selected should allow for accurate calibrations for the components of most importance to biorefining technologies.

4 Ion Chromatography (IC)

In researching for chromatographic equipment that could analyse monosaccharides with precision, accuracy and sensitivity, without the need for derivatisation or other lengthy preparations, the Author found that ion chromatography systems, and in particular high performance anion-exchange chromatography coupled with pulsed amperometric detection (HPAEC-PAD), offered many advantages.

4.1 Background to IC

Ion chromatography (IC) is a type of liquid chromatography; the sample is introduced to the system as a liquid, the mobile phase is liquid and the stationary phase is a solid material. In IC, separation occurs due to differing coulombic (ionic) interactions between analyte molecules and the stationary phase. These interactions involve the reversible adsorption of charged solute molecules to the immobilised ion exchange groups of opposite charge on the stationary phase (Behan and Smith, 2011). IC can be termed HPAEC (anion exchange chromatography) when anions are to be resolved and HPCEC (cation exchange chromatography) when cations are analysed. Hence, the stationary phase has ionic functional groups R-X, where X is a cation (X^+) in HPAEC and an anion (X^-) in HPCEC. This X is associated with mobile counterions that can be reversibly exchanged with other ions of the same charge without altering the matrix (Fritz and Gjerde, 1995). For example, anionic carbohydrates that are held on the ion exchange sites can be replaced with the hydroxide ions of the mobile phase. X can be different groups with different ion exchange capabilities and the total number of charged groups will determine the capacity of the column. In many HPAEC columns the active site consists of a quaternary ammonium ion ($-N^+R_3$) and HPCEC columns tend to use sulphonate groups. The capacity of columns is expressed in milliequivalents of exchangeable ion per gram of resin, with a lower capacity meaning that a lower eluent concentration will be needed to push the sample along the column. Most commercial columns tend to be of a low capacity (around 0.01 to 0.2 mequiv/g) (Behan and Smith, 2011). The columns tend to be made up of a polymer-based pellicular resin with resins based on styrene-divinylbenzene co-polymers being the most widely used ion exchangers, Figure 4-1 (a) shows an illustration of such a copolymer (Fritz and Gjerde, 1995). The function of the divinylbenzene (DVB) is to cross-link the linear chain of the polystyrene, an effect that increases the mechanical stability and decreases the solubility of the polymer. The

cross-linking does not inhibit ion-exchange because there are pores and channels that the ions can go through (Fritz and Gjerde, 1995). The nitrogen atom of the functional group is usually connected to the benzene by a single -CH₂ group but the number of groups (known as the spacer arm) between the two can be varied which may have differing effects for different analytes, facilitating chromatographic separation (Fritz and Gjerde, 1995). Figure 4-1 (b) shows an example of an anion-exchange group that can be attached to the styrene-divinylbenzene copolymer.

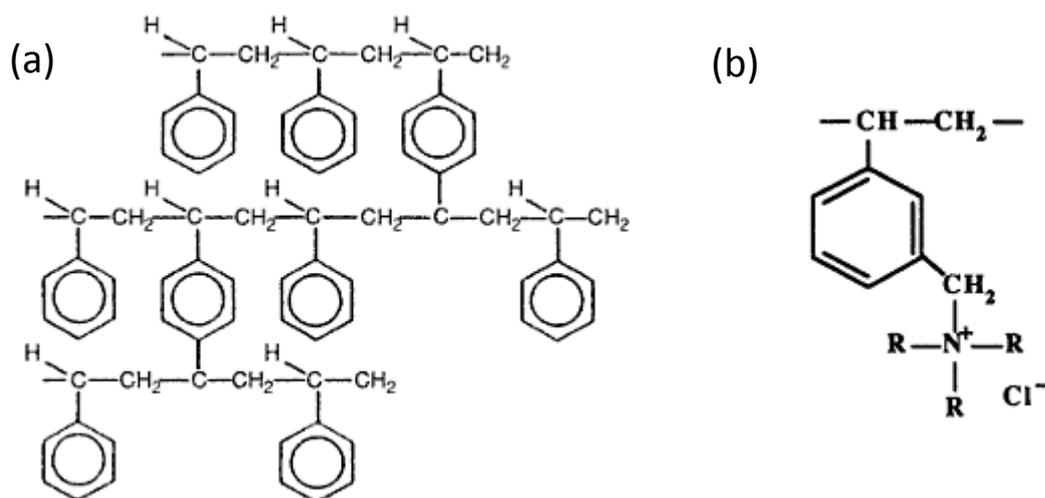


Figure 4-1: Polymers used in IC columns. (a) A styrene-divinylbenzene copolymer. Taken from (Fritz and Gjerde, 1995); (b) The functionalisation of a part of this polymer to enable anion-exchange. Adapted from (Haddad and Jackson, 2003).

A somewhat different form of ion-exchanger, and one used extensively in Dionex (principally carbohydrate-specific) columns such as the PA1 (see Section 4.6.1), involves the stationary phase being on a layer on the outside of a spherical substrate (which is typically polystyrene-DVB). The polymer can be sulphonated at the surface and the outside is then coated with a layer of latex particles that have been functionalised with quaternary ammonium groups (Behan and Smith, 2011). The result is that the positively charged latex particles are held electrostatically to the sulphonated substrate so strongly that even 4M NaOH cannot break the link (Fritz and Gjerde, 1995). In such systems the ion exchange capacity will be dependent on the size of the pellicular substrate, the abundance of latex beads that are coating it, and the size of these beads. Variations in the type of and cross-linking in the polymeric substrate, in the type of functional groups attached to the latex bead, and in the degree of latex cross-linking can aid in selectivity for anions (Fritz and Gjerde, 1995). Figure 4-2 shows the associations between the aminated latex particles and the sulphonated core polymer particle.

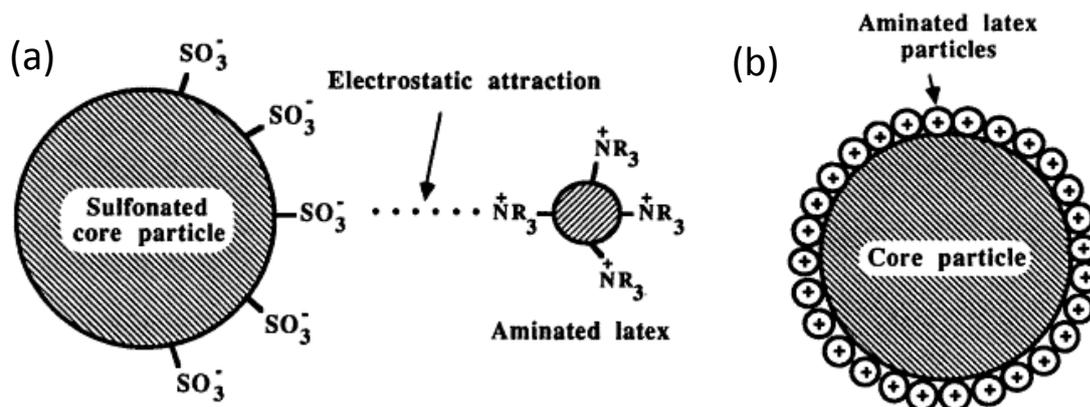


Figure 4-2: Electrostatic binding in anion exchange resins. (a) The formulation of an agglomerated anion-exchange resin using electrostatic binding; (b) A schematic of an agglomerated ion exchanger. Both adapted from (Haddad and Jackson, 2003)

4.2 IC for Carbohydrates

Carbohydrates can interact with the stationary phase because they are weak acids with pKa's above 11. The use of sodium hydroxide as an eluent promotes ionisation of carbohydrates to their anionic form (Dionex, 2003). Figure 4-3 shows an example of an oxyanion.

In studies where the hydroxyl in the 1 position was O-methylated it was found that the compound was poorly retained, while methylation of other hydroxyl groups had less of an effect (Koizumi et al., 1992). This indicated that the anomeric hydroxyl had the greatest acidity, attributable to the inductive effect of the ring oxygen. These studies also indicated that the order of acidity of the potentially ionisable hydroxyl groups of a monosaccharide is as follows 1-OH > 2-OH ≥ 6-OH > 3-OH > 4-OH (Cataldi et al., 2000).



Figure 4-3: An example of an oxyanion of mannose. Taken from (Behan and Smith, 2011)

HPAEC exploits the small differences in the pKa's of the OH groups of carbohydrates. Typically sodium hydroxide is used as the mobile phase for the elution of carbohydrates; however, there are papers

that describe the use of sodium acetate as an additional pushing agent, as described by Lee (1996). When employing a mobile phase with a 0.1 M sodium hydroxide concentration, the separation of some relevant isomeric monosaccharides, for instance galactose, glucose and mannose, cannot be achieved, as they exhibit very similar retention behaviour. On decreasing the OH⁻ concentration to a value lower than 20 mM, these compounds are more likely to interact with the stationary phase, thus better demonstrating their differences in ion-exchange behaviour. However, in order to separate xylose and mannose [OH⁻] concentrations as low as 2 mM are needed, or alternatively water can be used as the eluent (Cataldi et al., 2000).

4.3 Detection

There are various types of detectors that can be used in ion chromatography systems. At the University of Limerick our Dionex ICS-3000 system has two electrochemical detectors, a conductivity detector and a, recently purchased, diode array UV-Vis detector. Since only the electrochemical detector was used for the work presented in this Thesis, only this will be discussed.

4.3.1 Electrochemical Detection

The electrochemical detector (ED) operates with pulsed potentials (as shown in Figure 4-4). The pulsed amperometric detector (PAD) allows for sub nmol range sensitivity detection of carbohydrates without derivatisation (Lee, 1996). An amperometric detector measures current - as opposed to a potentiometric detector, which measures voltage, and a conductometric detector, which measures resistance.

The electrocatalytic oxidation of sugars occurs when they make contact with the gold electrode resulting in an electrical current that is linked to their concentration; however, unless these oxides are removed the detector response will drop over time. Hence pulsed potentials are necessary to remove these oxides from the electrode (the increased potential results in the formation of gold oxide at the electrode surface, removing the sugars in the process (Behan and Smith, 2011)). Following this, the potential is lowered and the coating is removed as a result of the gold oxide being reduced. Figure 4-4 (a) shows a three pulse ED set-up. The first pulse, E1, is the pulse for the oxidation of the analytes of interest. It should be noted that the step from one potential to another

results in a charging current that is not part of the current analyte oxidation current; hence the analyte oxidation current is measured after a delay that allows the charging current to decay.

The oxidation current is then measured by integrating the cell current after the delay – current integrated over time is charge (measured in coulombs) or the average current during the integration period can be presented (units of amperes). The second pulse, E2, causes full oxidation of the species on the electrode surface to their most soluble forms meaning that these oxides will be washed away (Behan and Smith, 2011). The final pulse, E3, then applies the reducing potential to regenerate the electrode surface. It has been shown, however, that with a standard three-potential waveform the peak areas for the analytes will decrease over time due to the recession of the gold electrode. This results from the dissolution of the gold that can occur at the high positive potential (Rocklin et al., 1998). These authors proposed a four-potential waveform whereby, rather than using a large positive potential to flush out the carbohydrate oxidation products, a large negative potential is used for a short time (10 milliseconds) prior to the (short) application of the gold oxidation potential. This waveform, which is of a shorter duration than the waveform in Figure 4-4 (a) is shown in Figure 4-4 (b).

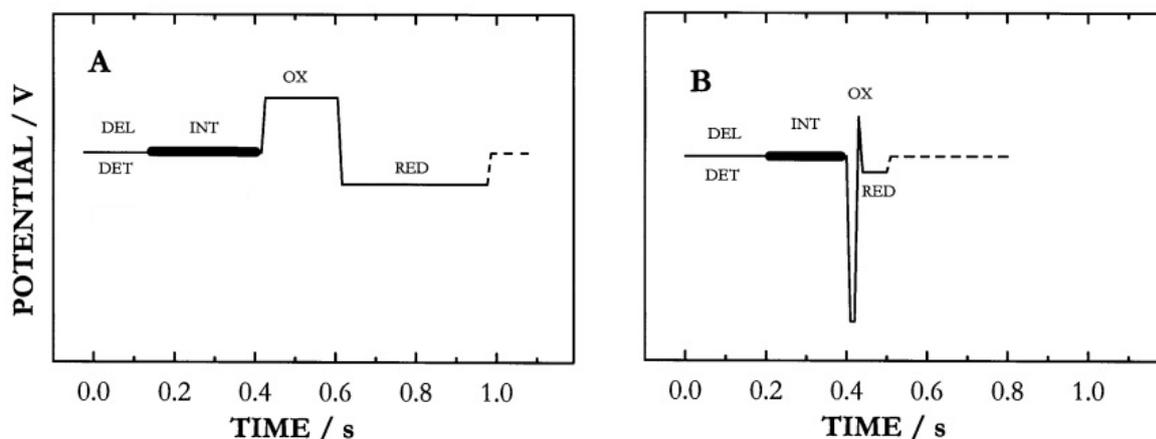


Figure 4-4: Two typical waveforms for PAD in alkaline solutions at a gold working electrode; (A) Standard three-potential steps waveform; (B) Four potential-steps waveform; DEL = delay time; INT = integration time; DET = detection potential; OX = oxidation potential; RED = reduction potential. Adapted from (Cataldi et al., 2000).

PAD will only detect the compounds that contain functional groups that are oxidisable at the detection voltage applied and neutral or cationic components in the injected sample will elute in approximately the void volume of the column meaning that, even if they were oxidisable, they would not interfere with the analysis of the carbohydrates (Dionex, 2000). Hence PAD detection can be very selective for the analytes of interest. However, care must be taken that the right potentials

are selected – for example the formation of gold oxide at the electrode will inhibit the oxidation of carbohydrates. Hence a measuring potential that is less than that required for gold oxidation should be employed (Dionex, 2000).

A pH of greater than 12 is required for the oxidation of carbohydrates to be favoured at the working electrode and the sensitivity and stability of the detection will be improved the more alkaline the mobile phase presented to the detector (Dionex, 2003). However, the potentials at which redox reactions will occur on the electrodes are pH dependent with the potential shifting -0.059 V per pH unit (Dionex, 2006). Therefore, if the hydroxide concentration of the mobile phase is changed there will be an effect on detector response, so it has been recommended that a solution of hydroxide could be added post-column in order to reduce the effect of any pH shift in the (pre-column) mobile phase (Dionex, 2006).

4.4 The Davis (1998) Paper

A wide literature review was conducted by the Author in order to determine what the ideal conditions would be for the separation of the monosaccharides liberated upon the sulphuric-acid-catalysed hydrolysis of lignocellulosic biomass. The target of this search was to find a procedure that was accurate, precise, and stable over long chromatographic runs but did not require extensive sample or eluent preparation procedures. It was found that there were many examples in the literature of difficulties, using dedicated carbohydrate columns such as the Dionex PA1, in resolving arabinose, galactose and (particularly) rhamnose while also being able to resolve xylose and avoiding tailing in the mannose peak (e.g. (Worrall and Anderson, 1993)). Many papers did not even consider the separation of rhamnose. That would mean that it would most likely contribute to either the glucose, arabinose or galactose peaks. The author considered the separation of rhamnose important, particularly given the planned work on the hydrolysates of peat samples (Section 13).

An extremely relevant paper by Davis (1998) appeared to solve many of the issues outlined above. It described a HPAEC-PAD protocol that had been used, for over 4000 samples, at the Forest Products laboratory (FPL) of the US Department of Agriculture Forest Service. The solution to the resolution, in a single injection, of fucose, arabinose, galactose, rhamnose, glucose, xylose and mannose on the Dionex PA-1 column was provided in the form of an acetate loading method. This method consisted of sugar elution with water for 11 minutes, followed by a minute ramp to 170 mM CH₃COONa in 200 mM NaOH, which was maintained for 5 minutes (column conditioning step), then a 1 minute return

to the water elution which continued for 9.5 minutes (equilibration step). The flow rate was 1.2 ml/min and the temperature was maintained at 22°C. Since only water was used for elution of the sugars it was necessary to include the post-column addition of base (300 mM at a flow rate of 0.3 ml/min).

The column conditioning step was considered necessary in order to elute any strongly retained compounds, such as uronic acids, that may have been in the hydrolysate. The acetate loading allowed for shorter retention times (all sugars were detected in less than 10 minutes) and this effect appeared to be different (less pronounced) for rhamnose compared to the other sugars of interest. This observation allowed Davis to fine-tune the acetate concentration in order to allow for the elution of rhamnose between galactose and glucose.

In contrast to advice from Dionex (Dionex, 2000) and previous studies (e.g. (Sullivan and Douek, 1994)), no neutralisation of the acidity of the sample or removal of the sulphate ions present in it (as a result of the hydrolysis with sulphuric acid) took place. The authors found that retention times were reduced and that the peaks for the sugars actually became sharper with an increase in the sulphate load of the sample meaning that resolution factors were slightly improved with 5 µl injections of 4% acid solutions compared with aqueous samples (Davis, 1998). It was found that the sulphate is not completely removed by the conditioning step. That means that the sulphate load and retention time changes that result from it required several injections in order to reach new steady state values. This meant that the data for the first few injections onto the system should not be used.

Solid phase extraction (SPE) is often considered necessary (Sullivan and Douek, 1994) (Worrall and Anderson, 1993, Dionex, 2000) in order to remove hydrophobic products, such as acid soluble lignin, from the sample since these could otherwise contaminate the analytical column and result in effects such as peak broadening and the associated loss of sugar resolution. This is generally an off-line step that occurs prior to sample injection. For example, an SPE cartridge can be attached to a sample syringe, the best extractant for hydrophobic organic material being solid polystyrene-divinylbenzene polymers or polyvinylpyrrolidone (Fritz and Gjerde, 1995). Davis (1998), however, proposed an on-line SPE system involving a Dionex NG1 guard column placed before the guard of the analytical column. The sample would be injected and pass through the NG1 which would retain any hydrophobic products meaning that these would not be eluted any further through the system. The method involved the use of a diverter valve that, one minute after sample injection, diverted the eluent flow away from the NG1 and directly to the analytical column. That would mean that the NG1 was only part of the eluent flow during the time of sample injection.

The NG1 guard column was washed with methanol and then re-equilibrated with water after, approximately, every 100 injections. In an experiment with the UV-detection of the methanol-wash from the NG1 it was demonstrated that UV-absorbing material was indeed retained by this column in its normal operation as a pre-PA1 guard (Davis, 1998). Davis demonstrated that the response factors of sugar standards in the presence and absence of the in-line NG1 were indistinguishable.

The use of the NG1 column as an additional guard column to remove the sample matrix and contaminants was more recently put forward in an Application Brief by Dionex (Dionex, 2009). In this system two NG1 columns were placed in loops connected to various ports on a 10-port switching valve. Depending on the position of the valve (A or B) one column would be in-line with the analytical column while the other was being back-flushed with, firstly, a 20% water and 80% acetonitrile solution, with the acetonitrile for cleaning and, latterly, a NaOH solution for equilibration. Switching the valve to the opposite position after the previous sample had been analysed would allow the next sample to pass through a freshly-cleaned NG1 while the previously used NG1 would be back-flushed. This configuration is illustrated in Figure 4-6, Section 4.6.2.

Due to the extended linear range of the detector, Davis used single-point calibration involving the intermittent injection (every 6 to eight injections) of a sugar standard which was used to calculate the relative response factors of all the sugars of interest compared with the internal standard fucose. He found that, with 5 μ l of such sugars standards, linearity was established for up to 1.94 mg/ml for glucose and 1.42 mg/ml for xylose.

Davis also found that the detector noise was normally under 4 nAmps meaning that the detection limits for 5 μ l injections of arabinose and mannose were 0.57 mg/L and 1.20 mg/L, respectively (i.e. 0.57 and 1.20 ppm). The repeatability and stability of the Davis method was high; he found that his procedure allowed sequences of over one hundred injections to be carried out with no loss in precision.

4.5 Important Statistics for IC

Some of the key points in chromatography are: that the analytes are well resolved; that the relationships between peak area/height and the concentrations of the analytes are well understood; and that the conditions/results are reproducible, precise, and stable over long periods. Peak resolution is improved if the peaks are sharp with no significant tailing or fronting demonstrated.

Some of the important statistics and quality parameters that can be used in the Chromeleon 6.80 software that is integrated with the Dionex ICS-3000 present in the Carbolea labs are listed below:

Width:

Various width statistics are available. The default one in Chromeleon is the peak width at baseline (BW). This is calculated by drawing tangents at the turning points of the leading and trailing edges of the peak and determining their point of intersection with the baseline, the distance (in time) between the two intersection points is BW. The software can also determine the peak width at 5%, 10%, and 50% height over the baseline. In these cases the peak width is measured as the distances between both sides of the peak at these respective heights. Left and right widths can also be determined; these are calculated by dropping a perpendicular line from the peak maximum to the baseline and measuring the distance between this line and the respective peak ends (depending on how the width is calculated, BW, $W_{50\%}$ etc.).

Asymmetry:

This statistic can provide a measure of peak fronting or tailing and so help to evaluate the performance of a column or set of conditions. The European Pharmacopeia (EP) standard formula is:

$$A = \frac{RW_{5\%} + LW_{5\%}}{2 * LW_{5\%}} \quad (4.1)$$

If there is no asymmetry then A will be equal to 1, while for tailing peaks A will be between 1.2 and 5 (Dionex, 2010).

If the asymmetry is calculated at 10% height instead of 5% height then the value is referred to as "skewness".

Retention Time:

The retention time, t , of a peak is defined as the time, from injection (where time = 0) until the peak maximum.

Dead Time:

The dead time, t_0 , is the time for the peak maximum of a non-sorbed marker, i.e. a substance that does not interact with the stationary phase, to reach the detector. If t_0 is known then the **adjusted retention time**, t' , of each analyte can be determined as $t - t_0$.

Resolution:

This is a very important statistic and is used to measure the separation between two peaks. It is determined via the following formula in the EP standard:

$$R = 1.18 \left| \frac{t_{RefPeak} - t_R}{W_{50\%,RefPeak} + W_{50\%,R}} \right| \quad (4.2)$$

Where t_R = retention time of the current peak; $t_{RefPeak}$ = retention time of the reference peak for the resolution; $W_{50\%,R}$ and $W_{50\%,RefPeak}$ = Widths of the two peaks at 50% of peak height. The next peak in the chromatogram can be selected as the reference peak (this is the default) or the previous peak can be used.

Capacity Factor:

The retention/capacity factor, k , is defined as the ratio of the amount of the analyte in the column stationary phases compared with the amount of analyte in the mobile phase but it is usually calculated based on the ratio of the adjusted retention time to the dead time:

$$k = (t - t_0) / t_0 \quad (4.3)$$

The differences in the k values between analytes need to be sufficient for good separation although k values should not be too large or chromatographic runs will become too long. A k range of 2 to 10 may be considered desirable (Fritz and Gjerde, 1995).

Height: The height from the chromatogram at the retention time to the baseline.

Signal to Noise Ratio:

This statistic is used to calculate the quantification accuracy of analytes and is calculated according to the formula below:

$$S/N = 2 \times \frac{Height}{Noise} \quad (4.4)$$

The noise is considered to be the baseline variation (height) that is attributable to noise rather than the presence of analytes. It can be calculated for the current sample by taking a peak-free interval to the left and/or right of the current peak with the larger of the left/right noise value being used; however, there may be cases where a peak-free time interval of suitable width cannot be found meaning that the S/N ratio cannot be determined. Alternatively the noise can be calculated from a

blank injection using the same column/detector conditions and then the time intervals that are located to the left and right of where the peak would be expected are used for calculating the noise.

Detection Limit:

Davis (1998) calculated the detection limit (LD) for each analyte as the system noise (determined as three times the SD around the baseline in a peak-free area of the chromatogram) divided by the ratio of peak/height to analyte concentration:

$$LD = \frac{Noise}{(Height/Concentration)} \quad (4.5)$$

Theoretical Plates Number:

This statistic is the measure of the separation capability of a column. It is calculated, using the EP method, according to the following formula:

$$TP = 5.54 \times \left(\frac{t}{W_{50\%}} \right)^2 \quad (4.6)$$

The greater the (unitless) TP value the greater is the separation power of the column. The statistic can be used to determine the **height equivalent of a theoretical plate** which is expressed in meters and calculated as (Fritz and Gjerde, 1995):

$$H = L/TP \quad (4.7)$$

Where L is the column length, in metres. However the reciprocal of this height statistic is more useful in providing an indication of the quality of the column, with a statistic for the number of theoretical plates per metre.

Peak Area:

This is defined as the area between the signal curve, the baseline, and the peak delimiters (see below for how peak limits are resolved).

Selectivity Coefficient:

This statistic can be used to describe the relative affinity for a column resin between two ions. The determination of this constant for ions A and B is determined by (Fritz and Gjerde, 1995):

$$K_B^A = \frac{[A]_r[B]_s}{[B]_r[A]_s} \quad (4.8)$$

Where the brackets indicate the ion concentration (in mmol/mL for the solution phase and mmol/g for the resin phase) and K_B^A is the equilibrium constant for the following equation (with subscript r denoting the resin-phase interaction of the analyte and subscript s denoting solution phase):



4.5.1 Methods for Peak Determination

The methods by which the peaks are delimited and the baseline is determined are of key importance in determining what the results will be for peak, height, width, area, resolution and other key statistics. In the Chromeleon software, in addition to rider peaks (which never occurred for the analytes of interest in this study), the following peak types are possible where there is no full baseline resolution between peaks.

- Baseline-Main-Baseline – The peak is integrated as a main (i.e. non-rider) peak with bilateral baseline contact on the left and right sides of the peak.
- Main – The peak is integrated as a main peak but there is no baseline contact with the peak. The peak delimiters are dropped down, to the baseline, from the local minima of the peak at either side.
- Baseline-Main – The peak is integrated as a main peak with baseline contact on the left side of the peak but not on the right.
- Main-Baseline – The peak is integrated as a main peak with baseline contact on the right side of the peak but not on the left.

These peak type definitions are abbreviated (e.g. Main is M, Baseline-Main-Baseline is BMB, etc.) with a further distinction made between B (where the main peak has direct baseline contact) and b (where the baseline point is defined as a local minimum). Examples are provided in Figure 4-5.

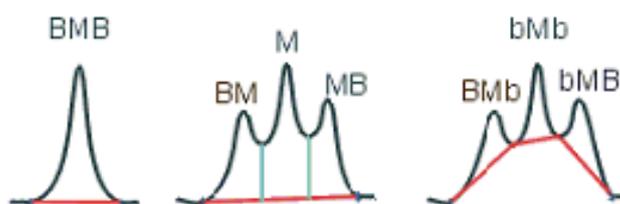


Figure 4-5: Some of the various peak types possible in Chromeleon. Adapted from (Dionex, 2010)

The software can be set to automatically determine what peak types to attribute to a sample or peak detection parameters can be set-up. These can specify which peak type is provided for each analyte.

4.6 Set-up of an Ion Chromatography System at UL

4.6.1 Purchase and Specifications of the System

In January 2008 the Author found that funding could be available to purchase, for the Carbolea laboratories, a chromatography system for the analysis of lignocellulosic biomass and, potentially, for the analysis of the products of any biorefining processes that may be developed at the Research Centre. Based on the Author's literature review of the state of the art it was decided that an IC system would be the most attractive of the options based on the criteria of: accuracy, precision, minimal sample preparation necessary; and flexibility to a variety of analytes and biomass/hydrolysate matrices. The Author then had to determine the build/capacity of the system since there are a wide variety of potential configurations. It was determined that the best system to would be the Dionex ICS-3000 and the system was installed on May 1st 2008. Some of the specifications of the system are listed below:

- Pumps: A Dual pump (DP) system. This incorporates two gradient pumps that can each deliver from up to four eluent bottles with flow rates from 0.001 to 10.0 ml/min and operating pressures up to 5000 psi.
- Two injection valves, space for at least two columns and their guards, and two slots for detectors, meaning that the system can run completely separate analytical methods (timebases).
- An autosampler (AS50) capable of accommodating 100 small (1.5 mL) vials. This autosampler includes a divertor valve that can choose which of the two injector valves to send the sample to.
- A conductivity detector with temperature control and detection range up to 15,000 μ S.
- A self-regenerating suppressor (ASRS 300)
- Two (PAD) electrochemical detectors – one had a detector cell with a permanent gold working electrode and the other detector cell accommodated disposable working electrodes.
- The detectors also allow the collection of 3D amperometry data – these data cover the detector response at a particular point in the waveform (not just the integration interval)

meaning that an extra dimension is added to a chromatogram. To date this option has not been used at UL.

- An automation manager (AM) that included a high pressure switching valve (for the integration of the NG1 pre-column guard in the system).
- Temperature control for the column (10-70°C) and detector compartments (15-40°C).

Also supplied with the system was: a PA20 column and guard for the separation of carbohydrates; an AS11 column with guard for the separation of organic acids; and two NG1 guard columns.

The NG1 guard column is usually used with the Dionex IonPac NS1 analytical column. The packing material of the NG1 is described as a macroporous ethylvinylbenzene polymer cross-linked with 55% divinylbenzene, having a 10µm particle diameter and a very high hydrophobic surface area. It is resistant to organic solvents and eluents from pH 0 to 14 (Dionex, 2004b).

The PA20 column is designed for the separation of glycoprotein monosaccharides. The column has a hydrophobic 6 µm diameter polystyrene-DVB (55% cross-linking) substrate agglomerated with 130 nm MicroBead quaternary ammonium functionalised latex (5% cross linking). It has an ion exchange capacity of 65 µeq per 4 x 250 mm column, is compatible over a pH range of 0 to 14, can sustain a maximum backpressure of 3500 psi, and is compatible with organic solvent eluents up to 100% (Dionex, 2005b).

The provision of the PA20 column was an error by the supplier since the Author made it clear that a PA1 was necessary for the separation of all the sugars of interest (for example by the Davis (1998) method). Section 4.6.2.1 illustrates the, unsuccessful, attempts that were made to resolve the important monosaccharides with the PA20 column. A few weeks after the installation of the system the supplier provided the PA1 column requested by the Author.

The PA1 column has a hydrophobic pellicular 10 µm diameter polystyrene-divinylbenzene (2% cross-linking) substrate agglomerated with 580 nm MicroBead quaternary ammonium functionalised latex. The MicroBead latex is 5% cross-linked. It has an ion exchange capacity of 100 µeq per 4 x 250 mm column, is compatible over a pH range of 0 to 14, can sustain a maximum backpressure of 4000 psi, but can only tolerate up to 2% organic solvent (Dionex, 2005a).

At a later point a second-hand Dionex GP40 gradient (up to four eluents) pump, with degas, was purchased to function as the post-column pump for the saccharides analysis method (Section 4.6.2.2). This would allow the second pump of the DP system to be used for the organic acids method. Some months after this another second-hand GP40, with no-degas, was purchased and

linked to the “B” configuration of the switching valve. This meant that back-flushing of the NG1 column to remove any sorbed hydrophobic material could take place without having to reconfigure the system.

In the last few months, while the Author has been writing up, a photodiode array detector (Dionex DAD-3000) has been purchased for the system. Its use will primarily be for the analysis of furans (e.g. furfural, hydroxymethylfurfural) that may be produced in the hydrolysis system being developed as part of the DIBANET project (see Section 18). However, the detector may also have use in the analysis of the acid soluble lignin products and the sugar degradation products from the analytical hydrolysis procedure. This detector is capable of measuring the absorbance spectrum from 190-800 nm with a deuterium lamp optimising the ultraviolet range (190 to 380 nm) and a tungsten lamp optimising the visible range (380 – 800 nm).

4.6.2 Determination of Optimum Conditions for Carbohydrate Analysis

4.6.2.1 Initial Dionex Conditions

In the tendering process for the IC system Dionex offered to analyse samples provided to them by the Author to demonstrate the capability of the system. The Author explained that reasonable resolution between fucose, arabinose, galactose, rhamnose, glucose, xylose, and mannose was of key importance and suggested the Davis procedure (Section 4.4) to achieve this. The Author sent 3 hydrolysates/solutions to the Dionex headquarters in Switzerland and he later visited this facility to see a demonstration of the system.

Dionex, however chose to use different conditions from those employed in the Davis paper. An NG1 was still used in order to remove hydrophobic components prior to passing the sample through the analytical column; indeed there were two NG1s placed in the switching-valve system with one being washed in a counter-current direction while the other was involved in the analytical set-up (as explained in Section 4.4). However, the samples were diluted between 20X and 100X and pre-treated (using two different cartridge-packs) in order to neutralise the acid and remove the sulphate from the sample. Furthermore, a PA20 column and different chromatographic conditions were employed:

Flow: 0.4 mL/min
 Temperature: 30°C (column)
 Injection Volume: 10 µl
 Gradient Conditions:

Time (min)	%A H ₂ O	%B 10mM NaOH	%C 500mM NaOH	%D 1mM NaOAc in 25mM NaOH
-17.00	70	30	0	0
23.25	70	30	0	0
24.50	69	11	20	0
26.00	69	11	20	0
45.00	48.4	6.9	20	24.7
45.60	50	0	50	0
51.00	50	0	50	0

NG1 Regeneration: Eluent A: 5mM NaOH B: water(20%)/acetonitrile(80%)
 Flow 0.4 mL/min

Time (m)	-17.0	-14.0	-11.0	17.0	20.0	51.0
%A	100	100	0	0	100	100
%B	0	0	100	100	0	0

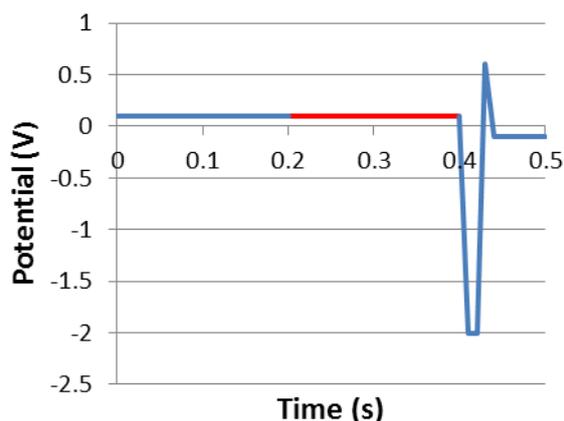
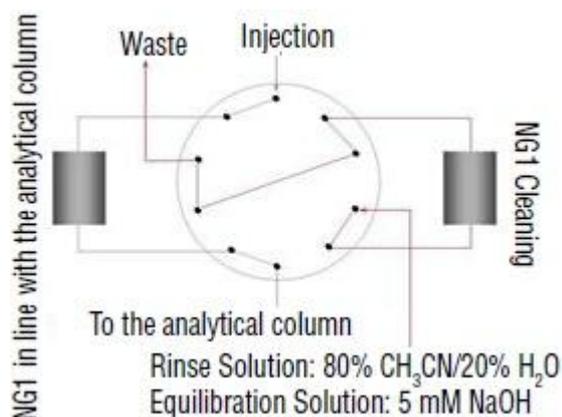


Figure 4-6: Dual NG1 columns in the IC system. Figure 4-7: The standard "carbohydrates" waveform. Red line = integration period
 Adapted from (Dionex, 2009)

Detection: Electrochemical – "Carbohydrates" waveform:

Time (s)	0.00	0.20	0.40	0.41	0.42	0.43	0.44	0.50
Potential (V)	0.1	0.1	0.1	-2.0	-2.0	0.6	-0.1	-0.1
Integration		Start	End					

These results of the analysis appeared to be good at the time, and the concentrations determined for one of the samples were in good agreement with the data for the lignocellulosic composition of the NIST biomass sample (NIST, 2011). The method also had the advantage that it could quantify for monosaccharides and UA in the same injection, although this did result in a relatively long analytical

time of 68 minutes per sample. The operator in Switzerland also tested the performance of the NG1 online cleaning system and noted that the chromatographic properties of the analytical column remained unchanged during the sequence when this system was employed.

Upon receipt of the ICS-3000 system at the Carbolea laboratories, the author was able to examine the chromatograms that resulted from the analysis of these samples in more detail. These are provided in Figure 4-8. Inspection of Figure 4-8 (c) shows that the resolution between some of the sugars (rhamnose/arabinose and xylose/mannose) is poor and that the separation between glucose and xylose is not ideal. Resolution improved in the case of sample 2, Figure 4-8 (e). However, this sample did not include rhamnose and the sugar concentrations as presented to the IC system were significantly lower. This lower concentration resulted in the noise of the baseline playing an increasing role in the determination of peak delimiters' locations; it can be seen that the baseline is not flat meaning that peak delimiter positions could be subject to variation, particularly between samples. Table 4-1 provides some peak data for the three injections of sample 1. It includes two statistics for the resolution of the peaks – the first involves the reference peak being the previous identified peak in the sequence and the second uses the next peak in the sequence as the reference. It can be seen that the separation between rhamnose and arabinose is so poor that some resolution statistics can not be determined.

Table 4-1: Average peak results (with standard deviations in brackets) for the three injections of sample 1. t = retention time; A = asymmetry; S/N = signal to noise ratio; Galact. Ac = galacturonic acid; Glucu Ac. = glucuronic acid; Peak Type = peak delimiter method. The standard deviation for the retention time was less than 0.01 minutes for all analytes.

Peak	t (mins)	Area (nC*min)	Peak Type	Height (nC)	A	Resolution (prev. peak)	Resolution (next peak)	S/N
Fucose	5.59	0.0636 (0.005)	BMB	0.314 (0.013)	1.053	3.01 (0.04)	2.38 (0.08)	11.9
Rhamnose	11.08	0.2084 (0.003)	BM	0.684 (0.003)	-	-	-	-
Arabinose	11.56	0.5721 (0.005)	MB	1.599 (0.013)	-	-	5.11 (0.03)	53.0
Galactose	14.64	1.0856 (0.005)	BMB	2.506 (0.005)	1.207	5.11 (0.03)	2.43 (0.01)	134.9
Glucose	16.78	59.1786 (0.196)	BMB	83.828(0.138)	1.607	2.43 (0.01)	3.02 (0.01)	223.4
Xylose	19.83	18.1502(0.072)	BMb	31.955(0.096)	1.153	3.02 (0.01)	1.26 (0.01)	85.2
Mannose	20.94	1.1226 (0.006)	bMB	2.219 (0.013)	1.430	1.26 (0.01)	3.60 (0.01)	5.9
Galact. Ac	41.98	0.9114 (0.013)	BMb	6.408 (0.078)	1.200	1.82 (0.05)	1.45 (0.07)	230.4
Glucu. Ac.	42.85	0.1116 (0.002)	bMB	1.400 (0.019)	1.883	1.26 (0.03)	4.78 (2.13)	50.3

Regarding sample 3, the peak for cellobiose, Figure 4-8 (f), is almost a rider on the unidentified peak that follows it, a far from ideal situation for an accurate quantification method that is comparable between samples.

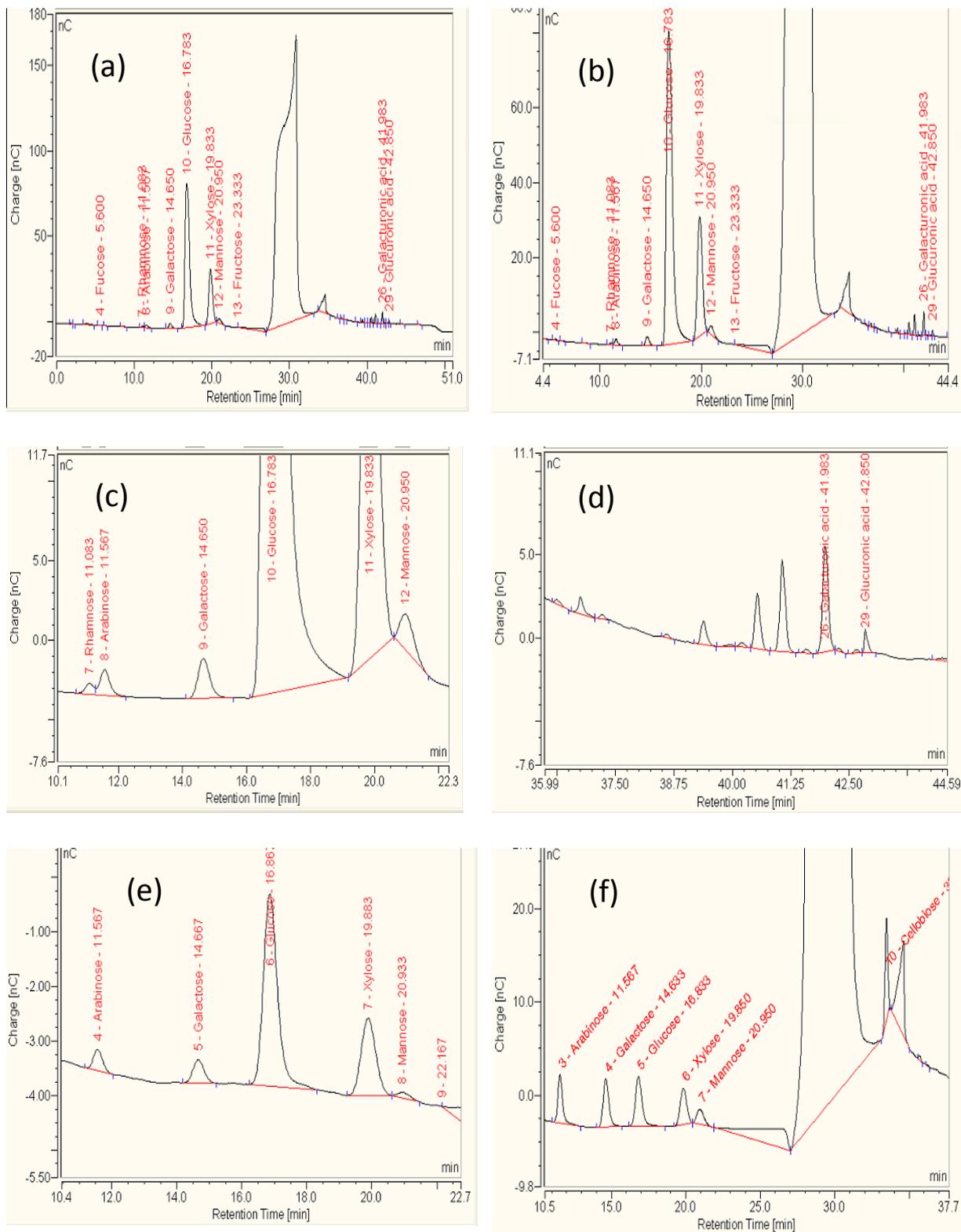


Figure 4-8: Chromatograms obtained upon the analysis of the three samples sent by the Author to the Dionex test laboratory in Switzerland. (a) The full chromatogram of sample 1; (b) The chromatogram from (a) is zoomed in so that the eluting analytes can be seen in more detail; (c) The chromatogram from (a) is zoomed in to see the monosaccharides; (d) The chromatogram from (a) is zoomed in to see the uronic acids; (e) The chromatogram for sample 2, zoomed in for the sugars; (f) The chromatogram for sample 3 zoomed in for the monosaccharides and cellobiose.

On a positive note, the resolution of the uronic acids (Figure 4-8 (d)) was reasonable, particularly considering their relatively small concentrations (a galacturonic acid concentration of 0.78% and a glucuronic acid concentration of 0.05% were calculated for sample 1) and the difficulties often encountered in resolving UA. Indeed, the presence of additional peaks in Figure 4-8 (d), peaks that were not present in the chromatograms of samples 2 and 3, indicate that the method was possibly able to discriminate between more UA species, although without the appropriate standards it is unknown what these are. However, as detailed in Section 3.2.4, the hydrolysis conditions that were employed to hydrolyse the polysaccharides of the Eastern Cottonwood sample are unlikely to be strong enough to result in the liberation of all UA as free monomers. Therefore the ability to resolve between UA and monosaccharides may be a moot point since additional hydrolysis steps on the sugar-hydrolysate are likely to be necessary for UA analysis and these conditions would most likely degrade much of the monosaccharides liberated from the standard hydrolysis method.

Upon the installation of the ICS-3000 system at the Carbolea laboratories, tests were undertaken with the PA20 column using variants of the method described above and other methods. These results will not be presented here, but, in summary, none of them improved significantly on the chromatograms presented in Figure 4-8 and they were considered unacceptable by the Author for incorporations as standard protocols for developing analytical results suitable for NIR calibrations.

Therefore, on the basis of the need for highly reproducible results (i.e. peak delimiters, area integrations) for samples of differing concentrations, and the preference for shorter chromatographic run times, and the avoidance of having to use sample-clean-up cartridges (which would cost approximately €5 for each injection) the Author decided that the Davis protocol (Section 4.4) would be attempted in the Carbolea labs. This required the supply of the PA1 analytical column from Dionex and this was installed in late May 2008.

4.6.2.2 Attempts to Replicate the Davis Paper

The Davis conditions involve sugar elution with water for 11 minutes, followed by a one minute ramp to 170 mM $\text{NaC}_2\text{H}_3\text{O}_2$ in 200 mM NaOH, which was maintained for 5 minutes, then a 1 minute return to the water elution which continued for 9.5 minutes. Flow rate is 1.2 ml/min, temperature is 22°C. There is also the addition of post-column base (300 mM NaOH) at a flow rate of 0.3 mL/min. Figure 4-9 shows the chromatogram that resulted from following these conditions (with the exception that the isocratic elution of water was doubled to 22 minutes in order to check for late-eluting components) for a 4% acid sugar-solution containing fucose, arabinose, galactose, glucose, xylose and mannose (all at 0.5 mg/mL concentration). The dotted purple line in Figure 4-9 represents the gradient for eluent B (i.e. the eluent used in the column conditioning step).

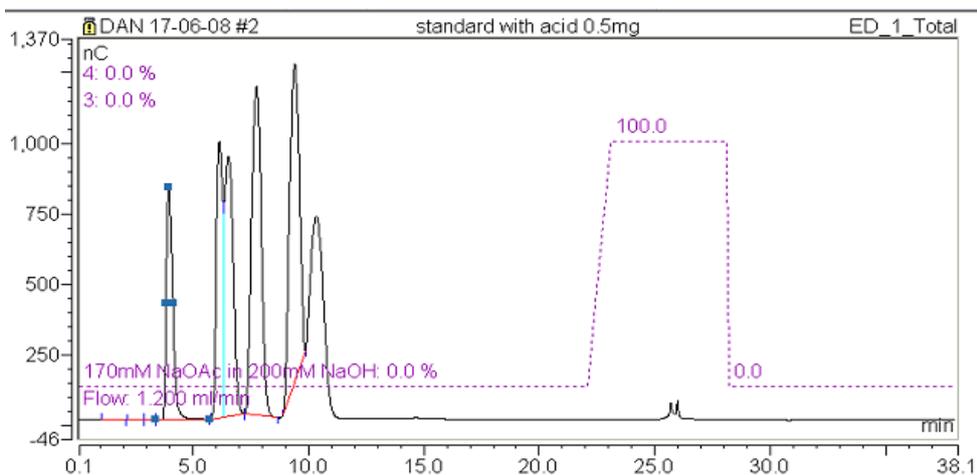


Figure 4-9: Chromatogram obtained from a sugar standard solution (containing fucose, arabinose, galactose, glucose, xylose and mannose in water; the peaks are eluted in this order) using the same conditions as the Davis (Davis, 1998) paper (except with a longer period for sugar elution).

It can be seen that the peak resolution is extremely poor, with the second and third peaks (arabinose and galactose) almost forming one peak due to their extremely similar retention times, and with xylose and mannose also poorly separated. The situation was similar when a non-diluted (as was the case in the Davis protocol) hydrolysate of the NIST standard biomass material Eastern Cottonwood was injected onto the system, Figure 4-10.

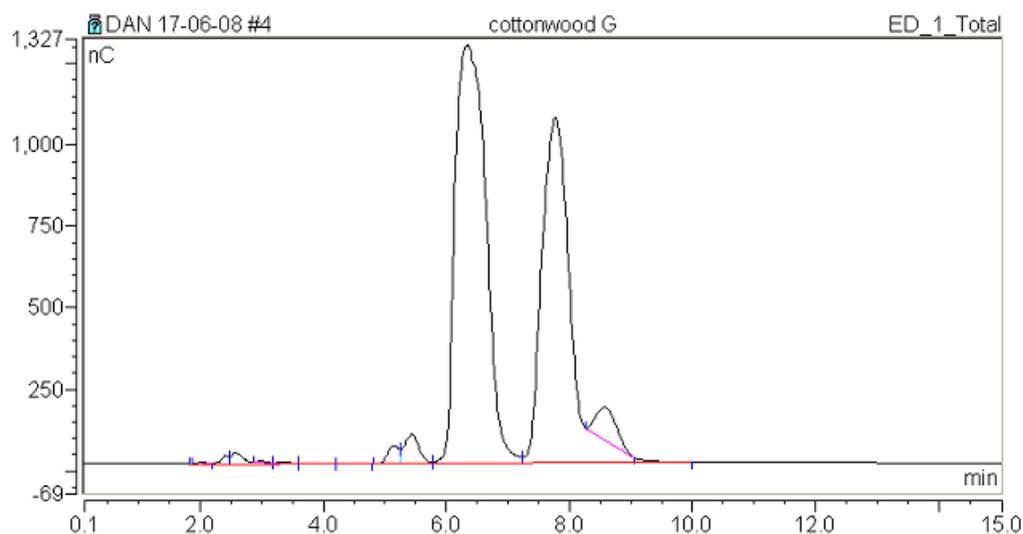


Figure 4-10: Chromatogram obtained when a non-diluted hydrolysate of Eastern Cottonwood (NIST biomass standard) was injected onto the system using the Davis conditions.

Following these results experiments were undertaken whereby each of the sugars were injected separately onto the system under methods where the eluent flow rate, during the period for the elution of sugars, was normal (1.2 mL/min), halved (0.6 mL/min), and reduced to a third (0.4

mL/min). It was considered that the slowest speed could potentially resolve all sugars and, upon injection of the six-sugar solution (1 mg/mL) all peaks were separated, with the exception of arabinose and galactose although their resolution did improve at concentrations of 0.2 mg/mL or less.

The Author also experimented with varying the acetate concentration in the column conditioning step. When using an acetate loading of 85 mM NaOAc (half the concentration of the Davis paper) it was seen that all the 6 sugars (each at 1 mg/mL concentration) were well resolved under a flow rate of 0.4 mL/min. However, the reduction in acetate loading and flow rate resulted in an increase in the retention times of all analytes meaning that 52 minutes was required for each injection. Putting the flow rate back to 1.2 mL/min but retaining the half-acetate loading was attempted and proved successful in providing reasonable resolution between the six sugars (1 mg/mL) with all sugars eluted in approximately 11 minutes. Unfortunately, once a solution containing these six sugars plus rhamnose was injected and analysed under these conditions, peak resolution problems returned with the rhamnose peak occurring only fractionally earlier than the galactose peak.

It had become clear that, as mentioned in the Davis paper, the acetate loading was critical in order to achieve satisfactory resolution of the sugars. However, it was also clear that the conditions outlined in the Davis paper did not result in comparable chromatograms in the Carbolea system. The reasons for this are unclear but, since the elution of sugars occurs with only deionised water as the mobile phase it is possible that differences in the water composition between labs, however slight, may affect ultimate chromatographic performance with such a weak eluent/pusher.

The Author decided that it would be necessary to investigate the effects of various acetate loading levels on each sugar individually. Many different experiments were conducted and results showing the effects of five different acetate loadings on the retention times of several sugars are presented in Table 4-2. Figure 4-11 presents the retention times (as a fraction of the retention time at 59.5 mM NaOAc) over these acetate loadings. It can be seen that, in contrast to what was stated in the Davis paper, rhamnose is not the least sensitive analyte to an adjustment in acetate loading (fucose is). Furthermore, the relative response of arabinose is closer to the relative response of rhamnose than it is to the relative responses of all other sugars (except fucose) whose responses are all extremely similar. What is key in the resolution of rhamnose, however, is its position relative to galactose and glucose, the two sugars nearest to it in the chromatogram and the results in Table 4-2 and Figure 4-11 indicated that there may be sufficient differences between the responses of rhamnose and these two sugars to allow fine-tuning of the acetate loading level for resolution between all sugars.

Table 4-2: The retention times (RT) of seven sugars under different sodium acetate loadings

Sugar	Retention Times of Sugars (min) Under different Acetate Loadings(mM)					% Difference in RT between 59.5 and 127.5 mM
	59.5 mM	85 mM	105 mM	116 mM	127.5 mM	
Fucose	3.617	3.017	2.967	2.883	2.75	23.97
Arabinose	7.217	5.5	5.35	5.067	4.767	33.95
Galactose	8.467	6.15	5.967	5.65	5.217	38.38
Rhamnose	8.384	6.5	6.333	6.067	5.65	32.61
Glucose	10.335	7.483	7.25	6.85	6.284	39.20
Xylose	12.767	9.25	8.917	8.417	7.7	39.69
Mannose	14.27	10.412	9.95	9.417	8.584	39.85

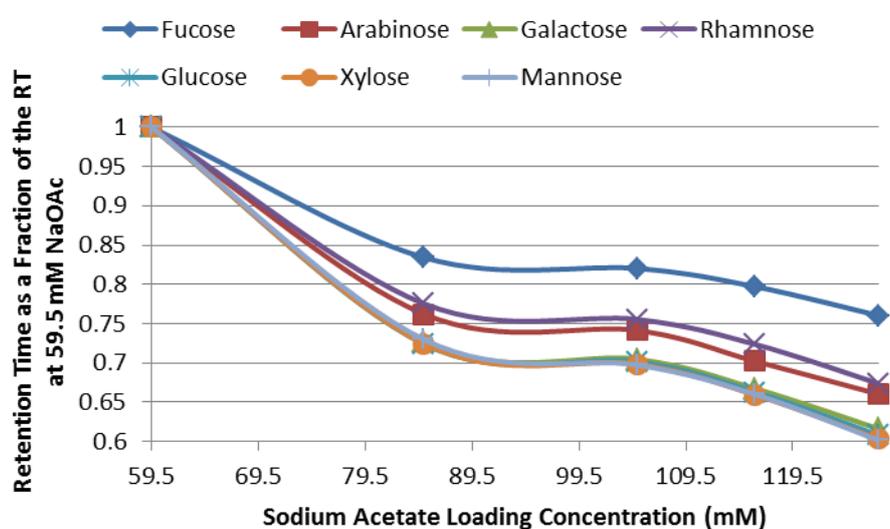


Figure 4-11: The retention times (as a fraction of the retention time associated with a 59.5 mM sodium acetate loading) for the seven sugars under various acetate loading concentrations.

The Author then undertook a series of experiments, with sugar standard solutions containing all seven sugars, under which various acetate loading levels were tested. It was found that an acetate loading of 127 mM, while not able to resolve between galactose and rhamnose when all the sugars were at concentrations of 1 mg/mL, was able to resolve all sugars for a solution containing them in the concentrations that would be expected from hydrolysates of wheat straw and Eastern Cottonwood samples (i.e. significantly lower concentrations for all sugars except glucose).

However, when these sugar standard solutions were acidified so that they had a 4% sulphuric acid content, it was noted that the retention times of the sugars changed with a resulting effect on resolution. Once again fine-tuning of the acetate loading was required and experiments were also conducted once again on flow rate and on changing the column and detector temperature. It was found that reducing the column temperature had a similar effect to reducing the flow rate and was helpful in improving resolution. A set of conditions were found in these experiments that resulted in

the reasonable separation of all 7 sugars in undiluted 4% acid solutions: 130 mM NaOAc loading (with 200 mM NaOH), 17°C, and a 0.6 mL/min flow rate for the sugar elution step.

In November 2008 the Author made contact with researchers at the US Forest Products Laboratory (FPL) who had worked on the ion chromatography system described by the Davis paper. These researchers had modified the chromatographic conditions in the column-conditioning step and were now using 240 mM NaOAc in 400 mM NaOH with modifications to the length of this step and the gradients between this and the isocratic water phases. They were also maintaining a column temperature of 18°C. The Author tried to use these conditions on samples that had been diluted 5X (see Section 4.6.2.3 for details on the dilutions) and found the resulting chromatograms to be excellent with good resolution between sugars. A standard method was then employed for the analysis of hydrolysates. Provided below is a summary of the chromatographic conditions that were used in providing all of the IC results presented in subsequent sections of the Thesis.

Column Pump:

Flow: 1.1 mL/min
 Gradient Conditions:

Eluent	Amount of Each Eluent (%) At These Times (min)					
	0.0	16.0	16.5	18.5	19.0	34.1
A – H ₂ O	100.0	100.0	36.0	36.0	100.0	100.0
B – 1M NaOAc	0.0	0.0	24.0	24.0	0.0	0.0
C – 1M NaOH	0.0	0.0	40.0	40.0	0.0	0.0

Post-Column Pump:

Eluents: A = H₂O, B = 1M NaOH
 Flow: 0.3 mL/min – 70% A, 30% B (i.e. 300mM NaOH)

Other Conditions:

Column: Dionex CarboPac PA1 column (4 x 250 mm) and PA1 guard (4 x 50 mm).
 Temperature: 18°C for the column and detector compartments.
 Injection Volume: 10.1 µl
 Detection: Electrochemical – “Carbohydrates” waveform (Figure 4-7)
 Hydrophobics removal: NG1 guard column included in line before the PA1 guard. A switching valve diverts eluent flow around this NG1 two minutes after sample injection. NG1 not used for SRS solutions.
 NG1 cleaning: After every approximately 100 injections via a back-flush (to-waste) with 80% acetonitrile followed by a wash with water.

Shutdown Procedure:

After the last sample is analysed the PA1 column is washed with 200mM NaOH for 30 minutes and then the Column Pump is turned off. The Post-Column Pump continues to pump 100% water at a flow rate of 0.3 mL/min for 10 minutes in order to clean any base from the working electrode.

The conditions outlined above have been tested on a wide variety of biomass feedstocks that have significantly differing concentrations and proportions of the 6 component sugars and resolution is of a high quality in all cases. Figure 4-12 presents some examples of these chromatograms and how the chromatographic method can resolve between the component sugars. Table 4-3 presents some chromatography statistics for several replicate analyses of a sugars standard solution (in acid) that was made up to have concentrations of sugars that are similar to those expected in the diluted hydrolysates of a Miscanthus sample. Statistics are presented for two different system configurations. In the first instance the external sugar standard (ESTD) was used to determine relative response factors (see Section 4.6.2.4) appropriate for the SRSs. These solutions only contain acid, the hydrolysed sugars, and any sugar degradation products (e.g. furfural etc.), and so do not have any of the hydrophobic components that may be present in the hydrolysates of solid biomass samples (e.g. fractions of the ASL) that would need to be removed by the NG1. Hence, these SRS solutions and their ESTDs bypass the NG1 column and travel directly to the PA1 guard and column. The second set of statistics correspond to a situation where the ESTD is being used to determine the relative response factors for the biomass hydrolysates and, since these may have hydrophobic components, passing through the NG1 column is necessary in order to avoid contamination of the PA1 columns. The Davis paper said that the chromatographic performance was the same whether or not an NG1 column was included. However the results in Table 4-3 indicate that this is not entirely the case for the Carbolea system. There is a slight reduction in the resolution and signal to noise statistics as a result of passing through the NG1 column and, obviously, the retention times are increased by a small degree. Nevertheless, the statistics in both instances are good and a clear improvement of the data in Table 4-1.

The Author has observed that the loss in resolution associated with passing through the NG1, although still present when a clean NG1 is used, can increase somewhat over the course of a chromatography batch, but this only occurs when particularly troublesome samples (primarily the hydrolysates of peats; see Section 13) are injected multiple times. The subsequent cleaning of the NG1 after a batch has finished eliminates the loss of resolution experienced from contamination by these samples. In normal analytical sequences involving less “dirty” samples (such as Miscanthus and bagasse hydrolysates) over 100 injections can be put through the system without any noticeable visual effect on the shape and resolution of peaks.

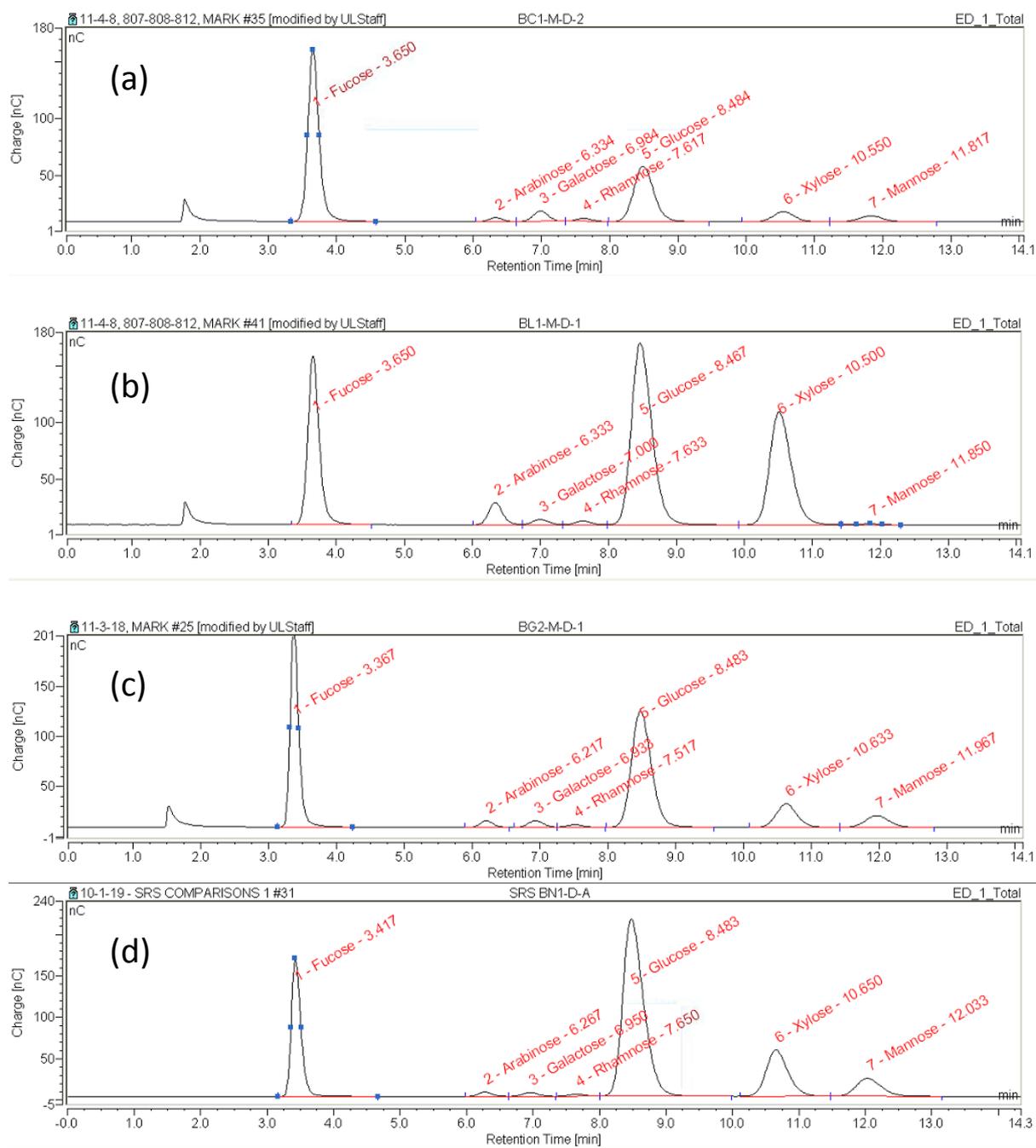


Figure 4-12: Four chromatograms showing the resolution between the major sugars for samples where the proportions of these sugars vary. (a) A sugar standard made with sugar proportions similar to many peat samples; (b) An external standard (ESTD) similar to many *Miscanthus* and other herbaceous samples; (c) An ESTD similar to some waste feedstocks; (d) An ESTD similar to many paper/cardboard samples.

Table 4-3: Chromatography statistics (after the system has equilibrated) for a sugar standard solution (Figure 4-12 (b)). The first set of numbers are for a set-up where the sample does not pass through NG1 prior to the PA1 guard/column (6 intermittent injections during a sequence, over a period of approx. 9 hours). In the second set of numbers the sample does pass through the NG1 guard (6 intermittent injections during a sequence, over a period of 12 hours). RSD = Relative Standard Deviation, RRF = relative response factor (to fucose the internal standard). Response factor units are nC.min/mg/ml (i.e. the area for a concentration of 1 mg/ml). The values for mannose are worse than the other sugars because mannose is at a very low concentration in this sample.

	Retention Time (mins) [RSD %]	RSD Area (%)	RSD Height (%)	Theoretical Plates [RSD %]	Asymmetry [RSD %]	Resolution (Prev Peak) [RSD %]	Resolution (Next Peak) [RSD %]	Signal/Noise [RSD%]	Response Factor [RSD %]	RRF [RSD %]
DOES NOT PASS THROUGH NG1										
Fucose	3.350 [0.005]	0.760	0.385	3598 [0.849]	1.31 [3.98]	8.88 [1.56]	9.95 [0.30]	3963 [17.06]	213.5 [0.76]	N/A
Arabinose	6.117 [0.22]	0.44	0.38	5357 [0.37]	1.13 [2.33]	9.95 [0.3]	1.95	442 [17.28]	205.5 [0.44]	0.9625 [0.55]
Galactose	6.817 [0.2]	0.66	0.45	5009 [0.74]	1.04 [3.20]	1.95 [0.0]	1.46 [1.52]	97.65 [17.9]	180.1 [0.66]	0.8432 [0.83]
Rhamnose	7.396 [0.21]	1.12	0.97	5186 [0.59]	1.09 [1.30]	1.46 [1.52]	1.93 [0.30]	66.75 [16.5]	135.2 [1.12]	0.6331 [0.52]
Glucose	8.296 [0.19]	0.65	0.42	3957 [0.21]	1.22 [2.16]	1.93 [0.3]	3.8 [0.21]	3321 [17.0]	193.4 [0.65]	0.9057 [0.13]
Xylose	10.354 [0.24]	0.86	0.62	5511 [0.28]	1.22 [2.41]	3.8 [0.21]	2.36 [0.60]	2134 [11.67]	218.6 [0.86]	1.0238 [0.11]
Mannose	11.713 [0.24]	2.19	1.34	6160 [0.96]	1.09 [3.04]	2.36 [0.6]	N/A	19.6 [11.55]	145 [2.19]	0.6792 [1.83]
PASSES THROUGH NG1										
Fucose	3.653 [0.19]	0.39	3.19	2513 [7.34]	1.21 [3.81]	8.09 [4.71]	8.13 [2.19]	2308 [20.75]	214.8 [0.39]	N/A
Arabinose	3.653 [0.14]	0.36	1.56	4773 [2.93]	1.12 [1.67]	8.13 [2.19]	1.73 [1.51]	301 [19.83]	201.2 [0.36]	0.9369 [0.16]
Galactose	6.986 [0.18]	0.85	1.51	4782 [1.89]	1.08 [2.26]	1.73 [1.51]	1.49 [1.68]	65.5 [19.67]	170.7 [0.85]	0.7946 [0.71]
Rhamnose	7.614 [0.22]	1.27	1.83	4809 [1.56]	1.06 [3.43]	1.49 [1.68]	1.71 [0.96]	45.5 [19.69]	132.5 [1.27]	0.6168 [1.03]
Glucose	8.45 [0.18]	0.48	0.96	3830 [1.40]	1.23 [2.15]	1.71 [0.96]	3.64 [0.60]	2371 [19.09]	194.9 [0.48]	0.9072 [0.16]
Xylose	10.486 [0.12]	0.41	0.91	5281 [1.31]	1.23 [1.99]	3.64 [0.60]	2.26 [1.21]	1471 [19.08]	221.3 [0.41]	1.0301 [0.10]
Mannose	11.82 [0.19]	3.21	2.17	6063 [4.40]	1.09 [6.36]	2.26 [1.21]	N/A	13.1 [20.77]	140.1 [3.21]	0.6523 [3.17]

4.6.2.3 Linearity Tests

The Author was looking to develop a robust quantitative chromatographic method that would be applicable to the hydrolysates obtained from a very diverse range of biomass samples. These included samples where the expected total sugar concentrations were relatively low (e.g. less than 10% total lignocellulosic carbohydrate was expected for some peats) as well as samples with close to 100% carbohydrates (as can be possible with some papers). Assuming that the dry mass of a sample was 285 mg (equivalent to 300 mg wet weight at 5% moisture content (wet basis)), then the resulting glucose concentration in the (undiluted) hydrolysate would range from 0.034 mg/mL for a sample with 1% glucan content to 3.24 mg/mL for a sample with a 95% glucan content (these calculations consider the hydration involved in the conversion from the cellulose polymer to the monomeric form and the sugar losses experienced in the secondary hydrolysis step).

Hence, most of the experiments that took place up till the point of contact with the researchers at the US FPL involved relatively strongly concentrated sugar solutions (either with or without acid present), or solutions that had been made up to approximate an (undiluted) biomass hydrolysate, or biomass hydrolysates themselves that had not been diluted. The reasons for not diluting the samples were that the Author ideally preferred to have a system that required minimal sample preparation. Furthermore, the Davis paper indicated that sugar resolution and quantification was possible on a Dionex system under these conditions.

However, simple tests carried out by the Author, involving several dilutions of glucose and mannose sugar solutions, indicated that linearity over the desired concentrations was not possible with the IC system as supplied by Dionex. A solution with a glucose concentration equivalent to that which would be expected from the undiluted hydrolysate of a biomass sample with 67% glucan was made and diluted 6 times with the lowest concentration being a 10X dilution. With mannose a solution corresponding to the hydrolysate of a sample with 36% mannan was made and 9 different dilutions were analysed on the IC system, with the lowest concentration solution being a 100X dilution. Figure 4-13 presents the curves for the areas of these solutions along with predicted curves based on the single-point calibration using the area of the most dilute sample. It can be seen that there are severe deviations from linearity, particularly for mannose.

It had become clear, therefore, that there was a significant difference between the electrochemical detector in the Dionex ICS-3000 system and the Dionex detector that was used by Davis in the DX-500 IC system (Davis, 1998). The ICS-3000-ED is a much more sensitive detector (personal communications

with Dionex representatives indicated that the detector in the ICS-3000 system is approximately 100 times more sensitive) meaning that it has a reduced linear response range for the sugars. The Author discussed the situation with Dionex and a wider (0.015 inch) flow gasket was purchased for the ED working electrode. This resulted in less eluent/sample passing over the detector and therefore a lower response for the same sugar concentrations, effectively increasing the linear range of the detector. Naturally, there will be an associated loss in sensitivity of the detector. However, that is unlikely to be an issue for most of the important sugars in the hydrolysates of lignocellulosic biomass samples.

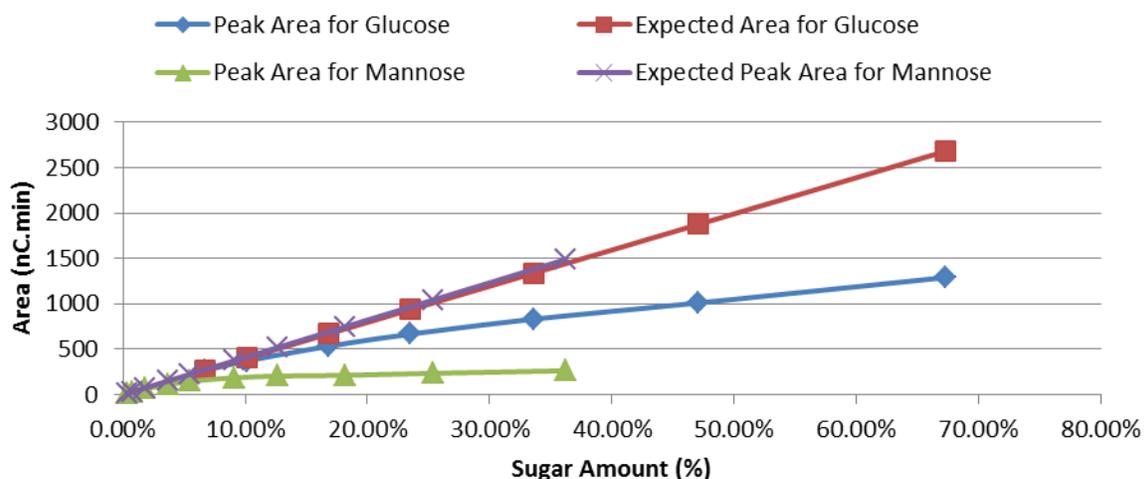


Figure 4-13: A plot of peak area for sugar solutions corresponding to (undiluted) hydrolysates containing various concentrations of mannose or glucose. The predicted areas are also plotted. The deviation from linearity is clear.

The researchers at the US FPL had informed the Author that the concentrations provided in Table 4-4 were the highest limit on their calibration curves (using 10 µl injections), with the figures in the subsequent row indicating the corresponding concentration (%) of the sugar in the biomass sample that would give rise to this concentration in the undiluted hydrolysate, assuming the hydrolysis method outlined in Section 11.5 was employed. These concentrations in Table 4-4 are too low to allow undiluted hydrolysates to be injected directly onto the system. However, a 5X dilution would allow for all biomass hydrolysates relevant to this study to be within the linear range.

The Author undertook a series of calibration experiments to determine if the maximum linearity concentrations provided by the FPL laboratory would be applicable to the IC system in the Carbolea labs and to examine whether the special gasket installed at the working electrode could potentially expand the linear range further. The results of these calibration curves are presented in Table 4-5. Note that the

calibrations for fucose and rhamnose were carried out using 5 µl injections and showed a linear response beyond double the maximum concentrations listed in Table 4-5. Hence it is considered that the responses will be linear for these sugars at the concentrations below when using a 10 µl injection. The value for the slope of the curve for the 5 µl injections is doubled in Table 4-5 so that it is on the same basis as the other sugars. However, it should be noted that these injections occurred on a different day to the 10 µl injections so the comparison between response factors should not be taken too strictly. The slope of the curve represents the area expected for a 1 mg/mL concentration of the sugar.

Table 4-4: The highest concentrations of seven biomass sugars that were used in the US FPL laboratory's calibration curve. Data provided by email by Frederick J. Matt at FPL. The % Biomass row converts these hydrolysate concentrations to the corresponding proportion that this sugar would form of the biomass (% dry matter) prior to hydrolysis.

	Arabinose	Galactose	Rhamnose	Glucose	Xylose	Mannose	Fucose
Conc (mg/mL)	0.073	0.068	0.04	0.82	0.37	0.42	0.3423
% Biomass	2.12	1.97	1.16	23.78	10.73	12.18	-

Table 4-5: The lowest and highest concentrations used in the calibration curves for the 6 biomass sugars. The corresponding percentage of biomass that these concentrations represent in an undiluted hydrolysate are also included, as is the percentage of biomass that the maximum concentration represents after the hydrolysate is diluted 5X. The multiple correlation coefficients and slope statistics of the (5-point) calibration curves are also provided.

Sugar	Range of Concentrations in Calibration Curve (mg/mL)		Max % Biomass with 5X dilution	R ²	Slope of Curve
	Lowest Conc. [% biomass]	Highest Conc. [% biomass]			
Arabinose	0.0104 [0.30]	0.1404 [4.06]	20.3	0.9992	254.52
Galactose	0.0054 [0.15]	0.1044 [3.05]	15.25	0.999	272.29
Rhamnose	0.0052 [0.15]	0.0315 [0.91]	4.55	0.9961	118.74
Glucose	0.2319 [6.73]	1.1603 [33.64]	168.2	0.9989	262.09
Xylose	0.0127 [0.37]	0.6295 [18.27]	91.35	0.9994	281.36
Mannose	0.0063 [0.18]	0.6272 [18.13]	90.65	0.9995	237.92

Linearity tests were also conducted for other sugars using 5 µl injections. It was noticed that the multiple correlation coefficients in these cases were lower than for the 10 µl injections. This was not due to any differences in linearity but instead a result of inconsistencies in the injection volume provided to the system by the autosampler.

The method of loading a sample onto the eluent stream involves the sample loop on the injection valve being filled with a user-defined volume of sample via a line from the autosampler. The sample loop installed at the time of the linearity tests was a 10.1 μl capacity loop (calculated based on the length of tubing cut to make the loop). Partial loading of the sample loop requires that the system is extremely well configured and the exact volume between the sample loop and autosampler diverter valve is known. Any discrepancies in that could lead to variations in the injection volumes. This would also be applicable to full-loop injections, although to a much lower degree. Even though the Author spent some time configuring the system, the reproducibility of partial loop injections was considered to be unacceptable. For all of the analytical results provided for biomass hydrolysates in subsequent chapters of this Thesis a 12 μl injection of each sample was specified by the user. This should ensure that the 10.1 μl capacity of the sample loop only contained the sample. As a further precaution an internal standard was included with each sample, helping to lessen any influence that any remaining variability in sample injection volume would have.

The next phase of the linearity/quantification tests involved the Author making up a sugar standard solution containing the seven sugars in similar concentrations to those listed in Table 4-4. This (full strength) solution was then diluted 2X, 3X and 5X. From the previous experiments on the individual sugars it was already known that the responses would be linear; however, the Author wished to see how well each solution could, by being used as an external standard to determine the relative response factor (RRF - see below) for each sugar relative to fucose, predict the sugar concentrations of itself and other solutions in subsequent injections. It was found that, when the RRFs from each solution were used to predict its own concentrations in subsequent injections of that solution, the accuracy was greater than 99.5% for the major sugars. However, when these solutions were used to predict the concentrations of the solutions with different dilutions factors the accuracy fell somewhat. This indicated to the Author that, while the response is linear over a relatively wide range, highly accurate analysis would require that external standards of similar concentrations to the analyte would be required.

4.6.2.4 Sugar Standards and Analytical Sequences

The principles of external standards (ESTD), relative response factors, and internal standards and their integration in a chromatographic sequence will now be discussed.

A sugar standard solution contains known concentrations of all of the relevant sugars, and when this sample is injected on to the IC system response factors can be determined as follows:

$$Response\ Factor_{sugar} (nC.\ min/mg\ mL^{-1}) = \frac{Area_{sugar} (nC.\ min)}{Concentration_{sugar} (mg/mL)} \quad (4.10)$$

Hence, the response factor (RF) indicates the area that would be expected for a peak corresponding to a 1 mg/mL concentration of the sugar (linearity assumed). A multiple point calibration involves the analysis of several sugar standards of varying sugar concentrations, resulting in a scatter plot, and linear or non-linear regression methods are then used to fit a calibration curve to these points. The extended linear range and reproducibility of the PAD detector allows single point calibration methods to be used, and these are widely employed in the chromatographic analysis of carbohydrates. This reduces the system time that must be spent on standard injections and also makes the intermitting injections of standards throughout a batch more feasible. Such intermittent injections are important because the RFs may drift over an extended period.

Table 4-3 contains, for the 7 sugars, some average RFs and their RSDs for multiple injections of a biomass standard. These RFs are based on the peak area, but there are other methods for calculating RFs; for example the peak height can be used. However it was found by the Author that RSDs increased somewhat when peak heights were used.

It can be seen in Table 4-3 that the RFs vary between the different sugars. The final column in that Table presents the relative response factor (RRF) of each sugar relative to the internal standard (ISTD) which is fucose in this instance. The RRF is simply calculated as:

$$RRF_{sugar} = \frac{Response\ Factor_{sugar}}{Response\ Factor_{ISTD}} \quad (4.11)$$

If samples with unknown sugar concentrations are diluted with known quantities of an ISTD (fucose) solution it will allow the response factor for the ISTD to be determined. This RF can then be multiplied by the RRFs determined for the recent ESTDs in order to calculate the estimated RFs for the sugars of the unknown sample. The advantage this has over the use of the ESTDs' RFs is that any variation in injection volume, or detector sensitivity, or other chromatographic conditions specific to the sample being analysed will be reflected in the RF determined for the fucose, and this response can then be transferred to RFs of other sugars by means of the RRFs. This usually means that the repeatability of analyses improves, as shown by the reduced RSDs for the RRFs in Table 4-3.

The Author was aware that fucose is a potential biomass component in some feedstocks, for instance in peats. However when these materials were analysed without using fucose as an ISTD it was found that the fucose concentration of the samples were so low as to be irrelevant and that the advantage offered by the use of fucose as an ISTD was of far greater importance than the very minor contributions of native fucose to these samples. Myo-inositol was examined as an alternative ISTD. However, it was considered that its RF was too different from that of the other sugars; hence fucose was used as an ISTD in all samples.

Sugar standard solutions were made up from dilutions of more heavily concentrated solutions, in order to ensure the concentrations were accurate, and then acidified so that the acid content was the same as that of the biomass hydrolysates (i.e. 4% sulphuric). Sugar standards and biomass hydrolysates were then diluted the same way; i.e. a 5X dilution with a 0.172 mg/mL fucose solution. Various sugar standards and ISTD solutions have been made up over the course of this study and their concentrations were all accurately recorded in order to ensure reproducibility between dilutions/ESTDs.

As shown in Figure 4-12, different types of sugar standard solutions have been made up in order that their concentrations and proportions of sugars should be comparable to the biomass hydrolysates for which they will be providing the RRFs.

It should be noted that it takes some time for the chromatographic system (detector response, pump pressure etc.) and for the RFs and RRFs to stabilise at the start of a batch, and this is particularly the case if a fresh set of eluents have been made up. Typically around 9 -10 injections of samples (usually SRS solutions or standards that do not need to go through the NG1 column) are put through the system before data are collected. In approximately half of the batches processed on the IC system the analytical sequence outlined in Figure 4-14 (a) was employed. Later the sequence was modified slightly to that of Figure 4-14 (b) so that fewer injections of sugars would be required and the RRFs calculated could be more representative of what the true RRFs may be for the hydrolysates encapsulated between the ESTDs. In either of these sequences the ESTDs that are used to calculate the RRFs for the unknown samples must be of a similar composition to those samples. If a chromatography batch contains vastly different samples (e.g. peats and Miscanthus samples) then these must be kept within their own "block" with the associated ESTDs. This is outlined in Figure 4-14 with ESTD 1 being used for samples of type 1 (e.g. peat). For the injection of biomass hydrolysates each "block" usually incorporates the two replicate hydrolysates of two biomass samples (i.e. 4 injections). The performance of the IC system is so stable that duplicate injections of the same hydrolysate are not considered necessary.

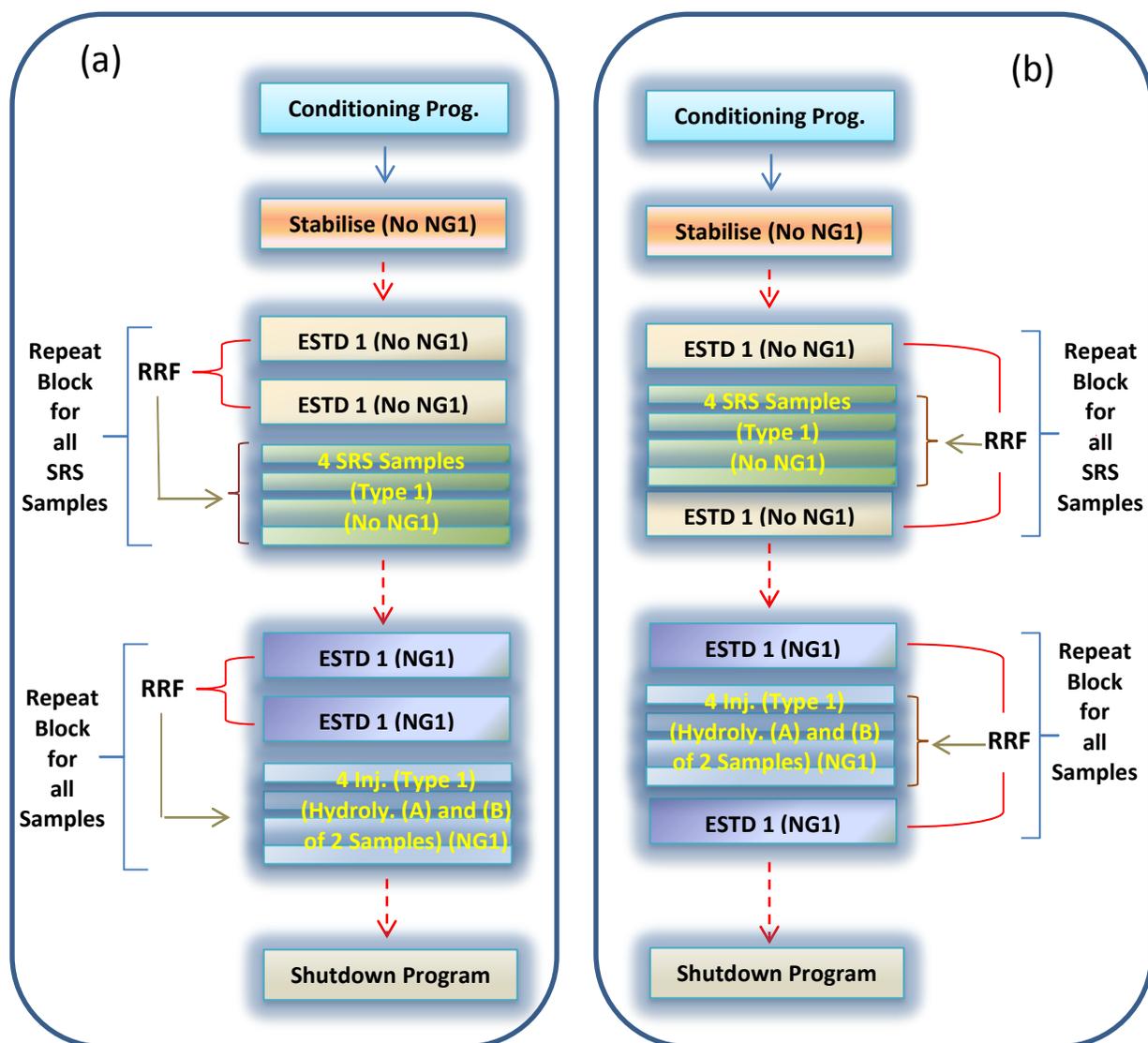


Figure 4-14: Two general chromatography sequences that were used in the research. (a) The average RRFs from two sugars standards are used for the next four unknown injections. (b) The average RRFs of the two ESTDS encapsulating the unknown samples are used for those unknown samples.

Sequences such as those outlined in Figure 4-14 have been successfully run over weekends with more than 100 injections taking place without any input required from the user. The regular injection of ESTDs helps to correct for any drift seen over this time and the use of more than one ESTD injection for the calculations of RRFs allows the user to check for any instability in the device in shorter time periods. Once the analytical results are available the peak areas observed for each of the sugars in the unknown samples are converted to their equivalent proportions of the dry, extracted, biomass sample by the following calculations:

1. Calculate the response factor of the sugar based on the RRFs of the ESTDs:

$$RF_{sugar} = RRF_{sugar} \times RF_{ISTD} \quad (4.12)$$

2. Then determine the concentration of the sugar in the hydrolysate, prior to dilution:

$$Conc_{sugar_{hyd}}(mg/mL) = Dilution_Factor \left(\frac{Area_{sugar}}{RF_{sugar}} \right) \quad (4.13)$$

3. This concentration is then corrected by the sugar recovery factor calculated from the SRS solutions and converted to the anhydro-polysaccharide form by the correction factor CF , which is 0.9 for hexoses and 0.88 for pentoses:

$$Conc_{sugar_{poly_corrected}}(mg/mL) = Conc_{sugar_{hyd}} \left(\frac{1}{SR_{sugar}} \right) \times CF \quad (4.14)$$

4. Finally the proportion that this sugar provides towards the total dry matter of the extractives-free biomass is calculated, based on Vol , the volume of the hydrolysate, and W , the dry weight of the sample (in mg).

$$Sugar\ (\%) = \frac{Conc_{sugar_{poly_corrected}} \times Vol}{W_{sample}} \times 100 \quad (4.15)$$

Section 4.5.1 details the different methods that can be used to classify peaks. The standard option in Chromeleon is for the software to automatically decide how to classify each peak on the basis of each chromatogram that is presented to it. This Automatic method was compared against a method whereby peak classification methods were assigned to each peak (each peak was set as a baseline-main-baseline, BMB). Another user-defined peak classification was also examined, named "TEST" in Table 4-6. This took the fucose peak as BMB and then drew a baseline from the start of the arabinose peak to the end of the mannose peak. Therefore, the arabinose peak was classified as BM, the mannose peak as MB and all other peaks as M (their areas were integrated by taking a line down to the baseline from the peak minima). Zoomed-in examples of the peak delimiters used in the different methods are shown in Figure 4-15, and the results of the analyses are presented in Table 4-6 for 6 injections of the same ESTD, through NG1, over the period of 12 hours. The results for the BMB method are the same as those provided in Table 4-3.

Table 4-6: Comparison of the RSDs of the RRFs for different peak classification methods, using 6 injections of the same ESTD. Also included are the peak classifications made using the automatic method, the nomenclature used is explained in Section 4.5.1. RSD = relative standard deviation.

Sugars	Peak Classification Made Using the Auto Method						RSD for RRF (%)		
	Inj. 1	Inj. 2	Inj. 3	Inj. 4	Inj. 5	Inj. 6	Auto. Peaks	“TEST” Peaks	All peaks as BMB
Fucose	BMB	BMB	BMB	BMB	BMB	BMB	-	-	-
Arabinose	BM	BM	BM	BM	BM	BM	0.42	0.28	0.16
Galactose	M	Mb	M	M	M	Mb	3.94	0.54	0.71
Rhamnose	M	bMb	M	M	M	bMb	6.83	1.19	1.03
Glucose	MB	bMB	M	MB	M	bMB	0.44	0.36	0.16
Xylose	BM	BM	M	BMb	M	BM	0.43	0.35	0.10
Mannose	MB	MB	MB	bMB	MB	MB	9.95	6.50	3.17

It can be seen that the repeatability of the RRFs for the automatic setting is significantly poorer than that of the other methods. This is attributable, as shown in Table 4-6, to different peak classifications being made by the software with different injections, even though it is the same standard that is being injected onto the system. The peak-classification situation under the automatic system is therefore likely to be even less stable when unknown samples are presented to the system and the RRFs for ESTDs, that might differ in sugar concentrations from the unknowns, are used.

The “TEST” method provides an improvement for all peaks. Since only two baseline points are drawn for all the non-fucose sugars it is probably the most stable method for analysing samples with significantly different sugar concentrations from the ESTDs. However, a problem was noted with this method in that the software sometimes automatically included a baseline contact between the glucose and xylose peaks because it considered these peaks to be baseline-resolved. This was particularly the case for samples with relatively low glucose/xylose contents, such as peats. Therefore the “TEST” method required more examination by the user in order to ensure that it was being consistently followed.

The BMB method required the least amount of checking in this regard, with observations only necessary concerning the ending of the mannose peak when that sugar was present in very low quantities. It can be seen from Table 4-6 that this method gave the most reproducible results for the repeat injections of a sugar standard. Whether this advantage would also occur for unknown samples would depend on their concentrations and the relative proportions of sugars. Samples that were similar to the ESTD would perform well, but differences in the height and angle of the baseline drawn between each peak (i.e. the valley-to-valley baseline system) would vary, and affect area calculations if sugar concentrations/proportions were significantly different.

It was decided that the BMB system would be used as standard, but care would be taken that ESTDs were of a similar composition to the samples being analysed.

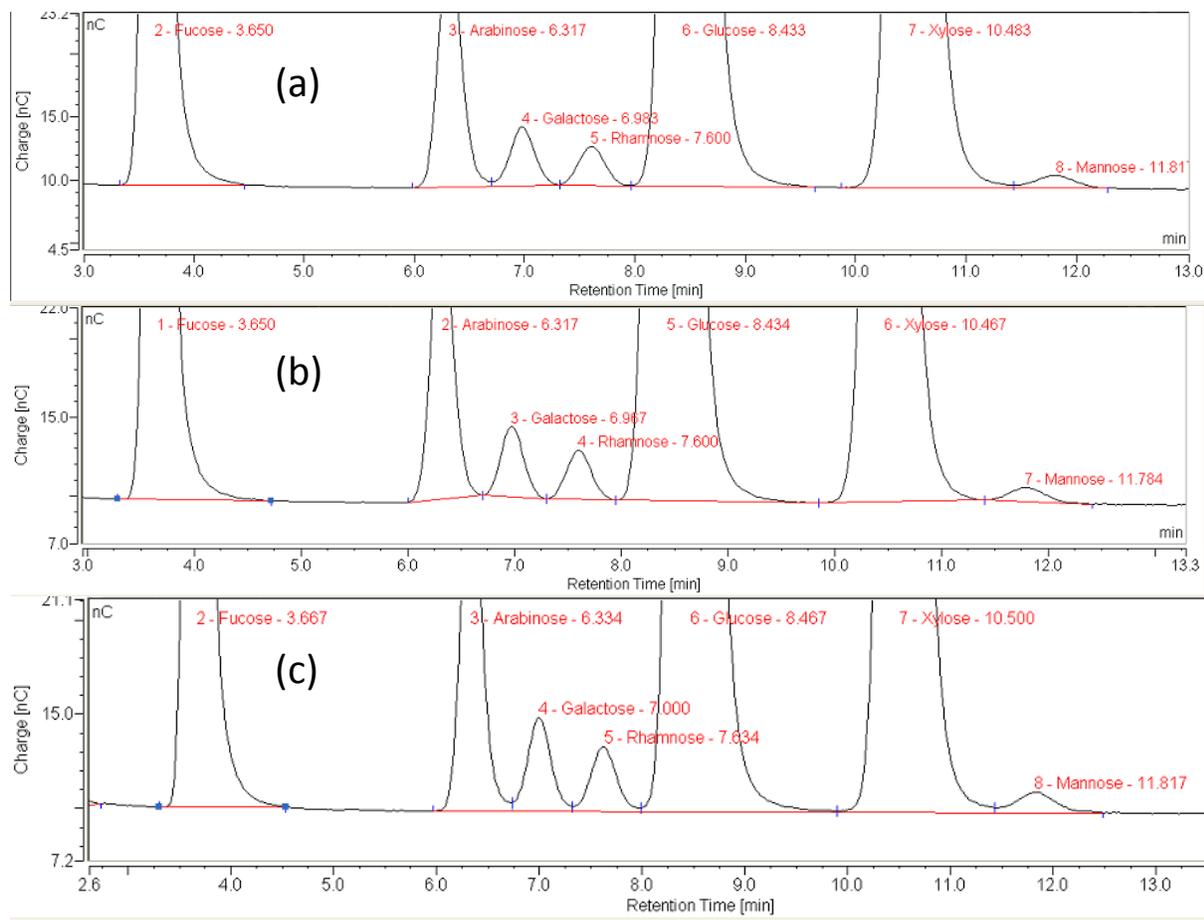


Figure 4-15: Zoomed-up chromatogram focusing on the peak delimiters for the seven sugars. (a) The classification made under the Automatic setting (injection 2); (b) The classification made under the “BMB” setting; (c) The classification made under the “TEST” setting.

4.7 Summary

This chapter has described the principles behind the operation of IC systems and focused particularly on the application of these technologies to the analysis of carbohydrates. This understanding of the advantages offered by IC for monosaccharide analysis resulted in the Author deciding that funding for such a system should be sought so that analytical data of the highest quality, accuracy, and precision

could be obtained. This application for funding was ultimately successful and a Dionex ICS-3000 system was selected as the instrument most suitable for the analytical requirements of the project. Following the installation of this system a considerable amount of time was spent trying to develop an accurate chromatographic method for the analysis of the monosaccharides in biomass hydrolysates. The end result was a monosaccharides analysis method that, at the time of writing, has been successfully run for over two years and allowed the development of successful NIRS calibration equations, with low standard errors of prediction, for the contents of these sugars in the original biomass feedstocks. The Author is of no doubt that if such a stable method had not been developed, and instead analytical results were based on GC data or on the carbohydrates protocol originally suggested by Dionex (Section 4.6.2.1), the quality of the ultimate NIRS calibrations would be significantly poorer.

5 Theory Behind Near Infrared Spectroscopy

5.1 Electromagnetic Spectrum and Spectroscopy

The electromagnetic (EM) spectrum of an object covers the range of frequencies of electromagnetic radiation that are emitted or absorbed from that object. Electromagnetic waves are typically described according to their wavelength (λ), wavenumber ($\bar{\nu}$), frequency (ν) and the energy of the photon (the unit of radiation). The wavelength is the distance (typically expressed as nanometres, nm, in infrared (IR) spectroscopy) between equivalent points on successive waves. The frequency is the number of waves per second (usually in Hertz) and is inversely proportional to the wavelength, as derived from the equation below:

$$\nu = c/\lambda \tag{5.1}$$

Where c is the speed of light in a vacuum (2.998×10^8 m/s)

Wavelengths of EM radiation can vary from a subatomic scale (in the order of 10^{-12} m with gamma rays) to several thousand metres in the case of radio waves. Correspondingly, the frequencies vary from 10^4 Hz with radio waves to over 10^{20} Hz with gamma-rays.

The wavenumber, $\bar{\nu}$, of a wave is equal to the reciprocal of the wavelength and is taken to represent the number of wavelengths in a given distance unit. For IR spectroscopy typically units of reciprocal centimetres (cm^{-1}) are used. The following equation shows how the wavelength, in nm, may be calculated from a given wavenumber (in cm^{-1}).

$$\lambda = \frac{10,000,000}{\bar{\nu}} \tag{5.2}$$

The energy of a photon is directly proportional to the frequency of the wave (and to the wavenumber), and hence inversely proportional to the wavelength as outlined in the equations below:

$$E = h\nu \tag{5.3}$$

$$E = \frac{hc}{\lambda} \quad (5.4)$$

$$E = hc\bar{\nu} \quad (5.5)$$

Where h is Planck's Constant (6.26×10^{-34} Joule seconds)

Also, by linking Equations (5.3) and (5.5):

$$\nu = c\bar{\nu} \quad (5.6)$$

$$\bar{\nu} = \frac{\nu}{c} \quad (5.7)$$

EM radiation is classified according to wavelength/frequency with the various parts of the spectrum assigned according to the effects that they have when they interact with matter. For example, ultraviolet radiation can result in the excitation of the valence electrons of a molecule (see Section 3.2), and infrared radiation can result in molecular vibrations and rotations. A small portion of the EM spectrum is classified as visible light (that which can be detected by the human eye). Infrared radiation is of a longer wavelength (and hence lower frequency and lower energy) than visible radiation.

Spectroscopy is the study of the interaction between matter and radiated energy (usually EM radiation). There are numerous types of spectroscopy according to the manner of interaction.

In absorption spectroscopy the EM radiation is emitted from a radiative source and is absorbed by the analyte of interest. When molecules are irradiated with EM waves there is the potential for an energy change within the molecule. The rotational, vibrational, or electronic energy of the molecule can be changed if the energy of the photon is exactly equivalent to the energy needed for this transition. Therefore whether a transition occurs will be dependent on the frequency of the wave (Murray and Williams, 1987). The devices used for spectroscopic analysis typically record spectra corresponding to the absorbance over the wavelength region of interest (for example between 180 and 400 nm for ultraviolet spectroscopy).

This thesis principally uses near-infrared spectroscopy (NIRS) and the principles of this technique and the interaction between NIR radiation and matter will be outlined in this Chapter. Ultraviolet spectroscopy is used to a lesser degree and its principles were outlined in Section 3.2.

5.2 Infrared Radiation

Infra means below in Latin, since infrared (IR) radiation is of a lower energy than the visible part of the EM spectrum. The IR region extends from wavelengths of 750 nm to wavelengths of approximately 1 mm (1×10^6 nm). IR radiation is typically further subdivided into three main groups.

- Far Infrared (wavelengths between 10,000 nm and 1×10^6 nm) - This has the longest wavelengths in the IR part of the EM spectrum (with even longer wavelengths being associated with the microwave region). Far IR radiation is usually absorbed by rotational modes in molecules that are in the vapour state.
- Mid Infrared, MIR, (wavelengths between 2500 nm and 10,000 nm) – Photons at these frequencies are typically absorbed by the fundamental transitions of molecular vibrations. It is also the radiation level at which hot objects (black-body radiators) radiate strongly (hence the use of IR heat-sensors within this range).
- Near Infrared, NIR, (wavelengths between 750 and 2500 nm) – This is the highest energy region of the IR spectrum, and it can excite overtone or combination vibrations in molecules.

5.2.1 Interaction of IR Radiation with Matter

Near-monochromatic light that is produced by a radiative source can interact with solid matter in a variety of different ways as outlined in Figure 5-1. There can be specular reflection (a) at the surface of the sample, or reflection can be diffuse (b) whereby the light travels partially through the sample but is ultimately reflected back out of the sample in a direction opposite to that of the incoming radiation. The light can also be transmitted (d) or refracted (e) through the sample, or its energy can be lost via internal scattering (f) within the sample. Alternatively, the energy of the light can be absorbed (c) by the chemical bonds of the material.

Whether this absorption occurs will depend on whether the photons have a frequency which is resonant to (matches) that of the rotations or the characteristic vibrations of the chemical bonds that constitute the molecules of the sample.

Due to the nature of chemical bonds and the electrostatic charges involved, molecules are constantly in motion. The two main types of motion are vibrational and rotational energy transitions (Murray and Williams, 1987). Rotational motions result from rotation about the molecular axes. However, rotational absorbances are only usually observable in spectroscopy when samples are in a vapour state since molecular collisions in the more condensed states (liquid, solid) tend to dampen out any resonant rotation (Murray and Williams, 1987). Vibrational motions involve a change in the distance between two atoms and can occur in two main ways: stretching and bending. Stretching vibrations involve motions along the axes of the bonds and can either be symmetrical or asymmetrical. Bending vibrations can involve changes in the bond angles between atoms, or relative changes in the bond angles between a group of atoms and the rest of the molecule. These bending vibrations can occur either out of plane (wagging and twisting motions) or in-plane (scissoring and rocking motions) (Shenk et al., 2008). Figure 5-2 shows some of the potential vibrations and absorptions of the alcoholic hydroxyl group. Note that these absorptions are the overtones of the fundamental transitions. Stretching vibrations tend to occur at higher frequencies (shorter wavelengths) than bending vibrations, with the frequencies decreasing from scissoring to wagging to twisting and rocking vibrations (Shenk et al., 2008).

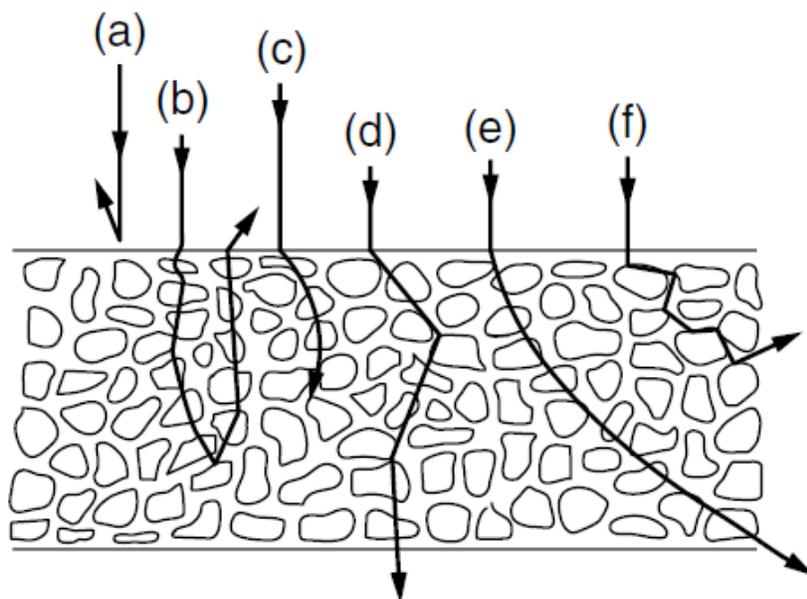


Figure 5-1: Potential interactions that may occur between IR radiation and a solid sample. Taken from (Shenk et al., 2008). (a) specular reflectance, (b) diffuse reflectance, (c) absorption, (d) transmittance, (e) refraction, (f) scattering.

Each type of vibration possible within a molecule is known as a vibrational mode and, as the number of atoms within a molecule increases, so do the total number of possible (fundamental) vibrational modes.

If n is the number of atoms in a molecule, there are $3n-5$ degrees of vibrational modes (vibrational degrees of freedom) in a linear molecule and $3n-6$ degrees of vibrational modes in a non-linear molecule. For this non-linear molecule $n-1$ of these vibrations are stretching motions and $2n-5$ are bending motions. While there are more bending vibrations than stretching vibrations, the stretching vibrations tend to be the most intense.

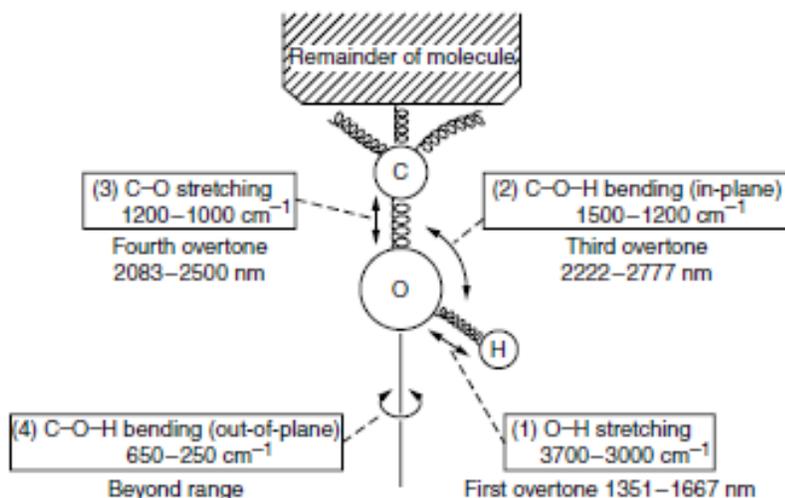


Figure 5-2: Some potential vibrations and absorptions of the alcoholic hydroxyl group. These absorptions are the overtones of the fundamental transitions. Taken from (Shenk et al., 2008).

It therefore follows that, for the complex molecules that constitute agricultural and lignocellulosic materials, there will be a large number of vibrational degrees of freedom that could potentially absorb photon energy that is resonant to their frequencies. For example, the non-linear monosaccharide glucose ($\text{C}_6\text{H}_{12}\text{O}_6$) in ring form has a total of 24 atoms and so has 66 degrees of vibrational modes. Furthermore, when glucose is a constituent unit of the polysaccharide cellulose the situation is likely to become even more complex since the vibrational frequencies of bonds are likely to be affected by whether the bond is in an amorphous or crystalline region, and upon the degree of hydration.

Not all of the degrees of vibration modes absorb infrared radiation, it is necessary for there to be a dipole (unequal distribution of electric charge) in the chemical bond. The presence of a dipole means that the bond possesses a local electric field that will strongly couple to the field of any passing light beam. Diatomic molecules like pure hydrogen have no dipole and do not absorb IR radiation whereas molecules such as hydrogen chloride form a dipole and absorb strongly (Murray and Williams, 1987). Furthermore, it is necessary for the displacement of the atoms in the vibrational mode to produce a

change in the dipole moment of the molecule, or in the local group of vibrating atoms, for IR absorption to be possible (Pasquini, 2003). An example can be provided by comparing a symmetrical stretch in a linear triatomic molecule with a symmetrical stretch of a non-linear triatomic molecule. Only in the non-linear case is there a net change in the dipole moment associated with this vibration, hence, there will only be two vibrational modes in a linear molecule (asymmetric stretch and bending) that can absorb IR radiation compared with three (asymmetric stretch, bending and symmetric stretch) in the non-linear molecule. Here the discrepancy can be seen between the number of vibrational degrees of freedom (4 in the linear molecule and 3 in the non-linear molecule) and the number of fundamental bands detectable for the vibrational spectra of these samples (2 in the linear molecule and 3 in the non-linear molecule).

The dipole change observed in the vibrational displacement of atoms is also of importance in that its magnitude is, in part, associated with the intensity of a given absorption band. The magnitude of a dipole, which is usually expressed in Debye units, is a function of either charge in the dipole multiplied by the charge spacing or molecular distance (Murray and Williams, 1987).

5.2.2 Harmonic Oscillator

A two-atom stretch can be considered to be approximately similar to the situation involving two spherical masses (m_1 and m_2) connected by a spring with a given force constant (k). The stretching vibrations, and their characteristic frequencies and associated energy, can be calculated based on these parameters (Pasquini, 2003). Assuming a harmonic oscillator (whereby sinusoidal oscillations occur about the equilibrium point), the potential energy, V , associated with a given displacement, x , of the two atoms from the equilibrium (minimum energy position) can be represented, according to Hooke's Law, by (Pasquini, 2003):

$$V = \frac{1}{2}kx^2 \tag{5.8}$$

Equation (5.8) is said to be applicable where the changes in the internuclear distances between the two atoms are less than 10%. The potential energy curve described is represented in Figure 5-3. According to Hooke's Law, the vibrational frequencies of the ground state, ν_0 , of two atoms vibrating as such a diatomic harmonic oscillator may be calculated as (Ciurczak, 2001):

$$\nu_0 = \frac{1}{2\pi} \sqrt{\frac{k}{\mu}} \quad (5.9)$$

Where μ is the reduced mass:

$$\mu = \frac{m_1 m_2}{m_1 + m_2} \quad (5.10)$$

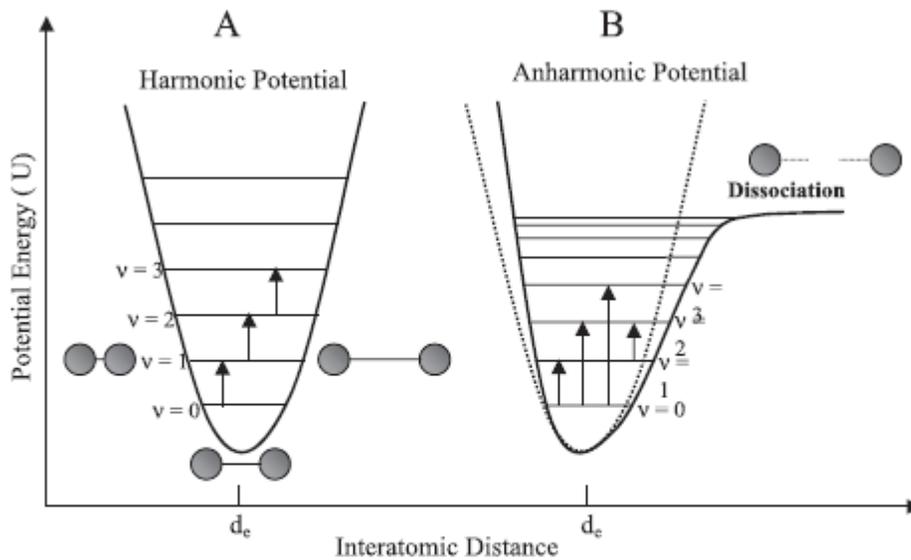


Figure 5-3: Representation of the harmonic (A) and anharmonic (B) models for the potential energy of a diatomic molecule. d_e represents the equilibrium interatomic distance. Taken from (Pasquini, 2003).

Given Equation (5.3), the energy level associated with this ground state can be written as:

$$E = h\nu_0 \quad (5.11)$$

Or

$$E = \frac{h}{2\pi} \sqrt{\frac{k}{\mu}} \quad (5.12)$$

The force constant, k , typically increases in value from single bonds ($\sim 5 \times 10^5$ dynes/cm) to double bonds ($\sim 10 \times 10^5$ dynes/cm) to triple bonds ($\sim 15 \times 10^5$ dynes/cm) (Ciurczak, 2001). Hence, for example, the stretching frequency for $C\equiv N$ is greater than that for $C-N$ (but still lower than a $C-H$ stretch).

Hydrogen bonding can also alter the force constant and so affect the frequencies of vibrations, generally $X-H$ stretching bonds move to longer frequencies and $H-X$ bending vibrations move to shorter frequencies (Murray and Williams, 1987).

The equations outlined above would imply that a wide range of photons could potentially be absorbed since there are a potentially infinite number of values for x (internuclear displacement). However in the microscopic level of molecular bonds, quantum mechanical laws are involved and these dictate that the system may only have discrete energy levels (Ciurczak, 2001).

Simplified versions of the values for the ground state ($n=0$), and subsequent excited states ($n=1$, $n=2$, etc.) can be calculated based on a derivation of the Hooke's law terms (Ciurczak, 2001):

$$E_n = \left(n + \frac{1}{2}\right) \frac{h}{2\pi} \sqrt{\frac{k}{\mu}} \quad (5.13)$$

Or

$$E_n = \left(n + \frac{1}{2}\right) h\nu_0 \quad (5.14)$$

A polyatomic molecule can contain a lot of different energy levels, but the molecule can be approximated as a series of diatomic, independent harmonic oscillators. The situation can then be generalised as (Ciurczak, 2001):

$$E(n_1, n_2, n_3, \dots) = \sum_{i=1}^{3N-6} \left(n_i + \frac{1}{2}\right) h\nu_0 \quad (5.15)$$

In this harmonic model transitions can only occur between adjacent energy levels ($\Delta n = \pm 1$); for example, the transition from the ground state ($n=0$, the state at which most molecules are at room temperature) to $n=1$. This transition is called the **fundamental**. These transitions will occur when the energy of the EM radiation matches the difference between the two adjacent energy levels. i.e. (Pasquini, 2003):

$$\Delta E = E_{n0} - E_{n1} = \Delta n h \nu_0 \quad (5.16)$$

Since in the harmonic model Δn must be equal to 1, a photon energy of $h\nu_0$ is required for the transition.

As well as the potential energy curve being symmetrical and parabolic, a further property of the harmonic model is that the difference in energy between two adjacent states is the same. This is illustrated in Figure 5-3 (a) where the harmonic model and the allowed energy transitions are illustrated.

5.2.3 Anharmonic Oscillator

The harmonic model is not entirely representative of the energy levels associated with variations in interatomic distances, particularly when these variations are large (i.e. greater than fundamental vibrations). This is due to: (1), repulsive forces between the vibrating atoms as they move closer together; and (2), the possibility of dissociation when the vibrating bond is strongly extended (Siesler, 2008). These effects are illustrated in the shape of the potential energy curve for an anharmonic oscillator, Figure 5-3 (b), being different from that of the parabolic and symmetrical potential energy curve for the harmonic oscillator, Figure 5-3 (a). The barrier for decreasing distances approaches rapidly while the barrier at the far end of the stretch slowly approaches zero (Ciurczak, 2001). Furthermore, the energy levels are not equally spaced, as they are in the harmonic model, but instead the energy differences between subsequent levels becomes progressively less with greater changes in intermolecular displacements as shown in Figure 5-3 (b).

The anharmonic model allows transitions of more than one energy level. Hence, in addition to fundamental transitions, “overtones” are possible. These are classified according to the change in the energy level that occurs, with a first overtone representing a transition from the ground state to the second energy level ($n=0 \rightarrow n=2$), a second overtone representing ($n=0 \rightarrow n=3$) and a third overtone representing ($n=0 \rightarrow n=4$). In the harmonic model, if overtones were permitted, it would be easy to predict the frequencies of irradiated light that would allow these transitions, with that for the first overtone being double that of the fundamental, that for the second overtone being triple that of the fundamental, etc. In the anharmonic model a correction is needed to accommodate the differences discussed in the potential energy curve and the quantum energy levels.

The Morse function, as given below, approximates the potential energy, as drawn in Figure 5-3 (b), of an anharmonic diatomic molecule (Pasquini, 2003):

$$V = D_e [1 - e^{-a(r-r_e)}]^2 \quad (5.17)$$

Where a is a constant for a given molecule, D_e is the spectral dissociation energy, r_e is the equilibrium distance between the atoms, and r is the distance between the atoms at a specific point in time. Solving this equation under quantum mechanical principles gives the following equation for the vibrational levels possible (Pasquini, 2003):

$$E_n = hv_0 \left(n + \frac{1}{2} \right) - x_m hv_0 \left(n + \frac{1}{2} \right)^2 \quad (5.18)$$

It can be seen that this equation represents a modification (the subtracted term) of the equation given earlier for a harmonic oscillator. The constant x_m is the anharmonicity constant of vibration.

If a calculation of the wavelength of an overtone is required based on the known wavelength of the fundamental, then the equation is:

$$\lambda_n = \frac{\lambda_1(1 - 2x_m)}{n - x_m n(n + 1)} \quad (5.19)$$

The effect of the anharmonicity constant is to give slightly longer wavelengths (or lower frequencies) for overtones than would be expected for a harmonic model. The constant typically varies between 0.5 and 5%.

Fundamental absorptions usually occur in the mid infrared region (2,500 nm to 15,000 nm). The overtones, however, can potentially occur in the NIR region. First overtones will occur at a little over half the wavelength of the fundamental (with the differential from half being dependant on the anharmonicity constant), while second overtones will occur at a little over a third of the wavelength of the fundamental.

The intensities of absorption bands partly depend on the value of the anharmonicity constant with vibrations that have low constants having lower overtone intensities (Siesler, 2008). Conversely, several bonds involving the hydrogen atom (e.g. O-H, C-H, N-H, and S-H) tend to have a high degree of anharmonicity (Pasquini, 2003). Given that these bonds also experience a large dipole change during

vibrations, the overtones of the fundamental vibrations of these bonds (the fundamentals occur at 3000-4000 nm) are some of the most intense regions in NIR absorption spectra (Pasquini, 2003).

However, the absorption bands for overtones are significantly weaker than for those for the corresponding fundamentals. This is because most molecular vibrations are in the ground state and the probabilities of transitions decrease greatly with each additional energy level. Murray and Williams (1987) suggested that, for a typical fundamental, the intensity of the absorbance for the first overtone would be approximately 10 times less, that for second would be approximately 300 times less and the intensity for the third overtone would be approximately 10,000 times less. It is suggested (Ciurczak, 2001) that the absorbances of overtones higher than the third would be so small as to be of little relevance to quantitative NIRS (particularly given the complex nature of the NIR spectrum of samples whereby the more intense absorbances of lower order overtones of other bonds may mask the signal).

Nevertheless, the presence of up to three overtones per fundamental vibration shows the added complexity of the potential absorbances associated with a molecule or sample (the $3N-6$ rule only applies to the number of fundamental vibrations, not their associated overtones).

5.2.4 Combination Bands and Resonance Effects

In addition to the absorbances of photons with resonant frequencies associated with fundamental transitions ($n=0 \rightarrow n=1$) and overtones ($n=0 \rightarrow n=2,3,4\dots$), the anharmonic model allows combination bands. In these two or more vibrations can combine (through addition or subtraction of the energies) to give a single band. In order for absorption to take place, the energy of the photon must equal the sum of the energy of the coupling vibrations (Murray and Williams, 1987). A combination band can involve both fundamentals and overtones. However, given the lower intensity of overtone absorbances, the absorbance intensities of combination bands involving overtones are likely to be significantly lower than those only involving fundamentals. Also, the absorbance intensities of combination bands decrease as more independent vibrations are used for its derivation (Siesler, 2008). In the NIR the strongest combination bands are found between the wavelengths of the C-H stretching vibration and its first overtone (Murray and Williams, 1987).

Importantly, the condition that there must be a change in the dipole moment associated with a bond displacement for IR absorption to occur, only need apply to one of the vibrations associated with the combination band. Hence, some vibrations that may not be seen in mid infrared spectra, may be observable, as part of a combination band, in NIR spectra (Pasquini, 2003).

A special case involving combination bands or overtones occurs in which they have the same symmetry and similar energy to a fundamental. The two absorbances interact, in a situation known as Fermi Resonance (Siesler, 2008). In a normal situation the absorption for the fundamental would be of a greater intensity than the nearby combination-band/overtone, but the interaction results in the two intensities becoming somewhat normalised and a shifting in frequencies to those higher and lower than the expected positions of the fundamental and overtone/combination-band (Siesler, 2008). The closer the two vibrations exist initially, the greater their eventual distance in the spectrum (Ciurczak, 2001). The phenomenon of Fermi Resonance can be easily identified in MIR (mid infrared) but may be difficult to observe in the NIR region as the effect may be hidden in the broad overlapping peaks associated with NIR absorbance spectra.

5.3 Near Infrared Spectroscopy

Typically most of the fundamental vibrational absorptions occur in the MIR region, at wavelengths between 2,500 and 15,000 nm. The MIR is often called the “fingerprint region” in that quite distinctive spectra can be obtained for specific molecules and compounds. Hence, the qualitative analysis of samples is relatively easy and can often be accomplished with a visual inspection of the spectra obtained. In contrast, the NIR region (750-2500 nm) consists of the much weaker absorbance intensity overtones and combination bands. The main absorbance bands in the NIR are the second or third harmonics of fundamental O-H, C-H and N-H stretching vibrations found in the MIR. The hydrogen atom is present in these main absorbance bands due to the high anharmonicity of the associated bond vibrations and the large dipole of the chemical bonds that it forms. The number and complexity of the NIR overtones and combination bands, and their interactions (resonances), result in spectra that cannot easily be evaluated on visual inspection.

However, the weaker absorptions that occur in the NIR can be considered to be an advantage in that they allow the radiation to penetrate further into the sample than with MIR; hence minimal or no sample preparation steps are needed (in contrast to IR spectroscopy) and signal to noise ratios are usually higher in the NIR (hugely important for accurate quantitative analysis). Hence, NIRS offers the potential to develop quantitative analytical techniques for samples that require minimal sample preparation; however, the challenge lies in developing methods for correlating the spectral variations with the physiochemical variations in the samples.

5.3.1 History of NIRS

The NIR spectral region was discovered by William Herschel in 1800 (Herschel, 1800). He used a prism to disperse sunlight and measure the differential heating effect of each part of the visible spectrum by use of a thermometer. In one experiment the thermometer was placed beyond the location of the red light and it was noted that the temperature change was greatest in this region. Little happened for decades following this although, in 1835, Ampere demonstrated, using a thermocouple, that (in contrast to Herschel's earlier beliefs) NIR radiation had the same optical characteristics as visible light. In 1881, Abbey and Festing, using the photographic plate, recorded the spectra of organic liquids in the range 1000-1200 nm (Abney and Festing, 1881) and, in 1905, Coblentz built a spectrometer, using a rock salt prism and a sensitive thermopile connected to a mirror galvanometer, and recorded the spectra of hundreds of compounds using the wavelength range of 1,000 to 15,000 nm (Coblentz, 1905). While conducting these analyses Coblentz noticed that, while no two compounds had the same spectrum, there were distinct patterns, for example all compounds with hydroxyl groups absorbed in the 2,700 nm (MIR) spectral region.

There were similar experiments to those of Coblentz by numerous researchers in subsequent years. However, much of the work focussed on the less complex MIR region and there was very little progress in the use of NIRS for quantitative analysis until the late 1960's. Qualitative research activity was spurred by technological advances, including the commercial availability in the 1950s of lead sulphide as a sensitive heat detector for the NIR region, and the development of tungsten filament lamps as NIR radiative sources.

The complexity of NIR spectra necessitates the use of automated methods for spectral simplification and the development of quantitative methods. Hence the increase, over the years, of the use and power of computers has facilitated the utilisation of NIRS as a quantitative analytical tool.

NIRS was identified in the 1960s by Karl Norris as an analytical tool for the rapid determination of wheat properties, and in 1968 he and Ben-Gera published work regarding the application of multiple linear regression to the calibration of agricultural products (Ben-Gera and Norris, 1968). Since then, and facilitated with advancements in chemometric methods, the use of NIRS in the agricultural sector has expanded greatly and there are now thousands of calibration equations that have been developed for various quality parameters (e.g. moisture, protein, oil content, sucrose content) for a wide range of agricultural products (wheat, soya beans, sugar cane etc.). NIRS has also been used for the quantitative analysis of a wide range of lignocellulosic properties for numerous feedstocks as outlined in Section 9. As well as a laboratory or at-line tool, NIRS is also used as an online process analytical tool in various agricultural sectors such as sugar mills (see Section 12).

5.3.2 NIRS Instrumentation

There are a wide variety of potential configurations for a NIR spectrometer. However, in the basic sense, all devices will require: a radiative source; a wavelength selection system; a means by which this radiation can interact with the sample; and a detector to record the reflected/transmitted radiation. There are a variety of NIR measurement modes but these are all based on either the detection of the radiation that is reflected from the sample (reflectance spectroscopy) or on the detection of the radiation that is transmitted through the sample (transmission spectroscopy).

Figure 5-4 outlines the main NIR measurement modes, that are currently employed in commercial devices. Transmittance, Figure 5-4 (a) occurs, as in Ultraviolet-Visible spectroscopy, whereby the sample is placed in a glass or quartz cuvette with a path length of between 1 and 50mm, the portion of the radiative light that is not absorbed passes through the sample and reaches the detector. Instruments that work in transmittance mode are preferred for analysis of whole grains and they generally operate over the low wavelength range, from 570-1098 nm (for example the FOSS Infratec 1241 Grain Analyser (FOSS, 2011)) .

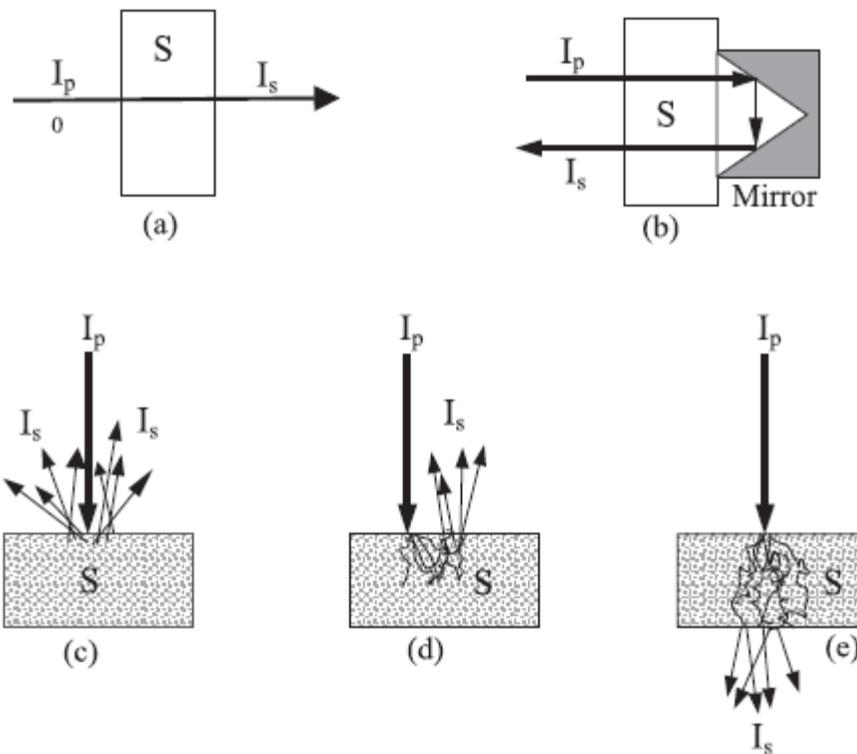


Figure 5-4: Some of the main modes of measurement employed in NIRS. (a) transmittance; (b) transreflectance; (c) diffuse reflectance; (d) interactance; (e) transmittance through a scattering medium. Taken from (Pasquini, 2003)

Figure 5-4 (b) illustrates a special mode of transmittance which is termed transreflectance. This mode usually involves an optical probe that is inserted into a sample and, after passing through the sample, the light is reflected back through the sample by a mirror, so doubling the path length. Figure 5-4 (e) also shows a more recently developed mode for the measurement of the transmittance through dense solid samples. Due to internal scattering, the effective path length is much greater than the actual thickness of the sample (approximately 65 times greater according to Johanson *et al.* (2002) who studied the analysis of pharmaceutical tablets) and so the resulting absorbance spectra may, in some cases, be more representative of the average sample content than diffuse reflectance spectra.

Diffuse reflectance, Figure 5-4 (c), is the NIRS measurement mode that is applied to many solid samples. It is the mode that is utilised by the FOSS XDS system that is installed at the Carbolea laboratory, see Section 5.3.2.3.

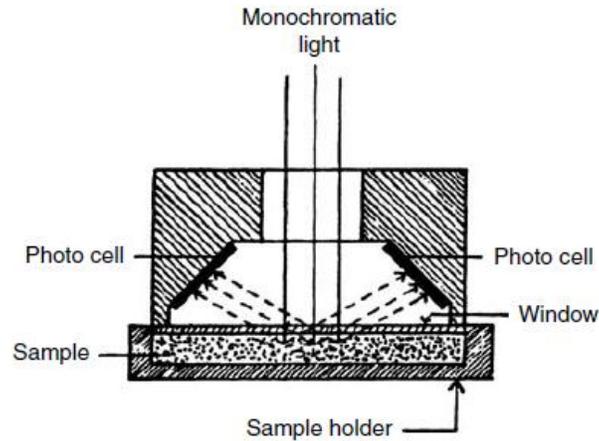


Figure 5-5: Representation of the sample presentation, light radiation, and reflected light detection of a diffuse reflectance system. Taken from (Norris, 1984).

Figure 5-5 illustrates the set-up of a diffuse reflectance system. Monochromatic light irradiates the sample at normal incidence (0° angle). Typically only approximately the first 1-4 mm of the sample will be penetrated. Radiation of a shorter wavelength is likely to penetrate further into the sample than radiation of a longer wavelength; however, the shorter wavelength NIR radiation (800-1400 nm) is also more susceptible to front surface scattering (Workman and Burns, 2001). A smaller particle size will also increase front surface scattering as discussed in Section 8.1. Two or more detectors are placed at 45 degree angles to the sample. These are typically lead sulphide for the 1100-2500 nm region while PbS “sandwiched” with silicon photodiodes are most often used for the visible region (Workman and Burns, 2001).

5.3.2.1 Wavelength Selection Systems

The main wavelength selection systems are listed below:

- Discrete filter
- Moving diffraction grating monochromator
- Fixed-grating monochromator with diode array detector.
- Acousto-optical Tunable Filters (AOTF)
- NIRS LEDs

- FT-NIRS interferometer.

Filter NIR instruments have specific filters (from 6 to dozens) that interfere with the incident light, only letting pre-specified wavelength ranges through. These can be rugged devices. However, if the right filters are not present some calibrations may not be possible. These devices tend to be used in commercial devices designed for specific applications where the calibrations have already been developed and so the required wavelength ranges and filter types are known. In exploratory calibration development for untested feedstocks and parameters (such as the work in this Thesis) such devices are unlikely to be suitable.

Dispersive (grating) systems use a rotating mirror/grating device to disperse the white light across a wider area. The controlled movement of the grating focuses a limited bandwidth of light on the sample at any one given period. Hence the detector can record the amount of light in this bandwidth that is reflected from the sample at that time. This is an example of a pre-dispersive instrument. These systems are explained in more detail when the NIR system used in this project is discussed (Section 5.3.2.3). Post-dispersive instruments also exist. In these the white light first goes to the sample and the reflected/transmitted light is then sent to the grating where the light is separated into the various wavelengths before striking the detector – the UV-VIS device instrument that is used in the Carbolea laboratory (see Section 3.2.2) is an example of a post-dispersive device.

An alternative to having a system with a moving dispersive grating and a single detector is to have a stationary grating that disperses the light to the sample and an array of detectors all tuned to specific wavelengths (diode array detectors). Since no moving parts are involved, the system is likely to be more robust and can be placed in harsh environments (e.g. on agricultural harvesters) that may be unsuitable otherwise. Such a system also allows the sample to be scanned in much less time than with traditional moving dispersive systems (a few milliseconds versus 20 seconds or more). Section 3.2.2 describes diode array detectors in the context of UV-Vis spectroscopy.

FT-NIR spectrophotometers use an interferometer to modulate the NIR signal. A beam splitter sends half of the irradiated light to a fixed mirror and the other half to a moving mirror. These beams are then reflected back to the beam splitter and interfere with each other with the degree of interference depending on the difference in path lengths. This results in the selection of appropriate wavelengths for sample analysis. FT-NIR was examined for use in this Thesis, but no system could be found with a large enough sample device to allow for the analysis of large biomass samples of heterogeneous particle sizes.

5.3.2.2 Reference

All NIR spectrometers need to scan a reference as well as the sample. This is due to the fact that the response of the instrument to an analysis is not only dependent on the properties of the sample but also on various environmental conditions (e.g. air humidity, temperature) as well as the condition of the instrument itself (lamp temperature, detector response, stray light, etc.). Taking a reference allows many of these external influences to be reduced or eliminated. Various reference materials can be used depending on the mode of analysis. The FOSS XDS unit has a ceramic standard as its internal reference material.

General practice in the Carbolea laboratory involves 2-3 reference scans being taken throughout the day. The laboratory is a closed room with no windows and therefore large variations in temperature are not experienced during the day.

5.3.2.3 FOSS XDS Unit

The author was responsible for the purchase at the University of Limerick of an NIR system suitable for the needs of the project (i.e. the prediction of various lignocellulosic properties of minimally processed biomass samples). After examining the various options it was decided that the (single beam) FOSS XDS monochromator, with the associated Rapid Content Analyser (RCA) module and its Solids Transport attachment, would be obtained. Figure 5-7 (a) shows a functional flow diagram for this monochromator and Figure 5-7 (b) provides a photograph of the inside. A tungsten halogen filament lamp (which has a life expectancy of approximately 8,000 hours) is used as the radiative source. Light then enters the monochromator enclosure via the lamp housing exit slit and is directed onto the grating by a mirror after the entrance slit.

The grating is a concave holographic grating, ruled with fine lines to separate the white light into individual wavelengths. It is rotated by a positioning motor, and hence the rotational position is correlated with the wavelength. The resolution of the grating is specified as 0.5 nm. An effect of the diffraction grating is that many orders of radiation are produced, each one comprising a spectrum over a

certain angle. There occurs a partial overlap of the different orders of radiation which will result in different wavelengths from different orders appearing in the same position. For example, the wavelengths 2400 nm, 1200 nm and 800 nm from the 1st, 2nd and 3rd orders, respectively, might be superimposed. The XDS only uses first order radiation, since this is the most intense. Hence it is necessary to block the radiation from the other orders. This is accomplished with order-sorter filters which are mounted in a paddle whose position is kept in sync with the grating position motor. Hence only first order radiation is let through. The four filters are: 400-700 nm, 700-1100 nm, 1100-1700nm, and 1700-2500nm.

The light leaves the grating and is reflected towards the order sorter filter off of a second mirror. The light then passes through the order sorter and fills the exit slit with the selected light to pass, via optical fibre, to the RCA module.

The XDS has a series of modules that are designed according to the type of samples or analysis that will be required. These connect to the monochromator and receive the light that emerges from its exit slit. These modules contain their own detectors and amplifiers. The RCA module has a reflectance detector. This incorporates (Figure 5-6) 4 silicon detectors (for the wavelength range 400-1100nm) and 4 lead sulphide detectors (for the wavelength range 1100-2500 nm). These detectors are mounted at a 45 degree angle to the sample surface (which is irradiated by the light at normal incidence). The signal from the detector is then sent to the amplifier where it is amplified and, digitised, and sent to the computer associated with the system.

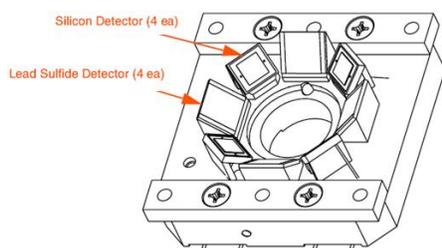


Figure 5-6: A schematic of the reflectance detector and the two types of sensors used in the FOSS XDS RCA. Taken from (FOSS, 2006b)

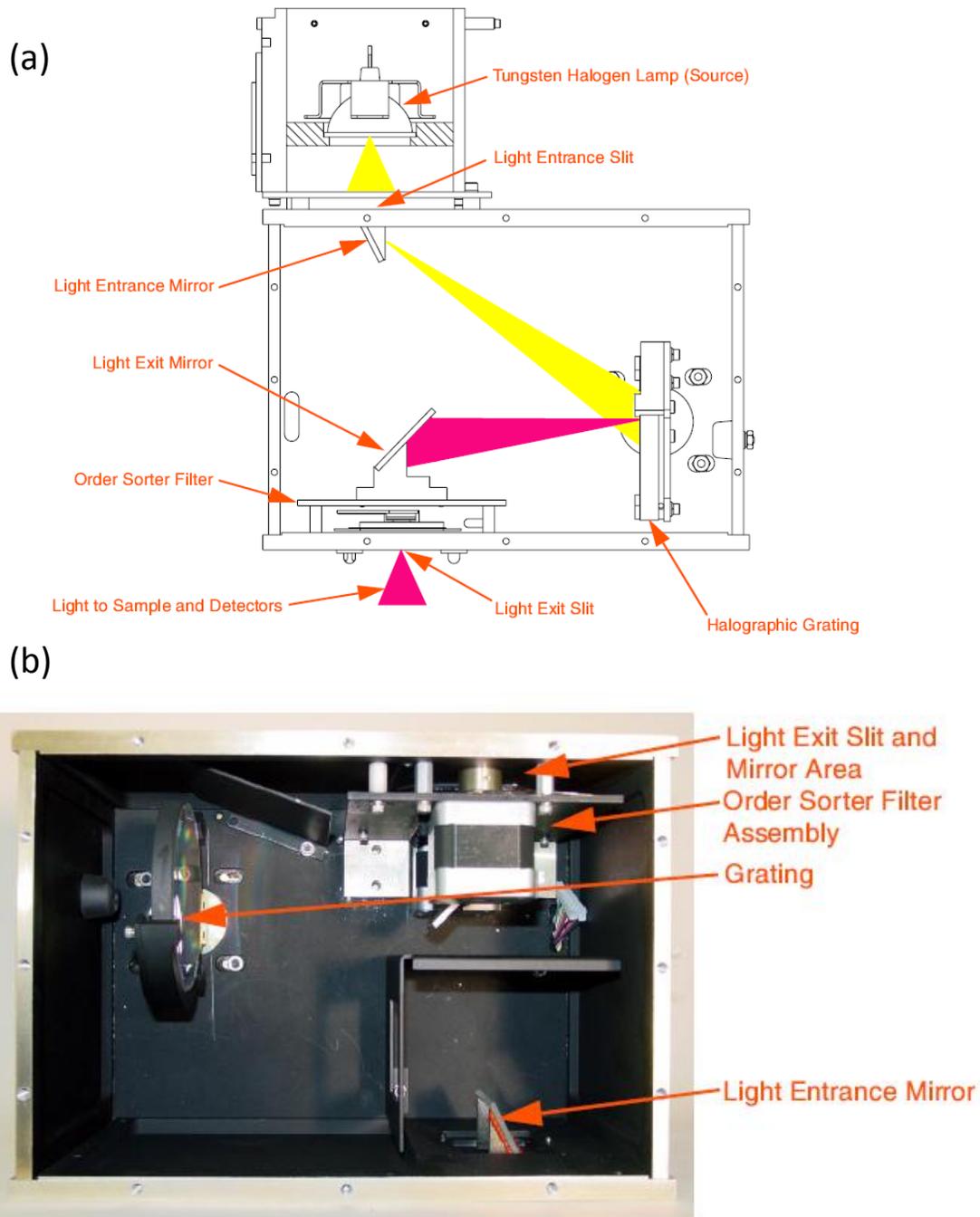


Figure 5-7: The FOSS XDS monochromator. (a) An interior schematic of the FOSS XDS monochromator; (b) A photograph of the insides of the monochromator. Both taken from (FOSS, 2006b).

In a conventional set-up an iris is locked in place in the RCA module and used to focus a small circular cell over the scanning window so that spectra can be collected from the sample contained within this cell. The capacity and surface area of this cell are very small and would preclude the analysis of anything but well comminuted and homogeneous samples. This would prevent the development of NIR calibration equations for more heterogeneous samples that have larger particle sizes. The RCA transport assembly solves this problem by allowing material within a much larger sample cell, the “coarse sample cell”, to be scanned. This rectangular cell is placed in the locking mechanisms of the wheeled solids-transport device and spectra are taken through the window. The transport assembly then moves, by means of a motor device, to a second position, which represents a separate part of the coarse cell to that which was previously scanned, and further spectra are collected. This process repeats six more times and at the final iteration the last part of the window of the coarse sample cell has been scanned and the transport assembly then returns the cell to its starting position.

Data collection with the FOSS XDS unit requires that a Data Collection Method (DCM) is set-up with all of the requisite parameters set. A summary of the parameters and the settings that were used in the Carbolea laboratories for the analysis of solid samples is included in Table 5-1.

Table 5-1: The data collection parameters that are available and the settings that are used in the FOSS XDS in the Carbolea laboratory.

Parameter	Explanation/Range	Settings Used
Wavelength Range	Region over which data is collected	400-2500 nm (full range)
Module	Module linked to monochromator	RCA with spot size
Cell type	e.g. moving, stationary, cell size	Moving full size
Spot size	For sample analysis (varies from 9.5 to 17.25 mm)	17.25 mm
Detector	Reflectance or transmission	Reflectance
Reference standardisation	Get standardized spectra	On
Instrument calibration	Adjust wavelength profile to an external, traceable wavelength standard	On
Use auto linearization	Software calculates fine tuning corrections to the linearization constants when a reference is scanned	On
Number of scans of the sample	The number of scans that will be taken to give an average spectra	32, equal to 4 per position with the solid transport device
# scans of reference material	This can be set to the internal reference material or an external	32 of the internal reference material

Calibration of the XDS

Approximately every month, or following the installation of a new bulb or change in configuration of the system, the XDS is calibrated to ensure that wavelength accuracy is maintained. This procedure involves the following steps.

1. Wavelength Linearization – This uses an internal wavelength standard to determine a set of peak positions that the instrument will use to maintain repeatability of wavelength response (FOSS, 2004). These peaks are plotted on a different axis (encoder pulses) that does not relate directly to wavelength but the linearization process fits wavelengths to this axis.
2. Reference Standardisation – This step is used to minimise any difference in instrument-to-instrument response that could be attributed to differences in the reflectance reference material. It involves a photometric standard of known reflectivity being scanned and the internal ceramic reference is then also scanned. The differences of the ceramic standard from 100% reflectivity are mapped and a photometric correction generated (FOSS, 2004). This correction is then applied to every spectrum collected by the device so that it would appear that each spectrum was taken with a reference of 100% reflectance.
3. Instrument Calibration – An external standard, with known peak positions, is scanned by the unit, and the method aligns the wavelengths to exact positions.

Performance Test

A performance test is carried out automatically by the system every night, it involves the testing of: photometric noise in several wavelength regions; internal wavelength performance (using an internal reference material); and the precision of these internal wavelength measurements.

Noise: Instrument noise should only be present as random spectral variability. If there is some structure in the noise spectrum, or if high noise is often caused by changes in the environment there may be an issue with the instrument that will need correcting (FOSS, 2004). The performance test involves taking the average of 32 scans and this is typically repeated ten times. Following the test a report is available, with several statistics related to noise, with a statistic presented for the three different wavelength regions:

- Peak to peak (PP) Noise: This is the difference between the largest and smallest value in the noise spectrum (expressed in milliabsorbance units).

- Bias: The average absorbance value of all points in the noise spectrum.
- RMS: The root mean square of the amplitude of the noise.

The DCM has critical limits for these statistics, and if they are breached the performance test will fail. This happened several times in the XDS unit at the Carbolea labs. However, since the performance test is taken daily there is no concern that “poor” spectra may have been collected while the unit was not performing to standard. Upon observing a failed performance test all data collection on the unit was stopped until the issue has been rectified.

Wavelength Accuracy/Precision: The internal standard on the XDS unit is comprised of polystyrene, Erbium Oxide and Samarium Oxide and six peaks across the range from 500-2310nm are used to determine the internal wavelength performance of the instrument (FOSS, 2004).

5.3.2.4 Important Terms for NIR Devices

Some important terms regarding the performance of an NIR instrument are listed below:

Bandpass/Bandwidth - In grating instruments this is defined as the convolution of the dispersion of the grating and the slit functions for both the entrance and exit slits. It determines the resolution of an instrument. It can be defined as the full width at half maximum (FWHM) of the bandshape of monochromatic radiation passing through a monochromator. The smaller the bandwidth, the higher the resolution.

Spectral Resolution – This can be determined as the measurement of the separation of closest pair of absorption lines that can be separately identified (Kaye, 1975). It is reported that this is 2 nm for the FOSS XDS.

Spectral Data Interval – The gaps over which data points are collected, in the XDS it is every 0.5 nm.

Wavelength accuracy - The difference between the measured wavelength of a wavelength standard and the wavelength reported for that standard by the equipment. This is reported as being less than 0.05 nm for the FOSS XDS.

Wavelength precision – Repeatability of wavelengths on the same instrument, given as < 0.005 nm on the XDS. Wavelength precision can also be measured based on a group of instruments in which case the value for the XDS is reported as < 0.02 nm.

Photometric accuracy - The difference between the photometric (brightness) value recorded on the analysis of a standard reference material and the nominal value for this analysis.

Photometric/dynamic range - The range from the highest useful absorbance that can be reproduced to the lowest detectable absorbance.

Baseline flatness - This is equal to the difference between the maximum and minimum values in a baseline scan (Workman and Burns, 2001).

Baseline drift - The change in photometric value of a spectrometer's baselines at specific wavelength with respect to time.

Scan speed – Time taken to cover all the analytical wavelengths. The XDS has a data acquisition rate of 2 scans per second.

Stray light – This is the sum of any energy/light, other than the chosen wavelength, that reaches the detector. It is reported to be $< 0.1\%$ at 2300 nm for the XDS.

5.3.3 Linearity in NIRS

The linearity of response in the NIR region, and hence the adherence to Beer's Law (see Section 6.1), may vary for several reasons.

- Linearity of response of the detector used in the instrument.
- Differences in path length caused by particle characteristics (see Section 8.1).
- Temperature effects – the effect of temperature can be complex with differing effects in different wavelength regions and with different absorbers. The effects can be bathochromic (whereby the absorbance bands are shifted to longer wavelengths) or hypsochromic (shifting of the band to a shorter wavelength) (Murray and Williams, 1987)

Sections 6.10 and 8 details the treatments available for correcting for deviations from linearity.

5.3.4 Important Regions of NIR Spectra

Absorption bands can be defined according to their location, height, and width. The absorption peak is the highest point in the band. However, since NIR absorption spectra are usually represented by $\log(1/\text{reflectance})$ (see Section 6.1) there tends to be only a few definable peaks. This is also due to the complexity of the NIR region with its numerous overlapping overtone and combination bands, as well as the interferences that occur between these. Quantitative analysis therefore requires the use of chemometric techniques that involve multiple wavelengths/factors and often involve spectral pretreatments (see Section 8). Qualitatively, though, some generalisations can be made:

- There tends to be confounding of the 2300 to 2500 nm baseline with mid infrared tails trailing into the NIR (Shenk et al., 2008).
- Most of the fundamental oscillations of the bound hydrogen atom fall just outside the NIR in the wavelength range of 2700-2600 nm (Murray and Williams, 1987).
- Hence, most of the overtones of these fundamentals are found between 400 and 2200 nm (Murray and Williams, 1987)
- There is a multiplicative response to changes in particle size (see Section 8.1).
- There is a confounding of the 750 to 1400 nm region with information from the visible region (Shenk et al., 2008).
- Specular reflectance is greater at shorter wavelengths.
- The area between 1,900 and 2,500 nm is referred to as the combination region, since many of the absorbances here come from combination bands (particularly those involving bonds with hydrogen atoms).
- The region between 1,100 nm and 1,900 nm is referred to as the overtones region (Murray and Williams, 1987).

There has been some work in trying to assign the positions of overtones and combination bands in the NIR region for various molecular bonds of importance in agricultural materials (Roberts et al., 2004, Marten et al., 1988, Clark and Lamb, 1986). This work usually involves the assignment of the positions of the fundamental absorbances in the MIR and then the calculation of where the associated overtones would be. This is an imprecise method due to the anharmonicity (described earlier) and the interaction

that occurs between the constituents of a material, as well as other effects, such as moisture, temperature, and hydrogen bonding, that will affect absorbance frequencies. Similarly, possible combination bands are estimated by considering which combinations are feasible and what wavelengths these would occur at (the sum/difference of the various energy levels of the various oscillations that could make up the combination band).

Regarding cellulose, hemicellulose, and lignin there have been some attempts to assign spectral regions to oscillations of the bonds of these molecules. Some of these are represented in Table 5-2.

For lignin, one of the most prominent absorption bands is the first overtone of the C-H stretching vibration of the aromatics at around 1672 nm, an absorbance region that coincides with the absorption band for the first overtone of the C-H stretch in xylan acetyl groups (Krongtaew et al., 2010a). There can also be a strong lignin band at 1446 nm, associated with the first overtone of the O-H stretch in the phenolic hydroxyl groups. However, once again this band is coincident with other bands, such as those involving the first overtones of O-H stretching in amorphous polysaccharides (Krongtaew et al., 2010a).

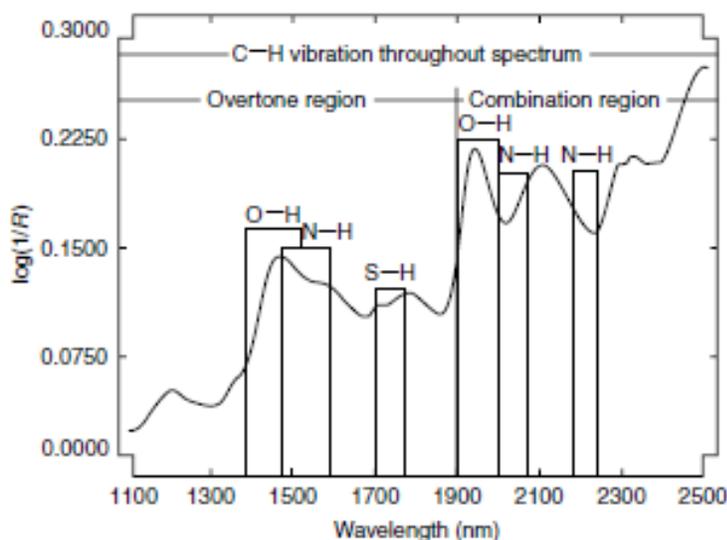


Figure 5-8: A general representation of the X-H vibrational bands that can be present in agricultural products. Taken from (Shenk et al., 2008)

Table 5-2: Tentative band assignments for lignin and the structural polysaccharides. * second derivative, ** overlaps with H₂O, *** these two overlap

Band Assignment	Structure	λ (nm)	Reference
Lignin			
C-H stretch 2 nd overtone		1170	(Shenk et al., 2008)
O-H stretch 1 st overtone		1410	(Shenk et al., 2008)
C-H stretch combination		1417	(Shenk et al., 2008)
O-H stretch combination		1420	(Shenk et al., 2008)
C-H stretch combination		1440	(Shenk et al., 2008)
C-H stretch 1 st overtone		1685	(Shenk et al., 2008)
O-H stretch 1 st overtone	Phenolic hydroxyl group	1446.5	(Mitsui et al., 2008)
2xC-H stretch + C-H deformation	Aromatic	1446	(Krongtaew et al., 2010a)
C-H stretch, 1 st overtone	Aromatic	1672	(Shenk et al., 2008)
C-H stretch, 1 st overtone	CH ₂	1724***	(Shenk et al., 2008)
Polysaccharides			
O-H stretch, 1 st overtone	Alcoholic	1410-1610	(Shenk et al., 2008)
O-H stretch, 1 st overtone	Amorphous polysaccharide Free, or weakly H bonded, OH	1430	(Tsuchikawa and Siesler, 2003a, Watanabe et al., 2006, Mitsui et al., 2008, Tsuchikawa and Siesler, 2003b)
	OH groups with H-bonds of intermediate strength	1434-1470	(Watanabe et al., 2006)
O-H stretch, 1 st overtone	Crystalline cellulose	1480	(Yonenobu et al., 2009)
O-H stretch, 1 st overtone	Semi-crystalline cellulose	1488	(Tsuchikawa and Siesler, 2003b, Tsuchikawa and Siesler, 2003a)
	O6-H6...O3" interchain H-bonds	1510	(Watanabe et al., 2006)
	O3-H3...O5" interchain H-bonds	1547	(Watanabe et al., 2006)
O-H stretch, 1 st overtone	Crystalline cellulose C _I	1548	(Tsuchikawa and Siesler, 2003b, Tsuchikawa and Siesler, 2003a)
O-H stretch, 1 st overtone	Crystalline cellulose C _{II}	1592	(Tsuchikawa and Siesler, 2003b, Tsuchikawa and Siesler, 2003a)
	H-bonds in the cellulose I _{β} allomorph	1591	(Watanabe et al., 2006)
C-H stretch, 1 st overtone	CH ₃ , acetyl	1669 and 1678*	(Fackler et al., 2007)
C-H stretch, 1 st overtone	CH, furanose or pyranose due to hemicellulose	1724***	(Tsuchikawa et al., 2005, Mitsui et al., 2008)
C-H stretch + H-O-H deform. combination	Cellulose and water	1780	(Shenk et al., 2008)
O-H stretch, OH bend	Polysaccharides	1920**	(Shenk et al., 2008)
O-H combination	Polysaccharides	2090	(Shenk et al., 2008)
O-H stretch/C-O stretch combination	Polysaccharides	2270	(Shenk et al., 2008)
C-H stretch/C-H deform combination	Polysaccharides	2329	(Shenk et al., 2008)

For hemicelluloses, there are various absorption bands listed in Table 5-2, including the O-H stretching first overtone and the first overtone of the C-H acetyl stretch referred to above, as well as the C-H

stretch first overtone of hemicellulosic furanoses/pyranoses at 1724 nm (however, once again this absorbance is coincident with a lignin-associated absorbance band).

Cellulose is also a much more complex absorber than may perhaps be expected. This is due to the variations in the crystallinity of the polysaccharide and, as shown in Table 5-2, some of the hydrogen bonds themselves have suspected absorption bands. There is an interesting paper by Krongtaew *et al.* (2010a) in which wheat straw and other lignocellulosic materials undergo various pretreatment methods (for example combination alkaline/H₂O₂ pre-treatment) and the NIR spectra of the resulting products are compared with the spectra of the virgin materials. Changes in the crystallinity of the samples are reflected in changes in the intensity of NIR absorption in related regions of the spectra.

It is important to reiterate the differences between NIR and MIR and the lack of distinct fingerprint regions in NIR spectra of lignocellulosic materials. This complexity necessitates the use of chemometric methods to determine the relationship between variations in the spectra of a material and variations in the concentrations of chosen components. Looking at the spectra or using literature values for wavelength values for absorption bands that are said to be characteristic of this component are unlikely to be adequate for any kind of meaningful analysis due to the matrix effects of complex materials and the environment. Hence, it is chemometrics, the latent variables that they can generate and the regressions that can result, that will inform us of correlations between a component and spectral wavelengths.

5.3.5 Effects of Water and Hydrogen Bonding

According to the $3n-6$ rule, there should be three fundamental vibrations with the water molecule: (1) symmetrical stretch of the O-H bonds, (2) asymmetrical stretch of the O-H bonds, and (3) a scissoring bending motion. Murray and Williams (1987) obtained transmittance spectra of liquid water and of water in its unassociated state (by dissolving it in the non-polar solvent carbon tetrachloride). In the dissociated state the fundamental O-H symmetrical and asymmetrical stretching vibrations were observed at 2,760 and 2,695 nm, respectively, with the first overtone bands being seen at approximately 1,390 nm and the combination band of the O-H bend and stretch (asymmetrical) being seen at 1,895 nm. However, the situation of water in its associated (hydrogen bonded) state is more complex. Figure 5-9 provides the NIR spectra (between 1100 and 2500nm) for pure water (in log 1/R form) and its fourth-order derivative. Large peaks can be seen at the previously referenced wavelengths. However,

these absorption peaks are very broad and the fourth derivative shows a more complex number of peaks (not all of which could be artefacts of the derivatisation). It is clear that the absorbances are not distinct easily identifiable absorbance bands.

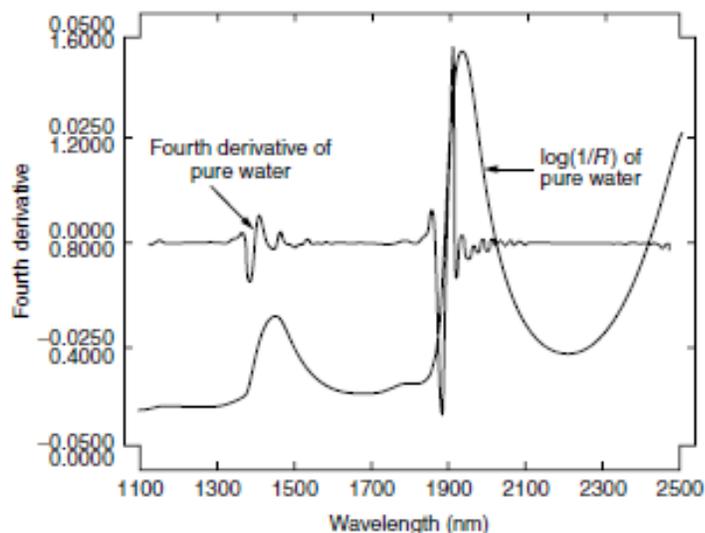


Figure 5-9: The spectrum of pure water in transmittance form ($\log(1/R)$) and its fourth-order derivative. Taken from (Shenk et al., 2008).

The influence of hydrogen bonding between the water molecules is substantial and it is said that water molecules should be considered as more like a polymer represented by $(\text{H}_2\text{O})_x$ (Murray and Williams, 1987). Hydrogen bonding results in changes of the force constants of X-H bonds and this can cause broadening and peak position shifts (Shenk et al., 2008). The change in the force constants of the fundamental vibrations is most pronounced in the X-H stretching vibrations. In the case of water the force constant of the O-H bond decreases and, hence, the frequency of the O-H absorption band will decrease meaning that it moves to a longer wavelength.

Furthermore, in their comparison of the absorbances of associated and unassociated water, Murray and Williams (1987) found that association tended to increase the intensity of the fundamental O-H band (which in effect can be considered as the O-H.....O band) relative to the unassociated O-H band. The opposite was the case for the first overtone of this band. Since the overtone is of a lower intensity it could be considered that the association reduces the anharmonicity of the vibration. Furthermore, hydrogen bonds can be disrupted by molecular collisions, and this results in a broadening of the O-H absorption band. The apparent band locations of these vibrations are also said to be isolinear and change position linearly with temperature (Shenk et al., 2008).

There are higher order overtones and more combination bands possible in the absorbance spectra of water than those mentioned above, and in longer pathlengths these extend into the visible region (hence the blue colour of water and ice (Murray and Williams, 1987)). All apparent water absorbances will in fact be composites of the individual bands that constitute them and, hence, the location of the composite bands will vary according to the composition of the mix of absorbance bands.

This complexity of the spectrum of water has clear implications for the analysis of solid samples that have varying levels of moisture and it is for this reason that many research papers that have focussed on the development of NIR calibration equations for lignocellulosic materials only take NIR spectra of dry materials. This limits the relevance of NIRS for rapid online analysis, however.

Regarding hydrogen bonding in solids, an increase in non-hydrogen bonded species within a sample may have several effects on the spectral pattern, such as shifts to shorter wavelengths and band narrowing. At the extremes, shifts of the magnitude 21-50 nm, may be observed when comparing samples with a high degree of hydrogen bonding and samples with significant hydrogen bond breakage (Shenk et al., 2008). The strength of the hydrogen bond will depend on both the electron withdrawing power of the atom to which the hydrogen atom is covalently linked as well as the electron donating properties of the atom to which the hydrogen is bridged. For example, nitrogen does not form as great a dipole with hydrogen as oxygen but nitrogen is a stronger electron donor (Murray and Williams, 1987).

5.3.6 Effect of Temperature on NIR Spectra

The temperature of both the sample and the environment also has effects on NIR spectra. Shenk *et al.* (2008) compared the spectra of hay samples at two different temperatures (23°C and 26°C). The fourth-order derivatives of these spectra, along with the fourth-order derivative for the transreflectance spectra for pure water, over the wavelength range 1890 nm to 1990 nm, are presented in Figure 5-10 (a). The corresponding spectra for corn grain samples are presented in Figure 5-10 (b). The authors focused on the effects on the O-H band positions and shapes. It can be seen in both cases that cooling causes the individual band at 1900 nm to be reduced and the individual band at 1928 nm to increase. This resulted in a shift of the composite 1930 nm band, it was considered that this change was a result of changes in hydrogen bond formation and bond breakage (Shenk et al., 2008).

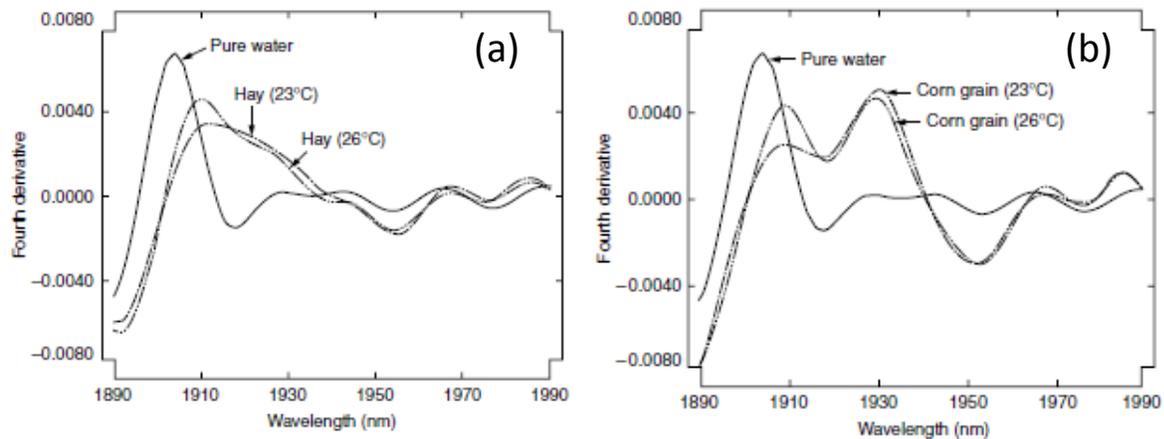


Figure 5-10: influence of a 3°C change in room temperature. (a) hay absorbance spectra; (b) corn grain absorbance spectra. Both Figures present the spectra as fourth-order derivatives and also include a fourth order derivative spectrum of pure water. Taken from (Shenk et al., 2008).

5.4 Summary

This Chapter has summarised the principles behind the absorbance of near infrared radiation by molecular bonds and described the spectrophotometer that has been used at the Carbolea laboratories to measure this absorbance. The key point regarding NIR spectra is that there is complexity due to the wide array of potential overtones and combination bands that will exist for the numerous bonds present in lignocellulosic materials. This complexity is increased when the variable effects of moisture and temperature on the spectra are considered. However, what can be seen as a disadvantage for NIRS compared with longer wavelength IR radiation is actually its greatest advantage, since radiation is less strongly absorbed in the NIR it is able to travel deeper into the sample and allow more representative analysis. The effect of water is strong but not as overpowering as it is at longer wavelengths and, as shall be shown in subsequent sections, quantitative analysis is still possible with high moisture materials. The coarse sample cell on the FOSS XDS unit is of key importance in allowing the user to develop such “dirty” calibrations since it provides a much greater surface area over which to collect the spectrum meaning that the sample need not be as homogenous as would otherwise be required.

There are numerous resources available in the literature citing characteristic absorbance frequencies of lignocellulosic bonds and these can certainly play a role in understanding the reasons why calibration

equations place such importance on particular wavelength regions. However these literature absorbance values cannot be taken as certainties, matrix effects will vary their intensities, locations, and relative importance to the total absorbance of the sample. Subsequent chapters will explain how the development of latent variables and the use of spectral preprocessing techniques can lead to case-specific determinations of what the most important wavelengths and absorbing groups will be in the discrimination between sample types or quantification of lignocellulosic components.

6 Quantitative Calibration Methodologies Applicable to NIRS

6.1 The Beer-Lambert Law

The basis of spectroscopic quantitative calibration methods for the determination of the properties (e.g. concentration of an absorbing constituent in a sample matrix) of a sample is that a relationship can be derived between those properties and the spectral variance within a data set. Various methods are used to mathematically relate the component concentration to the absorbance of electromagnetic radiation by molecules. In ultraviolet and visible (UV-Vis) transmission spectroscopy, the Beer-Lambert law is used to relate the transmission (T) of light through a liquid or gas to the concentration of an absorbing analyte according to the equations below:

$$T = \frac{I}{I_o} = 10^{-kcl} \quad (6.1)$$

Where I is the intensity of energy emerging from and I_o the incident energy to the sample, k is the molecular absorption (or molecular extinction) coefficient, c is the concentration of the absorber, and l is the path length of the energy through the sample (Murray and Williams, 1987).

This transmission is usually expressed in terms of absorbance, which is defined as:

$$A = \log_{10} \left(\frac{1}{T} \right) = \log_{10} \left(\frac{I_o}{I} \right) = kcl \quad (6.2)$$

Therefore, if the path length through which the light travels is constant, the absorbance is directly proportional to the concentration of the absorbing species. This is assumed to hold true for all wavelengths at which the species absorbs. There are conditions for this linearity assumption to be fulfilled, however. These include that there must be no scattering of radiation, absorbers must act independently, the incident radiation should be monochromatic, and the sample should be homogeneous throughout the path-length (Murray and Williams, 1987).

It is considered that, in NIR spectroscopy, reflectance (R) is analogous to transmittance, hence the above equation can be written as:

$$A = \log_{10} \left(\frac{1}{R} \right) \quad (6.3)$$

However, the situation with the NIR reflectance analysis of solid samples is clearly distinct from the ideal conditions used to derive the Beer-Lambert law. Firstly, the sample is usually very far from being homogenous throughout and, secondly, there is a significant degree of scattering associated with the solid particles that make up the sample (see Section 5.3). Hence, the detector(s) will measure both the specularly and the diffusely reflected radiation meaning that path length cannot be considered to be constant. The mean path length will be influenced not only by particle size but also by the shape of the particles, their crystallinity, the refractive index of the pore space (which will change depending on whether it is filled by water, or air, or something else), as well as the actual absorbance that may occur within the sample (Murray and Williams, 1987). Clearly therefore, pathlength will change between samples and also within the same sample following a repack of the NIR cell.

However, to date no definite reflectance theory has been formulated, presumably due to the complexity that would be involved in accounting for the near infinite number of light interaction effects associated with the packing of particles, and hence the Beer-Lambert law is used as an approximation (Workman, 2001). However, it is far from ideal and non-linearities are often encountered (Section 6.10). There are methods that can be used to reduce these non-linearities, including spectral preprocessing techniques (Section 8) and various calibration methods that will be outlined in this Chapter.

6.2 Univariate Calibration

A simplified model that only considers the absorbance at one wavelength, assumes that Beer's law applies perfectly, and ignores the path length term, considers the absorbance (A) at that wavelength to be equal to the molar extinction coefficient (k) multiplied by the concentration (c) of the absorbing species. Since there are only two variables, the relationship between A and c can be illustrated with a two dimensional scatter plot of absorbance (x) versus concentration (y) with each point representing the results of each experiment/analysis.

If the relationship is linear, there are two ways in which the relationship between A and c can be calibrated. The first is known as Classical Least Squares (CLS) where the instrument response is modelled as a function of analyte concentration:

$$A = a_0 + a_1c + f \quad (6.4)$$

Or, for sample i .

$$\hat{A}_i = a_0 + a_1c_i \quad (6.5)$$

Where f is a random error term and the coefficients (a_0 and a_1) would be determined by least squares regression. Once determined it will be necessary to invert the CLS equation in order to predict the concentration of an unknown substance based on its absorbance value. This is shown in the following equation (Naes et al., 2007):

$$\hat{c} = -(\hat{a}_0/\hat{a}_1) + (1/\hat{a}_1)A \quad (6.6)$$

The other method is known as **Inverse Least Squares (ILS)**. This inverts the Beer Lambert relationship at the first stage in order to model concentration as a function of instrument response (absorbance):

$$c = b_0 + b_1A + error \quad (6.7)$$

This model can then be fitted to the experimental data points using least squares and the resulting equation for the line of best fit be used directly for prediction:

$$\hat{c} = \hat{b}_0 + \hat{b}_1A \quad (6.8)$$

It is important to note that the least squares method for regression is only considered to be valid when the following four assumptions hold (Mark, 2001a):

- 1) There is no error in the independent (X) variables.
- 2) The error in the dependent variable (y) is normally distributed.
- 3) The error in the dependent variable is constant for all values of the variable.
- 4) The system of equations is linear in the coefficients.

6.3 Multivariate Calibration

As discussed in Section 5.3, the NIR spectra of samples are complex and peak positions and shapes can change according to variations in the sample matrix, temperature, particle size, etc. A calibration involving a single wavelength term (univariate) is therefore likely to be a poor predictor for variations in

component concentrations between samples. The ILS model for component concentration (y) can be modified to incorporate multiple explanatory variables (wavelengths in this case, represented as x_k) according to the equation below (Naes et al., 2007):

$$y = b_0 + \sum_{k=1}^K b_k x_k + f \quad (6.9)$$

Where K = the number of x wavelengths used. Hence there are $K+1$ calibration coefficients (the intercept term b_0 plus a coefficient to apply to the absorption value at each wavelength). Each experimental result provides values for the x and y terms in the above model, hence a series of observations (n) resulting from the analysis of various samples can, though the least squares method, provide a series of linear equations that can be solved for the coefficients. This is the basis of multiple linear regression (MLR).

An alternative representation of the model above, and one that allows model simplification in the multivariate case, is to put it in matrix form.

$$\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{f} \quad (6.10)$$

Where \mathbf{y} is a vector representing the component concentrations in each experiment, \mathbf{X} is a matrix of all the observations (samples \times variables), \mathbf{b} is a vector of all the coefficients, and \mathbf{f} is a vector incorporating the error for each prediction. This representation (as elaborated below) allows the separation of the predictor variables, in \mathbf{X} , from their coefficients, in \mathbf{b} (Naes et al., 2007):

$$\begin{pmatrix} y_1 \\ y_2 \\ \cdot \\ y_N \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & \cdot & x_{1K} \\ 1 & x_{21} & \cdot & x_{2K} \\ \cdot & \cdot & \cdot & \cdot \\ 1 & x_{N1} & \cdot & x_{NK} \end{pmatrix} \begin{pmatrix} b_0 \\ b_1 \\ \cdot \\ b_K \end{pmatrix} + \begin{pmatrix} f_1 \\ f_2 \\ \cdot \\ f_N \end{pmatrix} \quad (6.11)$$

The major advantage of the matrix representation is that it allows the use of matrix algebra to determine the least squares estimates of the coefficients according to Equation (6.12), which is discussed in more detail in Section 6.4.4:

$$\hat{\mathbf{b}} = (\mathbf{X}^t \mathbf{X})^{-1} \mathbf{X}^t \mathbf{y} \quad (6.12)$$

\mathbf{X}^t represents the transpose of matrix \mathbf{X} and $(\mathbf{X}^t\mathbf{X})^{-1}$ is the inverse of $(\mathbf{X}^t\mathbf{X})$. For the matrix inversion in the above equation to take place, it is necessary that the number of samples/observations (n) is greater than the number of variables (k). Furthermore, not all matrices have inverses. If a matrix maps two different vectors to the same result then it cannot have an inverse and is called **singular** (see Section 6.4.1). This will occur if the subspace spanned by its rows/columns is less than the full-space – i.e. vectors are mapped to a lower dimensionality. Hence, matrix inversion requires **non-singular** (or **full-rank**) matrices.

If there is a high degree of colinearity between the predictor variables, i.e. they are strongly correlated with each other, then the matrix is singular and any solution for the regression coefficients is non-unique and can be highly unstable.

Even if there are strong correlations between the predictor variables but no exact relationship, the matrix may have an inverse in principle, but its computation can be highly numerically unstable (Naes et al., 2007).

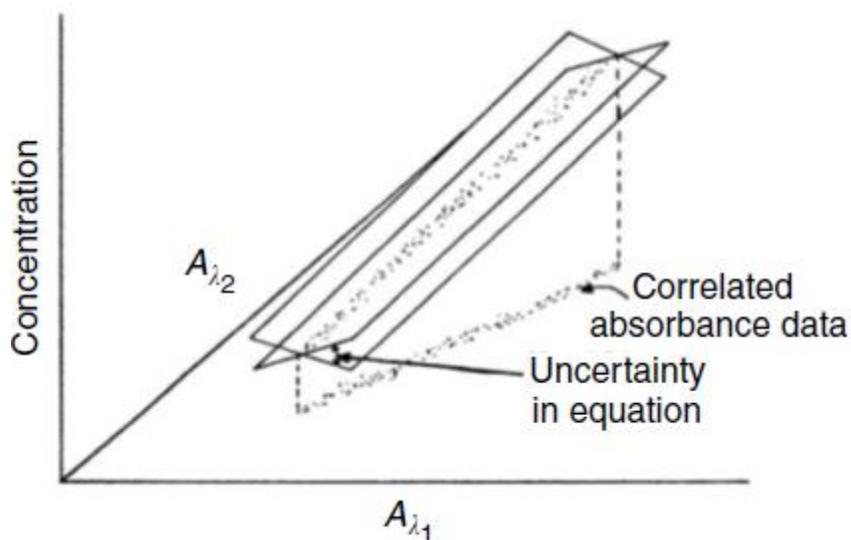


Figure 6-1: A regression plot for MLR using the absorbances at 2 wavelengths (A_{λ_1} and A_{λ_2}). The correlation between optical variables results in an unstable calibration plane meaning that calibrations between different data sets could give greatly different calibration coefficients. From (Mark, 2001a).

Colinearity between predictor variables is particularly relevant in NIR spectroscopy since there can be high correlations between different wavelengths, particularly with those that are close to each other. This is illustrated, for a two-wavelength calibration, in Figure 6-1. If the data from the two wavelengths

were not correlated the optical data would be spread cross the 2-wavelength plane. However, instead, the high intercorrelation that exists means that the reference data for the constituent is concentrated in a narrow region meaning that small differences between two sets of data (as may result from random lab error) would result in different tilts in the calibration plane, and thus in greatly different calibration coefficients (Mark, 2001a).

There is therefore an apparent conflict between the need to consider multiple wavelengths (in order to account for the complexity and variation in NIR spectra) and the requirement that calibration equations should be stable and applicable to samples outside of the calibration set. Fortunately there are chemometric techniques that can resolve this conflict.

6.4 Principal Components

The collinearity problem would clearly not exist if there was no relationship between the predictor variables. The fact that correlations do often exist between many wavelengths of NIR spectra means that these are not suitable variables for such a scenario. Instead we target the production of latent (new) variables from the manifest (original) variables such that these variables are orthogonal (uncorrelated) to each other. Two vectors are orthogonal if:

$$\sum X_i X_j = 0 \quad i \neq j \quad (6.13)$$

Principal components (PCs) are a type of latent variable and are constructed on the basis that they account for the largest possible amount of residual variance in the data.

The American Society for Testing and Materials (ASTM) define principal component analysis (PCA) as the following:

“A mathematical procedure for resolving sets of data into orthogonal components whose linear combinations approximate the original data to any desired degree of accuracy. As successive components are calculated each component accounts for the maximum possible amount of residual variance in the set of data. In spectroscopy, the data are usually spectra, and the number of components is smaller than or equal to the number of variables or the number of spectra, whichever is less” (ASTM Standard E131-10, 2010)

The “accounts for the maximum possible amount of residual variance” can be explained by considering that we can fit a PC to each spectrum from the dataset (by subtracting the scaled PC from each spectrum) and then sum all of these residuals according to the equation below:

$$\text{Residual sum of squares (RSS)} = \sum_{i=1}^N \sum_{j=1}^K R_{ij}^2 \quad (6.14)$$

Where K refers to the number of wavelengths in the spectrum, N to all the spectra in the dataset, and R is the residual (Mark, 2001a). According to the criterion outlined above the RSS obtained by using PCs is smaller than that obtained from any other possible set of fitting functions. This follows an order of hierarchy with PC1 explaining more variation in the original spectra than PC2 and so on. PC2 is computed from the residual (deflated) spectra that result after PC1 has been fitted and its computation follows the same principle, meaning that PC2 will give the smallest possible RSS for this set of residual spectra.

6.4.1 Important Statistics and Matrix Operations in PCA

The derivation of PCs requires understanding of several statistical terms.

Firstly, the spread of the data points for a variable, x , can be defined by its **variance** (s_x^2 or $\hat{\sigma}_x^2$) as outlined below:

$$s_x^2 = \text{var}(x) = \sum_{i=1}^N (x_i - \bar{x})^2 / (N - 1) \quad (6.15)$$

This variation can be expressed in terms of the original units of x by calculating the **standard deviation** (s_x or $\hat{\sigma}_x$), computed as the square root of the variance.

In multivariate analysis several variables exist for each sample/object with each variable having its own mean and variance. A **covariance** (degrees of covariation) between a pair of variables can be defined as below (Naes et al., 2007):

$$\text{cov}(x_p, x_q) = \frac{\sum_{i=1}^N (x_{ip} - \bar{x}_p)(x_{iq} - \bar{x}_q)}{N - 1} \quad (6.16)$$

Where x_{ip} is the value of variable p for object i and \bar{x}_p is the mean of variable p . Taking the example of spectra, p and q could be columns of the \mathbf{X} matrix, representing two different wavelengths.

Of course a spectrum will have more than two variables with the actual number, K , corresponding to the number of datapoints (wavelengths) recorded by the spectrophotometer. Since covariance is always measured between two dimensions, this means that there will be a large number of different covariance values needed to summarise a matrix representing spectra. These can be presented in a matrix form, known as the variance-covariance matrix:

$$\hat{\Sigma} = \begin{bmatrix} \text{var}(x_1) & \text{cov}(x_1, x_2) & \dots & \text{cov}(x_1, x_K) \\ \text{cov}(x_2, x_1) & \text{var}(x_2) & \dots & \dots \\ \dots & \dots & \dots & \dots \\ \text{cov}(x_K, x_1) & \dots & \dots & \text{var}(x_K) \end{bmatrix} \quad (6.17)$$

It can be seen that the variance-covariance matrix is a square matrix of size $K \times K$ (dimension K). The covariance matrix, $\hat{\Sigma}$, for a mean-centred matrix \mathbf{X} is equivalent to the cross-product matrix $\mathbf{X}^t \mathbf{X}$.

Similarly, a covariance matrix between the **rows** of \mathbf{X} (an $n \times p$ matrix) can be derived and is equivalent to the cross-product matrix $\mathbf{X} \mathbf{X}^t$ and, hence, will be a square matrix of dimension n .

Eigenvectors: A vector, \mathbf{v} , is defined as being an eigenvector of matrix \mathbf{X} if:

$$\mathbf{X} \mathbf{v} = \lambda \mathbf{v} \quad (6.18)$$

Where λ is a constant. Eigenvectors only occur for square matrices and not all matrices will have these. However those that do have the same number of eigenvectors as their rank. Eigenvectors are also all orthogonal to each other, a crucial requirement to avoid the colinearity problems mentioned in Section 6.3.

Usually eigenvectors are scaled to all have a length of one.

Eigenvalues: Each eigenvector has an associated eigenvalue. This is the constant (λ in the above equation) by which the eigenvector is scaled upon multiplication by the original matrix \mathbf{X} . They can be calculated by performing the following modifications to Equation (6.18):

$$\mathbf{X}\mathbf{v} - \lambda\mathbf{v} = \mathbf{0} \quad (6.19)$$

$$\mathbf{X}\mathbf{v} - \lambda\mathbf{I}\mathbf{v} = \mathbf{0} \quad (\text{where } \mathbf{I} \text{ is the identity matrix}) \quad (6.20)$$

$$(\mathbf{X} - \lambda\mathbf{I})\mathbf{v} = \mathbf{0} \quad (6.21)$$

e.g. if \mathbf{X} is a 2 x 2 matrix $\begin{pmatrix} x_{11} & x_{12} \\ x_{21} & x_{22} \end{pmatrix}$

$$(\mathbf{X} - \lambda\mathbf{I}) = \begin{pmatrix} x_{11} - \lambda & x_{12} \\ x_{21} & x_{22} - \lambda \end{pmatrix} \quad (6.22)$$

If $(\mathbf{X} - \lambda\mathbf{I})\mathbf{v} = \mathbf{0}$ where $\mathbf{v} \neq \mathbf{0}$, then $\mathbf{X} - \lambda\mathbf{I}$ must be not invertible (i.e. it is singular), meaning that $|\mathbf{X} - \lambda\mathbf{I}| = 0$.

Hence,

$$|\mathbf{X} - \lambda\mathbf{I}| = \lambda^2 - (x_{11} + x_{22})\lambda - x_{21}x_{12} + x_{11}x_{22} = 0 \quad (6.23)$$

This can then be solved for the eigenvalues, λ

Diagonalisable matrix: A diagonal matrix is one that has zeros in every entry except along the main top-left to bottom right diagonal. An orthogonal matrix will produce a diagonal matrix upon the product with its own transpose. (i.e. where \mathbf{X} is an orthogonal $n \times p$ matrix, $\mathbf{X}\mathbf{X}^t = \mathbf{D}_n$, where \mathbf{D}_n is a diagonal matrix of dimension n , and $\mathbf{X}^t\mathbf{X} = \mathbf{D}_p$ - in the first instance \mathbf{X} is row orthogonal, and in the second instance \mathbf{X} is column orthogonal). If the matrix \mathbf{X} is orthonormal, the result of the product is the identity matrix, \mathbf{I} (Massart et al., 1998b). An orthonormal matrix has the property that its columns and rows are orthogonal and any individual vector from it has a norm (scalar product with itself) of one.

The product of a matrix with a diagonal matrix results in the multiplication of the rows or columns (depending on the whether there is pre- or post-multiplication of the diagonal matrix) of that matrix by the constants of the diagonal, hence the rows/columns are scaled.

A matrix, \mathbf{B} , is diagonalisable if there exists a matrix, \mathbf{P} , such that $\mathbf{P}^{-1}\mathbf{B}\mathbf{P}$ is a diagonal matrix. i.e.

$$\mathbf{P}^{-1}\mathbf{BP} = \begin{pmatrix} \lambda_1 & 0 & 0 & 0 \\ 0 & \lambda_2 & 0 & 0 \\ 0 & 0 & \dots & 0 \\ 0 & 0 & 0 & \lambda_n \end{pmatrix} \quad (6.24)$$

Hence, by multiplying the above equation by \mathbf{P} , it is clear that the column vectors of \mathbf{P} are the eigenvectors of \mathbf{B} and they correspond to eigenvalues in the diagonal matrix.

Eigenvalue Decomposition/Spectral Decomposition: If a symmetric ($n \times n$) matrix \mathbf{A} is non-singular it means that its columns are linearly independent and, hence, there will be n positive eigenvalues and n normalised and mutually orthogonal eigenvectors. Therefore, the eigenvalue decomposition of such a matrix can be formed by multiplying Equation (6.18) by \mathbf{V}^t , hence (Massart et al., 1998b):

$$\mathbf{V}^t\mathbf{AV} = \mathbf{\Lambda}^2 \quad (6.25)$$

Where $\mathbf{\Lambda}^2$ is an $n \times n$ diagonal matrix where the diagonal values are the n eigenvalues associated with the eigenvectors in the $n \times n$ matrix \mathbf{V} . Equation (6.25) can be rearranged giving:

$$\mathbf{A} = \mathbf{V}\mathbf{\Lambda}^2\mathbf{V}^t \quad (6.26)$$

This is known as the spectral decomposition of \mathbf{A} .

Where \mathbf{A} is a symmetric matrix that is singular, i.e. is not full rank ($r < n$), there will be r positive eigenvalues and \mathbf{V} will be an $n \times r$ matrix.

Singular Value Decomposition (SVD) of a non-symmetric matrix: The SVD theory (Naes et al., 2007) states that a rectangular matrix \mathbf{X} can be decomposed into a diagonal matrix of singular values $\mathbf{\Lambda}$ and two matrices of singular vectors, \mathbf{U} and \mathbf{V} .

$$\mathbf{X} = \mathbf{U}\mathbf{\Lambda}\mathbf{V}^t \quad (6.27)$$

(where $\mathbf{U}^t\mathbf{U} = \mathbf{V}^t\mathbf{V} = \mathbf{I}_r$, with r being the rank of \mathbf{X})

This decomposition can be applied to the cross-products of the ($n \times p$) rectangular matrix \mathbf{X} , i.e. $\mathbf{X}\mathbf{X}^t$ and $\mathbf{X}^t\mathbf{X}$ (the covariance matrices \mathbf{C}_n and \mathbf{C}_p) as follows:

$$\mathbf{XX}^t = \mathbf{C}_n = \mathbf{U}\mathbf{\Lambda}\mathbf{V}^t \mathbf{V}\mathbf{\Lambda}^t\mathbf{U}^t = \mathbf{U}(\mathbf{\Lambda}\mathbf{\Lambda}^t)\mathbf{U}^t \quad (6.28)$$

$$\mathbf{X}^t\mathbf{X} = \mathbf{C}_p = \mathbf{V}\mathbf{\Lambda}^t\mathbf{U}^t\mathbf{U}\mathbf{\Lambda}\mathbf{V}^t = \mathbf{V}(\mathbf{\Lambda}^t\mathbf{\Lambda})\mathbf{V}^t \quad (6.29)$$

The size of the $\mathbf{\Lambda}\mathbf{\Lambda}^t$ matrix will be $n \times n$ and the size of the $\mathbf{\Lambda}^t\mathbf{\Lambda}$ will be $p \times p$, both of these matrices will have the squared singular values (i.e. λ_1^2, λ_2^2 etc.) along the diagonal.

We can therefore see that these two equations are equivalent to the eigenvalue decomposition of a square symmetric matrix, as shown in Equation (6.26). $\mathbf{U}\mathbf{\Lambda}^2\mathbf{U}^t$ is equivalent to \mathbf{C}_n , the cross product matrix of the rows of \mathbf{X} and $\mathbf{V}\mathbf{\Lambda}^2\mathbf{V}^t$ is equivalent to, to \mathbf{C}_p , the cross product matrix of the columns of \mathbf{X} .

Multiplying Equation (6.28) by \mathbf{U} and Equation (6.29) by \mathbf{V} results in the following:

$$\mathbf{XX}^t\mathbf{U} = \mathbf{U}\mathbf{\Lambda}^2\mathbf{U}^t\mathbf{U} = \mathbf{U}\mathbf{\Lambda}^2 \quad (6.30)$$

$$\mathbf{X}^t\mathbf{X}\mathbf{V} = \mathbf{V}\mathbf{\Lambda}^2\mathbf{V}^t\mathbf{V} = \mathbf{V}\mathbf{\Lambda}^2 \quad (6.31)$$

Since both of those matrices are square, \mathbf{U} and \mathbf{V} can be calculated as the eigenvectors of \mathbf{XX}^t , and $\mathbf{X}^t\mathbf{X}$, respectively. For example, Equation (6.30) can be written as a normal eigenvalue equation by defining the i th column of \mathbf{U} as \mathbf{u}_i and the eigenvalues as $\lambda_i = \Lambda_{ii}$, hence:

$$\mathbf{XX}^t\mathbf{u}_i = \lambda_i\mathbf{u}_i \quad (6.32)$$

The square root of the eigenvalues will be the singular values along the diagonal of the $\mathbf{\Lambda}^2$ (singular values) matrix. The square root of the largest eigenvalue being in the first position, the second largest in the second, and so on.

Taking the example of one eigenvector, \mathbf{v} , and its diagonal element (i.e. square root of the eigenvalue) as γ , the law of eigenvectors states that (Massart et al., 1998b):

$$\mathbf{X}^t\mathbf{X}\mathbf{v} = \gamma^2\mathbf{v} \quad (6.33)$$

Multiplying both sides by \mathbf{X} results in:

$$(\mathbf{XX}^t)\mathbf{X}\mathbf{v} = \gamma^2\mathbf{X}\mathbf{v} \quad (6.34)$$

This implies that the eigenvector $\mathbf{u} = \mathbf{X}\mathbf{v}$ and hence, the eigenvalue γ^2 must also apply for \mathbf{XX}^t . This shows that the two matrices \mathbf{U} and \mathbf{V} share the same eigenvalues.

Whichever is the least of n or p (regarding the original matrix \mathbf{X}) will be the number of diagonal values in the $\mathbf{\Lambda}\mathbf{\Lambda}^t$ and $\mathbf{\Lambda}^t\mathbf{\Lambda}$ matrices, with the larger of these matrices having zeros along the remainder of the diagonal. Hence, only the first $\text{MIN}(n,p)$ columns of the \mathbf{U} matrix are needed in the SVD to form the \mathbf{X} matrix, and similarly only the first $\text{MIN}(n,p)$ rows of matrix \mathbf{V}^t affect the product. \mathbf{U} is referred to as the matrix of row-singular vectors or as the matrix of left singular vectors. \mathbf{V} is referred to as the matrix of column-singular vectors or as the matrix of right singular vectors. Hence, the rows of \mathbf{V}^t are the right singular vectors.

Therefore, taking the SVD of \mathbf{X} as $\mathbf{X} = \mathbf{U}\mathbf{\Lambda}\mathbf{V}^t$, it can be said that the eigenvectors of \mathbf{U} and \mathbf{V} are identical to the eigenvectors of $\mathbf{X}\mathbf{X}^t$ and $\mathbf{X}^t\mathbf{X}$, respectively, and that the singular values in the diagonal matrix $\mathbf{\Lambda}$ are equal to the positive square roots of the corresponding eigenvalues in $\mathbf{\Lambda}^2$.

Most importantly, as outlined previously, if \mathbf{X} is singular then \mathbf{U} , $\mathbf{\Lambda}$, and \mathbf{V} will be of a lower dimension than \mathbf{X} with their dimensions equal to the rank, r , of \mathbf{X} . If \mathbf{X} is $(n \times p)$ then \mathbf{U} will be $(n \times r)$, $\mathbf{\Lambda}$ will be $(r \times r)$ and \mathbf{V} will be $(p \times r)$. This means that we have decomposed \mathbf{X} into matrices whose columns are orthonormal and whose dimensions are reduced - the matrix has been simplified.

Here is a summary of the steps needed for SVD for an $(n \times p)$ matrix \mathbf{B} . In order to achieve the decomposition more quickly, the SVD should be carried out for whichever of Equations (6.28) or (6.29) give the smaller dimensions for $\mathbf{\Lambda}^2$ (the diagonal matrix). If $n \geq p$ then (Massart et al., 1998b):

- 1) Compute \mathbf{B}^t and $\mathbf{B}^t\mathbf{B}$.
- 2) Determine the eigenvalues of $\mathbf{B}^t\mathbf{B}$ and sort these in descending order. Take the square root of these to get the singular values for \mathbf{B} .
- 3) Construct the diagonal matrix $\mathbf{\Lambda}$ by putting the singular values in descending order along the diagonal. Find the inverse, $\mathbf{\Lambda}^{-1}$. The inverse of the diagonal matrix is found by taking the reciprocal of each term along the diagonal.
- 4) Use the ordered eigenvalues from (2) to compute the eigenvectors of $\mathbf{B}^t\mathbf{B}$. Put these eigenvectors along the columns of \mathbf{V} and then get its transpose, \mathbf{V}^t .
- 5) Find \mathbf{U} from $\mathbf{U} = \mathbf{B}\mathbf{V}\mathbf{\Lambda}^{-1}$. The result can be checked by finding if $\mathbf{B} = \mathbf{U}\mathbf{S}\mathbf{V}^t$.

In the case of NIR spectroscopy, it is usually the case that the number of columns, k , corresponding to the number of wavelengths, are significantly greater than, n , the number of rows, corresponding to the number of samples in the calibration set. Hence the bottom $k - n$ rows of $\mathbf{\Lambda}$ will all be zero and can

therefore be removed. Correspondingly, only the first n columns of \mathbf{U} and \mathbf{V} need to be retained. It therefore makes more sense to find the eigenvalues of \mathbf{BB}^t since it will be a smaller matrix than $\mathbf{B}^t\mathbf{B}$. Hence, $\mathbf{BB}^t = \mathbf{U}\mathbf{\Lambda}^2\mathbf{U}^t$ will be solved and then \mathbf{V} computed.

SVD Considered as a Projection:

The product of a matrix with a vector can be considered to be an operation where a pattern of points is projected on an axis (Massart et al., 1998b). The product of two matrices can be considered to be the projection of a pattern of points on a set of axes. A projection is called a rotation when the projection matrix is non-singular and square. Orthogonal rotations occur when a matrix is multiplied with an orthonormal rotation matrix. Such rotations produce a new set of reference axes which are defined by the columns of \mathbf{U} and \mathbf{V} . The structural properties of the pattern of points is retained.

Hence, the SVD of \mathbf{X} can be considered to decompose the matrix into three simpler transformations: a rotation, \mathbf{V}^t , a scaling along the coordinate axes, $\mathbf{\Lambda}$, and a second rotation, \mathbf{U} .

If the dimensions of \mathbf{U} and \mathbf{V} are less than those of \mathbf{X} , which will be the case where \mathbf{X} is singular, then the r (rank number) columns of \mathbf{U} can be considered to be a basis of dimension subspace S^r of the original row subspace (S^n). Similarly the r columns of \mathbf{V} will be a basis of the dimension subspace S^r of the original columns subspace (S^p). S^r is known as the **factor space** (i.e. it is the space within S^n and S^p that is occupied by the orthogonal vectors).

Hence, \mathbf{U} and \mathbf{V} are projection matrices (each containing r projection vectors) and the symbols \mathbf{S} and \mathbf{L} are used to represent their images in the dual space. The projection of rows of a data matrix are called **scores** and the projection of columns are called **loadings**. In chemometrics the symbol \mathbf{T} is typically used for the scores matrix and the symbol \mathbf{P} for the loadings matrix.

$\mathbf{S} = \mathbf{T} = \mathbf{XV}$ for a projection in S^p producing an image in S^n , the image \mathbf{S} will be a $n \times r$ matrix

$\mathbf{L} = \mathbf{P} = \mathbf{X}^t\mathbf{U}$ for projection in S^n producing an image in S^p , the image \mathbf{L} will be a $p \times r$ matrix.

The scores can be considered to represent the coordinates for the n rows of \mathbf{X} in the factor space. Using the SVD it can be written in terms of \mathbf{U} and $\mathbf{\Lambda}$ (Massart et al., 1998b):

$$\mathbf{T} = \mathbf{U}\mathbf{\Lambda}^\alpha \tag{6.35}$$

The factor scaling coefficient α is usually 0, 0.5 or 1. Similarly, the loadings can be considered to represent the coordinates for the p columns of \mathbf{X} in the factor space and can be defined in terms of \mathbf{V} (Massart et al., 1998b):

$$\mathbf{P} = \mathbf{V}\Lambda^\beta \quad (6.36)$$

The factor scaling coefficient β is usually 0, 0.5 or 1.

6.4.2 Principal Component Algorithms

According to the original definition provided for PCs it can be considered that each column of the scores matrix \mathbf{T} represents a row principal component of \mathbf{X} . Where $\alpha = 1$, it can be seen that (Massart et al., 1998b):

$$\mathbf{T} = \mathbf{U}\Lambda \quad (6.37)$$

Hence, since $\mathbf{X} = \mathbf{U}\Lambda\mathbf{V}^t$, then multiplying by \mathbf{V} gives:

$$\mathbf{T} = \mathbf{X}\mathbf{V} \quad (6.38)$$

Therefore each column of \mathbf{T} is a linear combination of the columns of \mathbf{X} using the elements of \mathbf{V} as a weighting coefficient.

Similarly, where $\beta = 1$

$$\mathbf{P} = \mathbf{V}\Lambda = \mathbf{X}^t\mathbf{U} \quad (6.39)$$

Therefore, each column of \mathbf{P} represents a column principal component of \mathbf{X} and can be regarded as a linear combination of the rows of \mathbf{X} using the elements of \mathbf{U} as weighting coefficients.

Where $\beta = 0$ the columns of \mathbf{V} are the column principal components of \mathbf{X} .

In PCA typically $\alpha = 1$ and $\beta = 0$, hence the SVD of \mathbf{X} can be written as:

$$\mathbf{X} = \mathbf{T}\mathbf{P}^t \quad (6.40)$$

This model links loadings and scores to the original matrix and is said to be a **bilinear model** since it is linear in both loadings and scores.

The first principal component of \mathbf{X} is considered to be represented by the first column/eigenvector of \mathbf{P} (its loading) and the first column/eigenvector of \mathbf{T} (its scores). Since the eigenvectors are all of unit length and orthogonal to each other they can be considered to represent axes in multidimensional space with the direction of the first PC being orientated so that it covers the direction of greatest variance in the data points, and the direction of the second PC representing the direction of the greatest residual variance in the data. There will be as many axes as principal components (A) that have been chosen to decompose the original matrix. The effect that the rotation of axes has on how the datapoints are fitted is illustrated in Figure 6-2. There the black axes represent the original axes for three variables and the red lines indicate the two new axes provided by PCA. Given that PCs express variations in the data that are uncorrelated, each PC contains a representation of those variations in the data that are correlated with each other (Mark, 2001a).

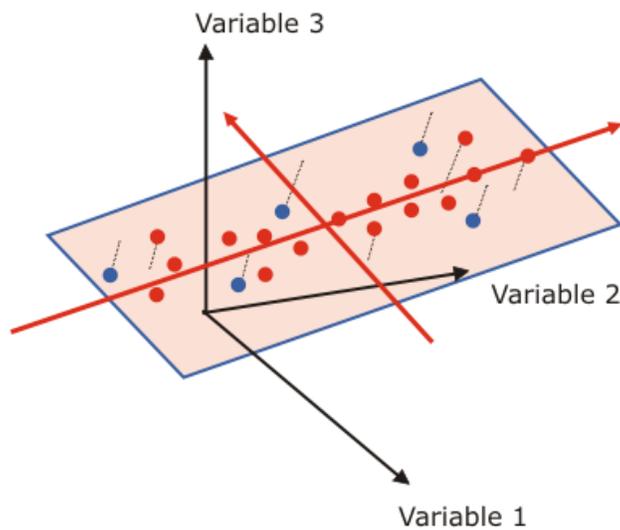


Figure 6-2: The new PCA loading vectors. In creating new, orthogonal, axes that are designed to cover the direction of greatest variance in the datapoints, PCs provide a simpler and easier to interpret model. The red axes are for PCs 1 and 2 and the black axes for the manifest variables. Taken from (CAMO, 2011).

Furthermore, since all manifest variables (wavelengths) are included in every PC there is no need for a wavelength search to pick wavelengths, as occurs in MLR. This is important because there are many effects that have influence on NIR spectra but have no specific absorbance bands (particle size being a clear example), and in the MLR situation some wavelength or some numbers of wavelengths would need to be used to correct for this effect whereas in PCA the effect upon the whole spectrum is considered (Mark, 2001a).

Since each PC is designed to minimise the residual sum of squares it becomes apparent that significantly fewer components are needed to satisfactorily model the data than original X variables. For example, NIR spectra taken between 1100-2500nm with datapoints every 0.5 nm (as is the case with the FOSS XDS unit) provide a total of 2800 variables in a non-decomposed X matrix; however, it is often the case that, for example, 99% or more of the variance seen across the spectra of all samples can be accounted for with less than 10 PCs. Hence, while a perfect SVD of X is possible if all the eigenvalues and eigenvectors are solved for, a good reconstruction of X can be accomplished with a lower number of eigenvectors than r, the rank of X.

6.4.2.1 NIPALS Algorithm

Most modern software use algorithms that target the determination of the greatest eigenvalue first. The NIPALS algorithm, which is used by The Unscrambler X software, produces row and column latent vectors one factor at a time from an $n \times p$ singular matrix \mathbf{X} with the first having the largest associated eigenvalue, the second the next largest, etc. The algorithm works (CAMO, 2010) by first centring the \mathbf{X} variables (mean-centring). A total of n random values are then selected for the scores vector $\hat{\mathbf{t}}_1$. Then:

1. The loading vector $\hat{\mathbf{p}}_1$ is then estimated for factor 1 by projecting the matrix \mathbf{X} on $\hat{\mathbf{t}}_1$, i.e.

$$\hat{\mathbf{p}}_1^t = (\hat{\mathbf{t}}_1^t \hat{\mathbf{t}}_1)^{-1} \hat{\mathbf{t}}_1^t \mathbf{X} \quad (6.41)$$

2. The length of $\hat{\mathbf{p}}_1$ is scaled to 1.0.

$$\hat{\mathbf{p}}_1 = \hat{\mathbf{p}}_1 (\hat{\mathbf{p}}_1^t \hat{\mathbf{p}}_1)^{-0.5} \quad (6.42)$$

3. The estimate of $\hat{\mathbf{t}}_1$ is then improved for this factor by projecting the matrix \mathbf{X} on $\hat{\mathbf{p}}_1$.

$$\hat{\mathbf{t}}_1 = \mathbf{X} \hat{\mathbf{p}}_1 (\hat{\mathbf{p}}_1^t \hat{\mathbf{p}}_1)^{-1} \quad (6.43)$$

4. See if the estimate of the eigenvalue $\hat{\tau}_1$ improves:

$$\hat{\tau}_1 = \hat{\mathbf{t}}_1^t \hat{\mathbf{t}}_1 \quad (6.44)$$

It is then noted whether the estimated eigenvalue is converging – if $\hat{\tau}_1$ minus the $\hat{\tau}_1$ in the previous iteration is less than a pre-specified constant then the method has converged for this factor. If not then the algorithm returns to step (1) with the new estimate for the scores vector and repeats.

Once a factor has been computed the residual spectrum is constructed as follows:

$$\mathbf{X}_a = \mathbf{X}_{a-1} - \hat{\mathbf{t}}_a \hat{\mathbf{p}}_a^t \quad (6.45)$$

And the process above is repeated, i.e. randomly choose the scores vector $\hat{\mathbf{t}}_a$ and then use this on the residual matrix \mathbf{X}_a .

6.4.3 Principal Components Analysis

There are many tools available in order to evaluate and design effectively a PC model that will represent a data-set of spectra.

6.4.3.1 Loading Plots

In NIRS it is common to present the loadings as a line with the wavelength variables on the x-axis and the loading values on the y-axis. Hence, each of the A vectors of \mathbf{V} can be plotted as if they were spectra. Every variable that is analysed will have a loading on each PC and this loading will reflect how much that PC considers the variation contained in that variable. A large loading (close to +1 or -1) will indicate a strong relationship.

It is possible for a loading plot to represent the spectrum of a pure chemical that exists in a mixture within a sample (e.g. water in a lignocellulosic sample); however, it is unusual for this to be entirely the case since this would require that chemical to vary in a manner that is entirely independent of the other constituents due to the orthogonality condition of PCs (Mark, 2001a).

6.4.3.2 *Score Plots*

The score for each PC for each sample can be found by projecting the point for the sample on the axis that this PC represents. It is the point on this axis nearest the point representing the sample in the hyperspace, i.e. it is the coordinate of the sample on that axis (Naes et al., 2007). Scores can be positive or negative and represent the distance from the mean. Samples that have similar scores on a particular PC are similar (at least in terms of the important variables for the associated loading of that component).

Typically PC scores are plotted in two dimensions with each axis representing a PC. There are some general points regarding the relationship between the PC score for a sample and the associated loading plot for that PC (CAMO, 2011):

- A variable with a positive loading means that samples with positive scores have higher than average values for that variable whereas all samples with negative scores have lower than average values for that variable. This relationship increases with the absolute value of the score (i.e. higher scores mean a higher variable value) and with the size of the loading.
- Conversely, where a variable has a negative loading, samples with positive scores have lower than average values for that variable and samples with negative scores have higher than average values for that variable. This relationship increases with the absolute value of the score and with the size of the loading.
- Variables with low loadings are not suitable for interpretation by the associated PC.

6.4.3.3 *Distance Measures in NIRS*

Before discussing the other analytical tools available in PCA, it is important to outline the methods that exist for measuring distance, both in the PC factor space, where scores will be used, and in the original space, where the manifest variables will be used. The **Euclidean Distance** between two objects (such as spectra) x and y with K coordinates (wavelengths or PC scores) is calculated as (FOSS, 2006a):

$$D_E = \sqrt{\sum_{i=1}^K (x_i - y_i)^2} \quad (6.46)$$

Where PC scores are used, the distance is sometimes calculated from a limited number of significant components (i.e. $K < r$). Where the data has been mean centred prior to the calculation of the PCs, the D_E of a sample, \mathbf{x}_i , from the centroid of the group is determined as simply (De Maesschalck et al., 2000):

$$D_e^t = \sqrt{\sum_{j=1}^K t_{ij}^2} = \sqrt{\mathbf{t}_i \mathbf{t}_i^t} \quad (6.47)$$

Figure 6-3 (a) and Figure 6-3 (b) show, in two and three dimensions (variables) respectively, how similar types of samples tend to group together in factor space. In Figure 6-3(a) the means of each group at each wavelength are indicated as vectors and where they cross is the central location of the group. The list of numbers describing the mean absorbance values for a given material (e.g. soft wheat) over all the variables is known as the **group mean** of that group and the set of vectors describing all the materials to be distinguished is called the **group mean matrix**. It can be seen that simply measuring the Euclidean distance of a spectrum from the centre of a group will not be the best test for whether that sample belongs to that group or not; due to the elliptical nature of the groups a sample orientated perpendicularly to the group would need to have a much lower D_E from the group centre to be considered outside the group compared with a sample that lies along the axis of the group.

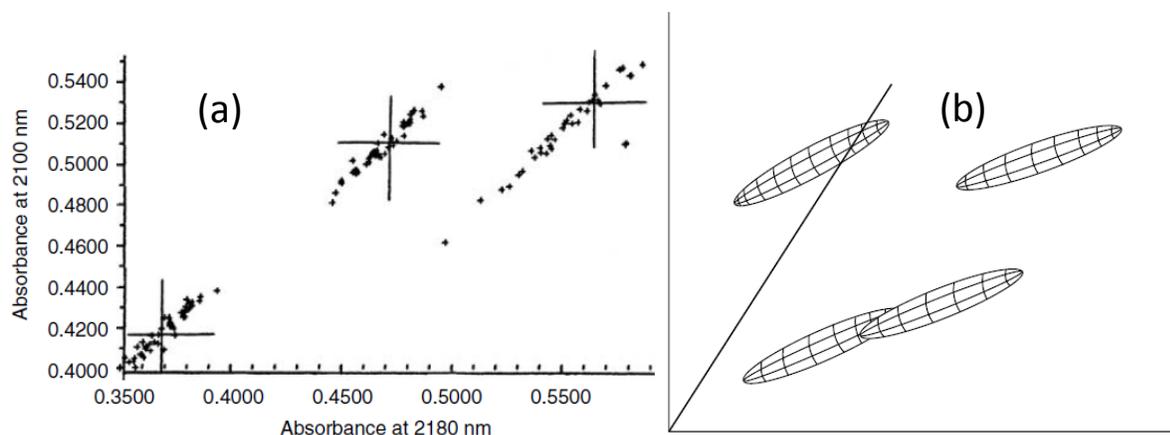


Figure 6-3: Absorbances of groups of materials in multidimensional space. (a) Plots of the absorbances of three materials (soft wheat, hard wheat and soy, from left to right) at two wavelengths. The samples tend to form elliptical groups in the 2D space. (b) In three- and higher dimensional space the groups are ellipsoidal in shape, as illustrated for four groups here. Figures taken from (Mark, 2001b).

To account for this another distance measure, known as the **Mahalanobis Distance** (MD), can be used. This technique allows the equivalent Euclidean distance to be large in those directions where the group is elongated. The MD can be considered to be an ellipse/ellipsoid that surrounds the centre of the group and spreads outwards with increasing distance. A MD of 1 typically represents one standard deviation of the data in the group. The MD required for this ellipse to stretch enough to meet the spectrum of a particular sample (sample A in Equation (6.48)) would be the MD for that sample relative to that group mean. This distance can be expressed as (De Maesschalck et al., 2000):

$$MD = \sqrt{(A - \mu)M(A - \mu)^t} \quad (6.48)$$

Where A is a multidimensional vector describing the location of sample A and μ is a vector of the same dimension describing the group mean. The spectrum A may or may not belong to this training set. M is a matrix that contains the distance measures used for computing the MD; it contains the information necessary to construct the ellipse. Mark (2001b) describes three forms, M1, M2 and M3 and these are illustrated in Figure 6-4.

M1 is simply a unit matrix meaning that that the ellipsoid becomes a sphere and the distance measure simply becomes the Euclidean distance. M2 uses the inverse of the variance-covariance matrix of the samples in the group (see Section 6.4.1 for the variance-covariance matrix). The M2 matrix will result in an ellipsoid being fitted to the data for each group. The M3 matrix applies the same ellipsoid (in terms of size, shape and orientation) to all of the groups in the calibration and it is represented by the inverse of

the matrix formed by pooling the within-group covariance matrices of all the groups meaning that there is only one inverse variance–covariance matrix for the whole factor space.

It should be noted that, while Euclidean space allows distance measures to be comparable for all directions, and the M3 matrix will allow distances measured in opposite directions (i.e. $D_{ij} = D_{ji}$) to be the same (with such spaces known as **metric spaces**), the use of M2 and a normalised M3 matrix (where there is more than one group) will mean that there will be different distance measures for each group. These all called **non-metric spaces**, meaning that there is no unique specification of the distance measure, even in a given direction (Mark, 2001b).

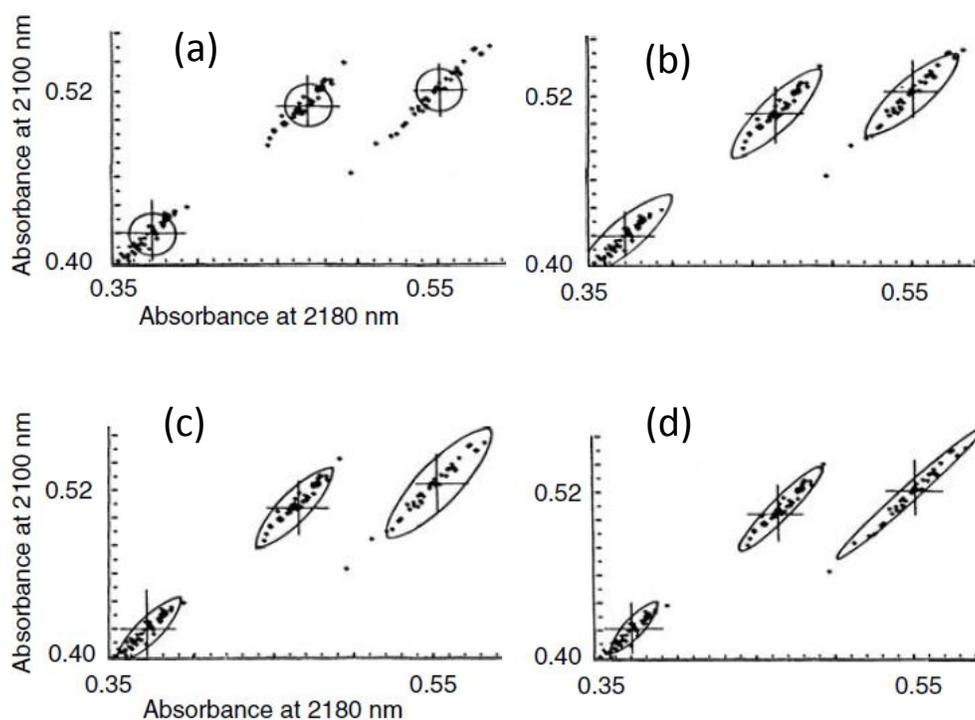


Figure 6-4: Illustrations of how the bounding ellipsoids can change in size, shape and orientation according to what matrices are used to calculate the MD. (a) The M1 matrix means that the space is Euclidean, (b) All groups are ellipsoids with the same size, shape, and orientation (M3 matrix), (c) All groups are ellipsoid with the same shape and orientation and the size is fitted to the root mean square group size (normalised MD from M3) (d) The M2 matrix is used, meaning that ellipsoids are fitted to each group, meaning that each has its own shape, size and orientation. Taken from Mark (2001b)

Where all of the spectra are only assigned to one group, clearly M2 and M3 will be the same and the use of the MD will be to detect outliers that lie far from the mean of the distribution. The methods involved in forming separate groups are discussed in the Section on qualitative analysis (See Section 7.1).

If the MD is computed based on the location of samples in a PC scores space, rather than the manifest variables space, the calculation is much simpler than in Equation (6.48). Since X has been mean-centred prior to PCA analysis the centroid of the group will be zero and Equation (6.48) will reduce to:

$$MD_i = \sqrt{\mathbf{t}_i \mathbf{C}_t^{-1} \mathbf{t}_i^t} \quad (6.49)$$

Where \mathbf{C}_t^{-1} is the inverse of the variance-covariance matrix \mathbf{C}_t which is equal to $\mathbf{T}^t \mathbf{T} / (n - 1)$.

The square of the MD is known as the T^2 statistic and can be calculated for sample i as described in the equation below (Deluzio et al., 1997):

$$T_i^2 = \mathbf{t}_i \mathbf{D}_t^{-1} \mathbf{t}_i^t \quad (6.50)$$

Where \mathbf{D}_t^{-1} is a diagonal matrix containing the variances of each scores vector. Equation (6.51) presents Equation (6.50) in an easier to understand form (Kourti and MacGregor, 1995):

$$T_i^2 = \sum_{a=1}^A \frac{t_{ia}^2}{s_{ti}^2} \quad (6.51)$$

Where s_{ti}^2 is the estimated variance of \mathbf{t}_i . A Hotelling's T^2 test, which is a multivariate form of a t -test can then be made to determine whether the sample lies within user-defined critical limit distances. Equation (6.52) shows the formula for calculating T_{crit}^2 for the training set (Candolfi et al., 1999):

$$T_{crit}^2 = \frac{(n - 1)^2}{n} B(\alpha, A/2, (n - A - 1)/2) \quad (6.52)$$

In Equation (6.52) B refers to the tabulated value for the confidence level, α , and for $A/2$ and $(n - A - 1)/2$ degrees of freedom (A is the number of PCs used). This value is found in the F distribution which is used in the situation where the training set samples are used to obtain the mean and variance-covariance matrix needed to calculate the MD's.

For the prediction of the T_i^2 value for a sample outside the calibration set, the mean and the variance-covariance matrix of the calibration set are used. For the calculation of T_{crit}^2 for these samples the F -distribution is used as shown in Equation (6.53) (Candolfi et al., 1999):

$$\begin{aligned}
T_{crit}^2 &= \frac{A(n-1)(n+1)}{n(n-A)} F(\alpha, A, (n-A)) \\
&= \frac{A(n^2-1)}{n(n-A)} F(\alpha, A, (n-A))
\end{aligned}
\tag{6.53}$$

F will be the tabulated value for the confidence level α and A and $(n - A)$ degrees of freedom.

If the T_i^2 value is greater than T_{crit}^2 then the sample is considered to be outside the critical limit. It follows that T_i^2 tests can be computed after each PC for each sample with the T_i^2 statistic and T_{crit}^2 increasing after each PC. This is an important consideration, since each PC score is scaled by the reciprocal of its variance (Equation (6.51)) each PC plays an equal role in the determination of T_i^2 irrespective of the amount of variance of the X matrix that each PC explains for. Hence, if too many PCs are included the effects of noise could become more influential to the calculation of the T_i^2 statistic.

6.4.3.4 Residuals and Variance

Residuals can be defined for samples and variables. For samples these can be considered to be the distance between the original location of the point and its projection onto the hyperplane generated by the model components, i.e. is considered to be the part of the location of the sample that is not modelled by the model components. The \mathbf{x} -residual vector $\hat{\mathbf{e}}_i$ for sample i is defined by (Naes et al., 2007)::

$$\hat{\mathbf{e}}_i^t = \mathbf{x}_i^t - \hat{\mathbf{x}}_i^t = \mathbf{x}_i^t - \hat{\mathbf{t}}_i^t \hat{\mathbf{P}}^t
\tag{6.54}$$

Equation (6.54) calculates $\hat{\mathbf{e}}_i^t$ as the position of the vector in real space minus the position of the vector in model-projected space. Large residual vectors mean that the sample is not well represented by the model and may contain other unexpected constituents compared to the rest of the samples.

The **studentised residual** is a residual that is adjusted by dividing it by an estimate of its standard deviation (its formulation, with a correction for the leverage of the sample, is described in Section 6.4.3.5). This is done because regressions yield different residual distributions at different data points, i.e. residuals do not have the same variance, with the variance decreasing as one progresses further from the average x value (a result of the relative influence that a sample has increasing with distance from the mean).

The **residual sample variance** is classified as the mean squared residual corrected for degrees of freedom. Correspondingly, the **explained variance** for a sample can be calculated from its residual variance as the residual variance given from a model with a certain number of factors divided by the residual variance when no factors are used.

The **total residual variance** is a representation of what the model cannot accurately represent, i.e. it is the variance of the error part of the data. It can be calculated as the sum of the residual variances for all samples. It can be computed for differing number of principal components and the data plotted. Similarly, **total explained variances** measure how much of the original variation in the data is described by the model. It can also be calculated for different numbers of PCs. An ideal model should have high (close to 100%) values for the total explained variance with as little PCs as possible although the number of PCs needed for this will increase as the heterogeneity and/or complexity of the samples does.

6.4.3.5 Leverage Plots

The leverage of a sample measures the distance from the **model-projected** sample to the model centre. A high leverage indicates that a sample has a higher influence on the model than other samples with lower leverages. The sample leverage, h_i , is usually determined, in PCA and PLS, from the scores that are used in the regression equation meaning that the leverage is a distance measure in the space covered by the A components used in the factorisation of \mathbf{X} . It can be expressed, for samples, in terms of the normalised-score vectors of the sample, $\hat{\mathbf{u}}_i$ (Equation (6.55) (Boysworth and Booksh, 2001)), or in terms of the $\hat{\mathbf{t}}_i$ scores of the sample and the variance of component vector $\hat{\mathbf{t}}_a$ (Equation (6.56), the formula that The Unscrambler X uses (CAMO, 2010), which is also equivalent to Equation (6.57)(Naes et al., 2007)).

$$h_i = 1/N + \hat{\mathbf{u}}_i^t \hat{\mathbf{u}}_i \quad (6.55)$$

$$h_i = 1/N + \sum_{a=1}^A \hat{t}_{ia}^2 / \hat{t}_a^t \hat{t}_a \quad (6.56)$$

$$h_i = \frac{1}{N} + \sum_{a=1}^A \hat{t}_{ia}^2 / \hat{\lambda}_a \quad (6.57)$$

In equations (6.56) and (6.57), \hat{t}_{ia} is the score along component a for sample i and the denominator is the sum of squared score values for the calibration samples corresponding to component a . It is equivalent to the a th eigenvalue of the matrix $\mathbf{X}^t\mathbf{X}$ (see Section 6.4.1).

It can be seen that the leverage is closely related to the Mahalanobis Distance. Points with equal leverage values are positioned on ellipsoids centred at the mean of the group/product and the h_i constant is the equation of an ellipse, whose shape is defined, via $\mathbf{X}^t\mathbf{X}$, by the configuration of the calibration spectra (Boysworth and Booksh, 2001).

The leverage for calibration samples can be between $1/N$ (which can be close to zero for a very large calibration set) and 1 with the average leverage for all samples being $(1+A)/N$ (Naes et al., 2007). For samples that are not in the calibration set h_i can be any value greater than $1/N$. The equation above also shows that each component provides its own contribution to the leverage of a sample; this can be of use to determine whether a sample is in fact primarily responsible for that component (which could be the case where the contribution was very high).

The Unscrambler X allows an **influence plot** to be used in PCA: this plots the leverage of a sample against its residual X-variance for the total number of components chosen for the analysis. An alternative leverage plot is one based on the Hotelling T^2 statistic as described in Section 6.4.3.3.

The **studentised leverage corrected residual** is used to correct for the weight each sample has in determining the calibration model meaning that the studentised residual is increased for samples with a high leverage. It is calculated by (Boysworth and Booksh, 2001):

$$r_{ij} = \frac{e_{ij}}{\hat{\sigma}_j \sqrt{1 - h_i}} \quad (6.58)$$

Where r_{ij} is the studentised leverage corrected residual of sample i at wavelength j , e_{ij} is the non-studentised residual of that sample at that wavelength, h_i is the leverage of sample i , and $\hat{\sigma}_j$ is the standard deviation of the residuals at wavelength j as described in equation (6.59):

$$\hat{\sigma}_j = \sqrt{\frac{\sum_{i=1}^N e_{aj}^2}{N - A - 1}} \quad (6.59)$$

6.4.3.6 Identification of X-Outliers

Outliers are considered to be samples that appear highly different from the others. The result of this may be that the sample has a large residual on the model or that the model itself is highly influenced by that sample (high leverage).

There are several types of outliers; spectral outliers are evaluated purely on how their spectral properties differ from the PC model. There can be some instances where they differ so greatly from the other samples in the data set that one or more PCs will be required simply to include that variation into the model; this may often be to the detriment of a robust model that is relevant to the vast majority of samples. This is illustrated in Figure 6-5, where the axis of PC1 changes significantly after the removal of outlying sample number "7". For reasons such as this it is important that the PCA model is recalculated following the removal of any outliers.

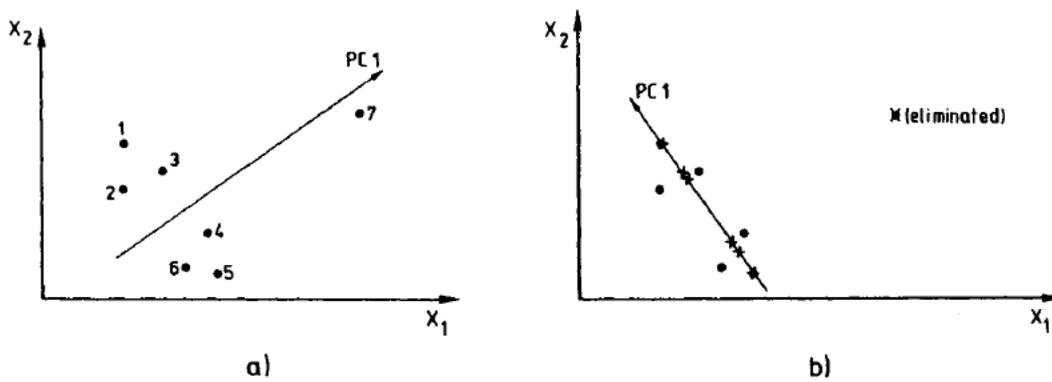


Figure 6-5: Effects of an outlier on a PCA loading vector. (a) PC1 for a 2D data set containing an outlier (point 7), (b) PC1 for the same data set after elimination of point 7. Taken from (Massart et al., 1998a).

There are several ways to manually find and assess outliers:

- 1) Leverage plots (Section 6.4.3.5) can be used – there are no strict rules about when a sample should be considered to be a leverage outlier but a possible warning may be given when the leverage is two or three times larger than the average of $(1+A)/N$ (Naes et al., 2007). Alternatively a Hotelling T^2 plot with its critical limit can be used and samples lying outside the limit can be considered as potential outliers.
- 2) The residuals can be examined, the cut off for the studentised residual (or the studentised leverage corrected residual) is generally considered to be 2 or 3, since the probability for a

residual to have such large values would be 2.3% or 0.13%, respectively, assuming a normal distribution (Massart et al., 1998a). The use of the Q-residuals and its associated critical limit can also be applied.

- 3) An influence plot can be used. Samples with a high leverage and a high residual are most probably outliers, as illustrated in Figure 6-6. An alternative to a standard influence plot would be a Hotelling T^2 and a Q-residuals influence plot that includes the critical limits for both statistics. Samples lying beyond the critical limits for both statistics are strong candidates as outliers. The use of an influence plot is important since it looks at both residuals and leverage and these are both important given that they contain information about different subspaces (the model projected space and the manifest variable space).
- 4) One dimensional or two dimensional scores plots can indicate where a sample is not similar to other samples in the data set.

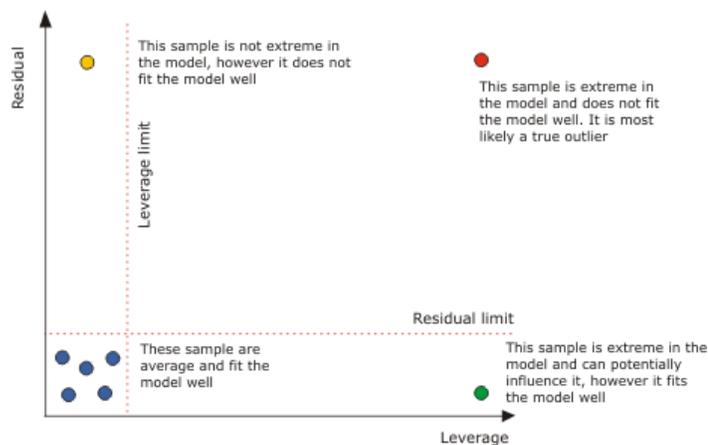


Figure 6-6: An influence plot with theoretical examples. It shows samples with low influence/leverage that fit the model well, i.e. low residuals (in blue); a sample with a high influence that fits the model well (green), a sample with a low influence that does not fit the model well (yellow), and an “influential outlier” which has a large influence on the model but does not fit it well (red). Taken from (CAMO, 2011)

6.4.4 Principal Component Regression

The relationship $\mathbf{X} = \hat{\mathbf{T}}\hat{\mathbf{P}}^t + \mathbf{E}$ (in the instance where $a < r$) results in the matrix \mathbf{X} being decomposed into orthogonal components and we can consider the scores matrix to represent the co-ordinates of the samples in the new factor space. Hence, regression for a vector \mathbf{y} where each entry corresponds to a

property (e.g. cellulose content) of a sample whose corresponding spectrum is recorded in \mathbf{X} (and therefore its scores in $\hat{\mathbf{T}}$) can be done via conventional MLR as outlined in Equations (6.60) and (6.61):

$$\mathbf{y} = \hat{\mathbf{T}}\mathbf{q} + \mathbf{f} \quad (6.60)$$

Or,

$$\hat{\mathbf{y}} = \hat{\mathbf{T}}\mathbf{q} \quad (6.61)$$

The regression coefficients, stored in \mathbf{q} , can then be estimated by least squares with the residuals stored in \mathbf{f} .

As mentioned in Section 6.3, MLR can be considered to be a matrix/projection problem. The n rows of \mathbf{T} will form a pattern \mathbf{P}^n of points that are projected on the unknown vector (axis) \mathbf{q} . Hence the axis is imaged by \mathbf{T} in the dual space \mathbf{S}^n at the point $\hat{\mathbf{y}}$. Since \mathbf{y} has the same dimension it is also represented in the space \mathbf{S}^n (Massart et al., 1998b). Therefore MLR seeks to define \mathbf{q} such that $\hat{\mathbf{y}}$ is as close as possible to \mathbf{y} (i.e. for the mean distances between the two vectors to be the minimum possible under \mathbf{q}). MLR allows the calculation of \mathbf{q} through matrix operations as outlined in Equation (6.62) (Massart et al., 1998b).

$$\mathbf{q} = (\mathbf{T}^t\mathbf{T})^{-1} \mathbf{T}^t\mathbf{y} \quad (6.62)$$

Hence, including this term in Equation (6.61) leads to:

$$\hat{\mathbf{y}} = \mathbf{T}(\mathbf{T}^t\mathbf{T})^{-1} \mathbf{T}^t\mathbf{y} \quad (6.63)$$

The term $\mathbf{T}(\mathbf{T}^t\mathbf{T})^{-1} \mathbf{T}^t$ is known as the **orthogonal projection operator** (Massart et al., 1998b).

The result of Equation (6.62) allows predictions of the compositions of unknown samples to be made based on their PC scores, i.e.:

$$\hat{y}_i = q_0 + q_1t_1 + q_2t_2 + \dots \quad (6.64)$$

In order to be used for the prediction of unknown samples that were not part of the calibration set and whose scores are not known, Equation (6.64) will require knowledge of not only the calibration coefficients but also the loading vectors (\mathbf{P}) since the PC scores for the unknown samples will need to be calculated via Equation (6.38).

However, it is possible to construct a calibration equation that only requires calibration coefficients and the spectral data of the unknown samples. This is because the PC scores themselves are computed from the product of the optical data and eigenvectors of \mathbf{P} :

$$t_i = p_{1i}x_1 + p_{2i}x_2 + \dots \quad (6.65)$$

Hence, Equation (6.64) can be written as (Mark, 2001a):

$$\begin{aligned} \hat{y}_i = & q_0 + (q_1p_{11} + q_2p_{21} + q_3p_{31} + \dots)x_1 \\ & + (q_1p_{12} + q_2p_{22} + q_3p_{32} + \dots)x_2 + \dots \end{aligned} \quad (6.66)$$

Since all of the items within the brackets are constants they can be replaced by individual constants and the equation becomes:

$$\hat{y}_i = b_0 + b_1x_1 + b_2x_2 + \dots \quad (6.67)$$

This means that it is not necessary to retain the values of the PCs once the regression coefficients have been determined. Furthermore, there will be a coefficient for each wavelength meaning that the calibration can be plotted as a spectrum or treated in any way in which a spectrum can. The regression coefficients will show how each wavelength is weighted when predicting a response in \mathbf{y} .

The main difference between PCR and MLR of the original \mathbf{X} matrix is that the scores matrix is orthogonal and, hence, there will be none of the collinearity problems that frequently occur with MLR of spectra whose absorbances at different wavelengths tend to be highly correlated. Therefore, the estimation of the regression coefficients under the PCR method is much more stable and allows for the better prediction of the composition of unknown samples.

It should be noted that it is still possible to overfit a calibration with PCR by including too many components in the model; however, there are statistical techniques to assess the importance of the components and which should be retained. These and other important statistics are outlined in Section 6.8.

6.5 Validations of Models

All models should be validated in order to ensure that the model is not overfitted to the samples that were used to derive the model and that it can be applicable to unknown samples outside of the calibration set.

There are several ways to validate a model with the most common ones being test set validation and cross validation. With test-set validation a group of samples that were not included in the development of the model are used in the validation set. Cross-validation involves the use of calibration-set samples for the validation process. The method involves removing a sample (full cross validation), or group of samples (segmented cross-validation), from the calibration set after which the model development takes place on the remaining samples. Various statistical techniques can then be applied (see Sections 6.7 and 6.8) to measure how well the model fits the sample(s) that were excluded from the model. Following this the sample(s) are put back into the calibration set and other sample(s) removed from the set and the process repeated. This continues until all samples have, at some stage, been validated.

A more developed form of validation involves the separation of the samples into three separate subsets. The first is used to construct all of the models to be considered (calibration set). The second (fitting set) is used to choose the best fit of the model in terms of accuracy or precision (for example, selecting the most appropriate number of PCs to incorporate in the model), and the third being used as a test-set (Boysworth and Booksh, 2001).

It is very important for the validation process to be effective, and for it to be a good predictor of the future performance of the model on unknown samples. For this to happen the structure of the calibration and test sets (and fitting sets if these are to be used) should be similar. Hence there should be a similar distribution regarding the spectral diversity of the samples and regarding the variation in physical and chemical properties of the samples in each set. Therefore for complex sample sets that cover wide differences in spectral and physiochemical properties the size of the calibration, fitting and test sets will need to be significant. It is for this reason that often the model is fitted via cross-validation methods within the calibration set and, hence, only two sets (calibration and test) are used (Boysworth and Booksh, 2001).

Section 6.8 describes how models are fitted and Section 6.11 details important procedures in the development of models and how the samples should be distributed between calibration and test sets.

6.6 Prediction of Unknown Samples

Section 6.4.4 and Equation (6.67) detail how prediction for the composition (\mathbf{y} vector) of unknown samples can, following the development of regression equation coefficients for the calibration set, be achieved purely from the spectral data of these new samples and the coefficients. This method, which the Unscrambler X software refers to as **short prediction**, will provide values for \mathbf{y} but will not enable many sample/variable outlier diagnostic tools to be applied. **Full prediction** on the other hand employs projection of the spectra onto the PCR/Partial-Least-Squares (PLS, see Section 6.9.3) model components in order to predict \mathbf{y} and also to give a greater range of diagnostic tools.

In PCR $\hat{\mathbf{X}} = \hat{\mathbf{T}}\hat{\mathbf{P}}^t$ and $\hat{\mathbf{y}} = \hat{\mathbf{T}}\mathbf{b}$ are used and in partial least squares regression (PLSR) $\hat{\mathbf{X}} = \hat{\mathbf{T}}\hat{\mathbf{P}}^t$ and $\mathbf{Y} = \hat{\mathbf{T}}\mathbf{B}$ are used. With these methods the scores are used and not \mathbf{X} directly. Since the scores are calculated from \mathbf{X} (the spectra of the new samples) these methods allow values for the leverage and \mathbf{X} -residuals to be determined for the new samples and so outlier detection is possible, i.e. a sample with a high leverage and/or high residual may be a **prediction spectral outlier** and this means the predicted \mathbf{y} values may not be trusted.

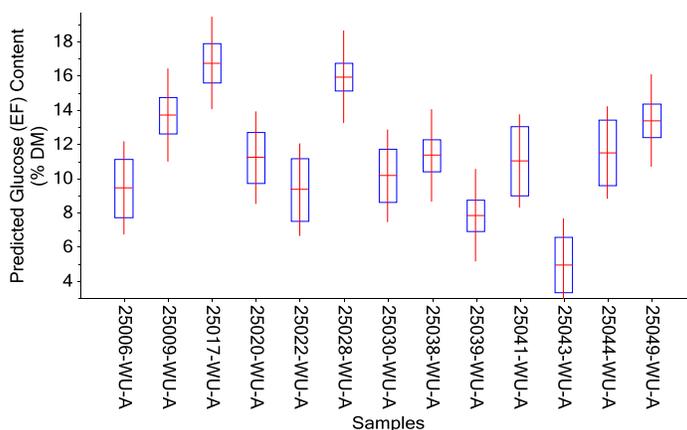


Figure 6-7: A prediction with deviation plot for the extractives-free glucose content of 13 peat samples based on a 2-component PCR calibration model developed on the (untreated) wet spectra of 40 peat samples.

The leverage and X-residual variance data for samples can be used to produce a **prediction with deviation** plot for the estimate of **y** for those samples. The deviation (which is a form of a 95% confidence interval around the predicted y value) expresses how similar the sample to be predicted is to those samples which constitute the calibration set. Figure 6-7 shows a prediction with deviation plot for the extractives-free glucose content of 13 peat samples based on a 2-component PCR calibration model developed on the (untreated) wet spectra of 40 peat samples. The predicted glucose value is a horizontal line and the box around it indicates the deviation.

The deviation for the y-variable *j* in prediction object *i* is calculated from the formula below (De Vries and J.F. Ter Braak, 1995, CAMO, 2010):

$$Dev(i, j) = \sqrt{V_{y, val} \left(\frac{V_{xi, pr}}{V_{xTot, val}} + H_i + \frac{1}{I_{cal}} \right) \left(1 - \frac{A + 1}{I_{cal}} \right)} \quad (6.68)$$

Where: $V_{y, val}$ is the residual variance in the validation set; $V_{xi, pr}$ is the X-residual variance in the prediction sample; $V_{xTot, val}$ is the average x-residual variance in the validation samples; H_i is the leverage of the prediction samples with respect to the *A* principal components/factors; and I_{cal} is the number of calibration samples.

The full prediction process also allows methods to examine whether the samples to be predicted are inliers, based on the **inlier statistic**. An inlier, also known as the nearest neighbour distance inlier, has been defined by the ASTM as “a spectrum residing within a gap in the multivariate calibration space, the result for which is subject to possible interpolation errors” (Workman, 1996). Inliers will not be found by classical methods of outlier detection (Jouan-Rimbaud et al., 1999). The Unscrambler X calculates the inlier statistic by computing the Mahalanobis distance between the prediction sample and all of the samples in the calibration set and takes the minimum MD found as the inlier distance. This is then compared against a critical inlier distance in order to determine whether the sample is an inlier or not.

This critical value is based on the inlier distances of the calibration samples, it is the maximum inlier distance found for these samples. If a prediction sample has an inlier distance greater than the critical value then it is considered an inlier.

6.7 Important Calibration and Regression Statistics

These statistics concern the relationship between the value for y predicted by the model, i.e. \hat{y} , and the actual value for y . They are applicable for various calibration methods, including PCR and PLSR. As mentioned previously there can be multiple subsamples of the original sample set. For example there can be a calibration set, a fitting set, and a test set. For each of these a plot of \hat{y} versus y can be drawn (providing the actual values for y are known) and a regression line of best fit included. The equation of this line can then be written as:

$$\hat{y} = b_0 + b_1 y \quad (6.69)$$

Where b_1 is the slope of the line and b_0 is the offset, or bias.

Some of the most important statistical terms are listed below and some discussions are given regarding their importance and use. Unless otherwise noted the equations are from Workman (2001).

Y-Residual:

$$f_i = y_i - \hat{y}_i \quad (6.70)$$

Where \hat{y}_i is the estimated value for y_i , for sample i , as derived from the calibration equation.

Total sum of squares:

$$SS_{TOT} = \sum_{i=1}^N (y_i - \bar{y})^2 \quad (6.71)$$

Where N is the number of samples in the set.

Sum of squares for residuals:

$$SS_{res} = \sum_{i=1}^N (y_i - \hat{y}_i)^2 \quad (6.72)$$

Sum of squares for regression:

$$SS_{regr} = SS_{TOT} - SS_{res} = \sum_{i=1}^N (\hat{y}_i - \bar{y})^2 \quad (6.73)$$

Mean square for regression:

$$MS_{regr} = \frac{SS_{TOT} - SS_{res}}{d.f. \text{ for regr.}} = \frac{\sum_{i=1}^N (\hat{y}_i - \bar{y})^2}{A + 1} \quad (6.74)$$

i.e. SS_{regr} divided by the degrees of freedom for regression, which is equal to A (the number of factors used in the regression, e.g. PCs, PLS factors, or wavelengths in MLR) plus one.

Mean square for residuals:

$$MS_{res} = \frac{SS_{res}}{d.f. \text{ for residual}} = \frac{\sum_{i=1}^N (y_i - \hat{y}_i)^2}{N - A - 1} \quad (6.75)$$

i.e. SS_{res} divided by the degrees of freedom for the residual, which is equal to the number of samples (N) minus A minus one. This statistic is also known as the estimated error variance, $\hat{\sigma}^2$ (Naes et al., 2007).

Pearson Correlation Coefficient (r):

$$r = \frac{\sum_{i=1}^N (y_i - \bar{y})(\hat{y}_i - \widehat{\bar{y}})}{\sqrt{\sum_{i=1}^N (y_i - \bar{y})^2} \sqrt{\sum_{i=1}^N (\hat{y}_i - \widehat{\bar{y}})^2}} \quad (6.76)$$

This is the covariance of two variables divided by the product of their standard deviations. This statistic shows the degree to which two sets of data (e.g. y and \hat{y}) agree with each other. It can vary between -1 and +1.

t-test for the coefficient of correlation:

$$t = \frac{r}{\sqrt{\left(\frac{1 - r^2}{n - 2} \right)}} \quad (6.77)$$

Note that, since the value of t that can be found in statistical tables is dependent on n , statistically significant correlations can be found at even low r values where n is large.

Coefficient of Multiple Determination or Multiple Correlation Coefficient (r^2):

$$r^2 = 1 - \left(\frac{\sum_{i=1}^N (y_i - \hat{y}_i)^2 / (N - a - 1)}{\sum_{i=1}^N (y_i - \bar{y})^2 / (N - 1)} \right); r^2 = \left(\frac{\sum_{i=1}^N (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^N (y_i - \bar{y})^2} \right) \quad (6.78)$$

The equation above on the left is the r^2 adjusted or proper degrees of freedom; and the equation above on the right is for the unadjusted r^2 (general notation).

This equation shows the proportion of variance in the y data that is attributable to the variance in \hat{y} as a total fraction of 1.0. Hence, the unadjusted r^2 can be simplified to:

$$r^2 = \frac{SS_{regr}}{SS_{TOT}} \quad (6.79)$$

Bias:

$$BIAS = \sum_{i=1}^N (\hat{y}_i - y_i) / N \quad (6.80)$$

This is defined as the average difference between \hat{y}_i and y_i in the subset (e.g. test set).

Root Mean Square Error of Calibration:

$$RMSEC = \sqrt{\sum_{i=1}^N (\hat{y}_i - y_i)^2 / (N - A - 1)} = \sqrt{MS_{res}} \quad (6.81)$$

This statistic only considers the residuals of the samples that were in the calibration set that was used to develop the model. It is an estimate of the best accuracy possible for a specific calibration model. The subtraction of A in the degrees of freedom term is an attempt to correct somewhat for the overfitting that can occur when an excess of factors are used in the model. If there is no error with the NIR instrument and method of calibration the RMSEC would be a measure of the error involved in the reference analysis.

Root Mean Square Error of Cross Validation:

$$RMSECV = \sqrt{\sum_{i=1}^N (\hat{y}_{CV,i} - y_i)^2 / (N)} = \sqrt{\frac{SS_{res}}{N}} \quad (6.82)$$

Where $\hat{y}_{CV,i}$ is the estimate for y_i based on the calibration with sample i deleted.

Root Mean Square Error of Prediction:

$$RMSEP = \sqrt{\sum_{i=1}^{N_p} (\hat{y}_i - y_i)^2 / (N_p)} \quad (6.83)$$

This statistic is based on the test set with N_p being the number of samples in that set.

Standard Error of Prediction:

$$SEP = \sqrt{\sum_{i=1}^{N_p} (\hat{y}_i - y_i - BIAS)^2 / (N_p - 1)} \quad (6.84)$$

This is defined as the standard deviation of the predicted residuals. Since SEP is defined as standard deviation $N_p - 1$ is used in the denominator instead of N_p . Equation (6.84) can be simplified to:

$$SEP^2 \sim RMSEP^2 - BIAS^2 \quad (6.85)$$

The SEP measures the precision of prediction (i.e. the difference between repeated measurements) while the RMSEP measures its accuracy (the difference between the true and estimated y value). If only the SEP data are presented with no report of the BIAS then the results can be misleading and give an over-optimistic impression of the predictive ability of a model (Naes et al., 2007).

The **Standard Error of Calibration (SEC)** and **Standard Error of Cross Validation (SECV)** are calculated via bias adjustment in the same way as for the SEP.

Standard Error of Laboratory (SEL):

If duplicates of a sample are analysed then then the SEL of the reference (wet-chemical) analysis can be determined as:

$$SEL = \sqrt{\frac{\sum_{i=1}^m (y_1 - y_2)^2}{m}} \quad (6.86)$$

Where $y_1 - y_2$ is the absolute difference between the two values (it can also be the difference between the results from two laboratories/users) and m is the number of batches. The SEL can be compared against the SEP and other predictive statistics; it is rare for the SEL to be less than the SEP.

Student's t test for the residual:

$$t = \frac{\text{Residual}}{\text{SEC}} = \frac{(y_i - \hat{y}_i)}{\text{SEC}} \quad (6.87)$$

This statistic compares the residual of a particular sample compared against the SEC. Absolute values greater than 2.5 are considered to be statistically significant and the sample could potentially be a y -outlier. This is most likely to be the result of a laboratory error so the experimental procedure should be repeated. Alternatively the sample may not have been presented well to the NIR cell and the NIR analysis should be repeated (Workman, 2001).

RPD:

RPD stands for the **R**atio of standard error of **P**erformance to standard **D**eviation and is calculated as (Fearn, 2002):

$$RPD = \frac{s_y}{SEP} \quad (6.88)$$

The RPD is a dimensionless statistic meaning that it can be compared on the same basis between calibrations that use different units for other statistics (e.g. % dry-matter versus pH for the units of the SEP in two calibrations). If the SEP is equal to the standard deviation of y (RPD=1.0) then the calibration model is not predicting the reference values at all.

RER:

RER stands for Range Error Ratio and it is equal to the range of the reference values divided by the SEP (Fearn, 2002):

$$RER = \frac{Range_y}{SEP} \quad (6.89)$$

The numbers obtained for the RER will typically be around four to five times larger than those for the RPD; however, the exact relationship between the two will depend on the distribution of samples in the validation set (Fearn, 2002). The result of the RER will also be more influenced by single extreme results than the RPD.

6.8 Techniques for Comparing Calibration Models

In a regression or factorisation model it is important that an appropriate number of factors (e.g. PCs, PLS factors etc.) are chosen in order that the calibration set is well modelled but not overfitted. Hence the model will be relevant to future unknown samples. If there is only a calibration set, with no validation (cross, fitting or test set) techniques employed there will be a problem in that increased numbers of latent variables will reduce the SEC and increase the correlation coefficient until it is equal to one (usually where the number of factors, A , is equal to the rank of \mathbf{X}). There are techniques, such as using Mallows's statistic (Naes et al., 2007), to penalise additional variables, however, more useful methods for determining the number of latent variables to use become available when validation techniques are available. An important statistic that can be used in this case is the **Prediction Error Sum of Squares** (PRESS).

When this statistic is used on the calibration set the process usually starts with the development of a single factor model from a subset which consists of the calibration set with one sample excluded. The model is then used to predict the excluded sample and the residual recorded. This procedure is then repeated for the entire sample set (as with cross validation) and a sum of squares for the residuals (SS_{res}) is reported, i.e. the PRESS. The algorithm then adds another factor and the sequence is repeated. The process will continue until all factors have been computed or a predetermined factor limit has been reached (Li et al., 2002). The result will be a series of PRESS statistics, one for each factor, that can be

plotted in in order to visualise the significance of each factor. Such a plot is included in Figure 6-8. There are various different criteria that can then be used to determine, from the PRESS statistics, the number of latent variables (A) that would be optimal for the model. A common approach is to set A as equal to the number of variables that give the minimum PRESS. Alternatively Wold's R criterion (Wold, 1978) can be used. Equation (6.90) shows the calculation of the R statistic.

$$R = PRESS(m + 1)/PRESS(m) \quad (6.90)$$

The process will increment m until the value for R is greater than one and at this point will set $A=m$. This method takes the first local minimum in the PRESS plot and can sometimes result in a value for A that is lower than when the minimum PRESS is selected as shown in Figure 6-8 which plots the PRESS and SECV for PLS models (for the extractives-free glucose content of wet *Miscanthus* internode sections) covering between 1 and 16 factors. The Vision software package uses Wold's R criterion and so would choose point A as the local minimum and therefore automatically choose 5 factors as the optimum number for the model. The global minimum of both plots occurs at 11 factors, however. It is therefore important to observe the PRESS plot before accepting the choice made by the software especially since, in extreme cases, it can occur that Wold's R criterion can stop at Factor 1 if there is a small increase in the PRESS for Factor 2, and any significant improvement in the statistic in subsequent factors would be ignored by the software.

The adjusted Wold's R criterion uses 0.95 and 0.90 as thresholds for the value of R , the reasoning being that an additional latent variable will only be included in the model if it results in a significant improvement in the PRESS (Krzanowski, 1987). Alternatively an F -test criterion was suggested by (Osten, 1988) with:

$$F = \frac{PRESS(m) - PRESS(m + 1)}{K} \bigg/ \frac{PRESS(m + 1)}{NK - (m + 1)K} \quad (6.91)$$

Where K is equal to N divided by the number of samples excluded in each cross-validation iteration for PRESS (i.e. in full cross validation $K=N$). Where the critical F -value was defined as $F_{K,NK-(m+1)K,0.05}$.

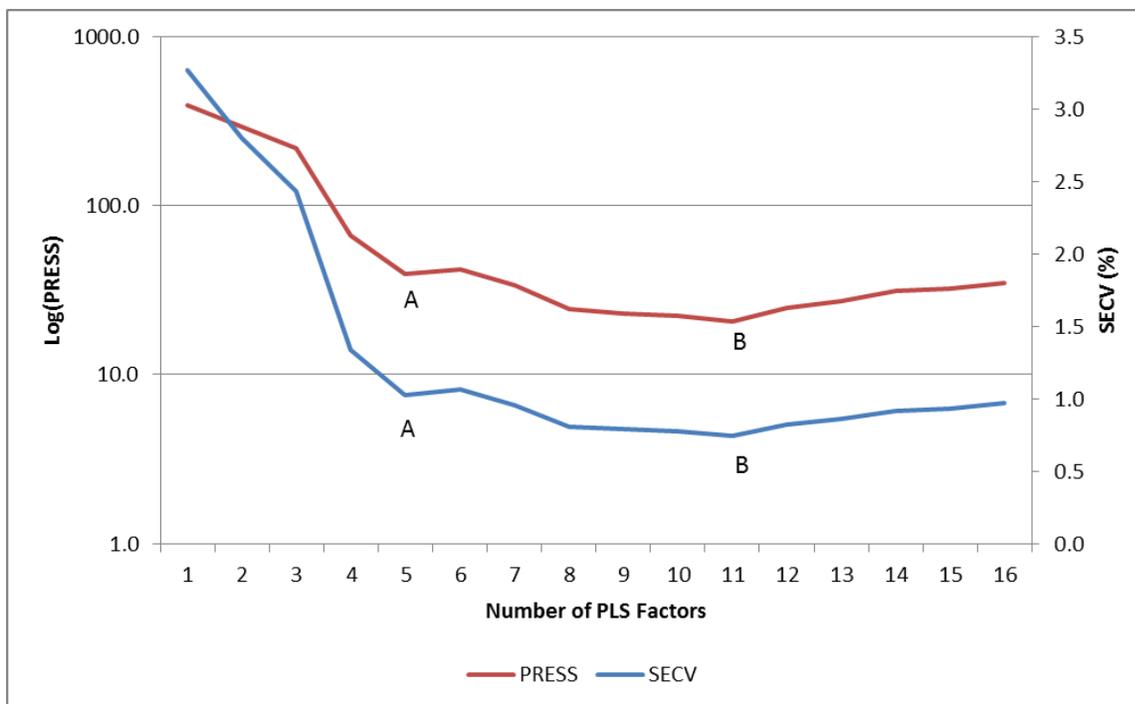


Figure 6-8: A chart plotting the PRESS statistic and SECV for PLS models incorporating various numbers of factors. The models were based on the glucose content (extractives-free basis) of wet miscanthus samples. Vision uses Wold's R-criterion and would select the value for A corresponding to the first PRESS minimum (i.e. 5 factors, point A). The global PRESS and SECV minimum occurs at 11 factors (point B).

Another PRESS-based criterion was introduced by Haaland and Thomas(1988). This uses an F -test to compare the number of factors for a model that yields the minimum PRESS, m^* , with all models with fewer factors. The number of factors chosen will be the minimum value for m for which the PRESS associated with that model is not significantly greater than for m^* . Therefore the F -value is calculated as:

$$F(m) = \frac{PRESS(m)}{PRESS(m^*)} \quad (6.92)$$

This is then compared with a tabulated value of $F_{N,N,\alpha}$ and the smallest m is chosen such that $F(m) < F_{N,N,\alpha}$. The value for α can vary but a value of 0.25 (i.e. a 75% probability level) is often selected (Iñón et al., 2003). The advantage of this method can be seen in that early local PRESS minima will only be chosen if they are reasonably close to the global PRESS minimum.

Blanco *et al.* (1996) compared various PRESS methods, including those mentioned above, for determining the optimal number of PCs to include in a model for determining the concentration of a

syrup flavouring agent. They found that the Haaland and Thomas criterion gave a model with the lowest SEP, while the methods involving the first local minimum resulted in marked underfitting.

A **RMSECV** algorithm follows a similar process to the PRESS statistic except the RMSECV (i.e. $\sqrt{MS_{res}}$) is the statistic that is determined.

Both the PRESS and SECV methods, as described above, involve fitting the model using the calibration set, however, (as mentioned earlier) an independent set can be used to determine the optimum model. This should not be the test set, since that set should ideally be completely independent from any model development; instead a separate fitting/training set should be used. It is crucial, however, that the structure of this training set is similar to that of the calibration set and the expected distribution of future samples that may be predicted. Using a test/fitting set the determination of the optimum model is achieved through the same process as mentioned above except the SEP is often used as the deciding statistic. This can be plotted over A just like the SECV and PRESS.

6.9 Partial Least Squares (PLS) Regression

The PLS methodology takes a different approach from PCA in the decomposition of the matrix \mathbf{X} (representing the spectra of the samples in the calibration set). PCA obtains PCs purely based on the variability in \mathbf{X} (eigenvectors for the covariance matrix $\mathbf{X}^t\mathbf{X}$ are sought) with no consideration made for \mathbf{Y} (the matrix, or vector if only one constituent is used, containing the reference lab values for the samples). The \mathbf{Y} data are only introduced at the PCR (regression) stage, i.e. once the PCs are already developed, meaning that the first few latent variables (which incorporate the least errors in factorisation methods) may explain little variance in the \mathbf{Y} matrix. This can often be seen in an Explained Y Variance plot with PCR where only later PCs start to account for the significant variance. This will be particularly true for minor components in the matrix that will be responsible for relatively small variations in the \mathbf{X} data.

PLS attempts to introduce the variation in \mathbf{Y} into the factorisation process while still developing latent variables (scores) that are orthogonal to each other, so avoiding the problems associated with MLR (which is a process totally dependent on the variation in \mathbf{Y}). This results in predictive accuracies that are comparable to PCR but involve less latent variables that incorporate less of the (unrelated) \mathbf{X} variance.

The PLS process is bilinear (as PCA) because the \mathbf{X} and \mathbf{Y} data can be constructed from linear combinations of their scores and loadings. The term partial in PLS refers to the fact that various parameters are estimated in the different partial modelling steps rather than in one main parameter evaluation process (Bjørsvik and Martens, 2001). Occasionally PLS is also referred to as Projection to Latent Structures.

PLS theory first defines the **outer relation**, which can be considered to be the separate bilinear decomposition of the \mathbf{X} and \mathbf{Y} matrices into their individual scores and loadings matrices as outlined below:

$$\mathbf{X} = \mathbf{TP}^t + \mathbf{E} \quad (6.93)$$

$$\mathbf{Y} = \mathbf{UQ}^t + \mathbf{F} \quad (6.94)$$

Where the matrix \mathbf{T} represents the factor scores for the new coordinates of the data points in the \mathbf{X} -space and matrix \mathbf{P}^t contains the \mathbf{X} loadings vectors. Similarly, matrix \mathbf{U} contains the factor scores and \mathbf{Q} the loadings from the factorisation of \mathbf{Y} . Matrices \mathbf{E} and \mathbf{F} are the residual matrices, for \mathbf{X} and \mathbf{Y} respectively, which will be needed if full factorisation does not occur.

6.9.1 NIPALS Algorithm

The conventional NIPALS PCA procedure (see Section 6.4.2.1) would follow the following steps for each block (Geladi and Kowalski, 1986):

X-block

1. Take $\mathbf{t}_{\text{start}}$ as a column of \mathbf{X} .
2. $\hat{\mathbf{p}}_1^t = (\hat{\mathbf{t}}_1^t \hat{\mathbf{t}}_1^t)^{-1} \hat{\mathbf{t}}_1^t \mathbf{X}$
3. Normalise $\hat{\mathbf{p}}_1$, i.e. $\mathbf{p}^t \mathbf{p} = 1$
4. Determine $\hat{\mathbf{t}}_1$ by projection of \mathbf{X} on $\hat{\mathbf{p}}_1$:

$$\hat{\mathbf{t}}_1 = \mathbf{X} \hat{\mathbf{p}}_1 (\hat{\mathbf{p}}_1^t \hat{\mathbf{p}}_1)^{-1} = \mathbf{X} \hat{\mathbf{p}}_1 \text{ (since } \hat{\mathbf{p}}_1 \text{ is normalised)}$$

5. Compare the new value of \mathbf{t} with the old one and if they are the same stop the algorithm, otherwise go to (2)

Y-block

1. Take $\mathbf{u}_{\text{start}}$ as a column of \mathbf{Y} .
2. $\hat{\mathbf{q}}_1^t = (\hat{\mathbf{u}}_1^t \hat{\mathbf{u}}_1)^{-1} \hat{\mathbf{u}}_1^t \mathbf{Y}$
3. Normalise $\hat{\mathbf{q}}_1$
4. $\hat{\mathbf{u}}_1 = \mathbf{Y} \hat{\mathbf{q}}_1$
5. Compare the new value of \mathbf{u} with the old one, and if they are the same stop the algorithm.

This above process is clearly separate for \mathbf{X} and \mathbf{Y} . Because of this the score matrices \mathbf{U} and \mathbf{T} are likely to be quite different. The relationship between these scores can be represented by what is known as the **inner relation**.

$$\mathbf{U} = \mathbf{T}\mathbf{D} + \mathbf{H} \quad (6.95)$$

This relation is a linear regression model between \mathbf{T} and \mathbf{U} , with the regression coefficients stored in the diagonal matrix \mathbf{D} and the residuals in the matrix \mathbf{H} .

This inner relation can be combined with the outer relation to give the mixed relation as outlined below:

$$\begin{aligned} \mathbf{Y} &= \mathbf{U}\mathbf{Q}^t + \mathbf{F} = (\mathbf{T}\mathbf{D} + \mathbf{H})\mathbf{Q}^t + \mathbf{F} = \mathbf{T}\mathbf{D}\mathbf{Q}^t + (\mathbf{H}\mathbf{Q}^t + \mathbf{F}) \\ &= \mathbf{T}\mathbf{D}\mathbf{Q}^t + \mathbf{F}^* \end{aligned} \quad (6.96)$$

Where \mathbf{F}^* is the new residual matrix

It can be seen that the outer relation mentioned previously, which used separate scores for \mathbf{X} and \mathbf{Y} has now been simplified by using the scores matrix \mathbf{T} for both. The target in PLS is to determine scores, coefficients and loadings that will minimise \mathbf{F}^* . Some articles in the literature replace the $\mathbf{D}\mathbf{Q}^t$ term in the above equation with a new loading vector \mathbf{C}^t (i.e. $\mathbf{Y} = \mathbf{T}\mathbf{C}^t + \mathbf{F}$) whereas others do not include the diagonal matrix and simply write the equation in terms of the \mathbf{Y} loading weight vector \mathbf{Q}^t and the scores matrix \mathbf{T} , i.e. $\mathbf{Y} = \mathbf{T}\mathbf{Q}^t + \mathbf{F}$, this is reasonable since the inner relation regression coefficient for each factor is simply a constant.

The inner relation of scores can be improved by exchanging information between the two block algorithms, by letting \mathbf{t} and \mathbf{u} change place in Step (2). This allows the algorithms to be written in sequence as follows (it is generally the case that these steps are preceded by the mean centring of \mathbf{X} and \mathbf{Y}):

1. Take $\mathbf{u}_{\text{start}}$ as a column of \mathbf{Y} .
2. $\hat{\mathbf{p}}_1^t = (\hat{\mathbf{u}}_1^t \hat{\mathbf{u}}_1)^{-1} \hat{\mathbf{u}}_1^t \mathbf{X}$ $[\hat{\mathbf{w}}_1^t = (\hat{\mathbf{u}}_1^t \hat{\mathbf{u}}_1)^{-1} \hat{\mathbf{u}}_1^t \mathbf{X}]$
3. Normalise $\hat{\mathbf{p}}_1$ $[\hat{\mathbf{w}}_1^t \hat{\mathbf{w}}_1 = 1]$
4. $\hat{\mathbf{t}}_1 = \mathbf{X} \hat{\mathbf{p}}_1$ $[\hat{\mathbf{t}}_1 = \mathbf{X} \hat{\mathbf{w}}_1]$
5. $\hat{\mathbf{q}}_1^t = (\hat{\mathbf{t}}_1^t \hat{\mathbf{t}}_1)^{-1} \hat{\mathbf{t}}_1^t \mathbf{Y}$
6. Normalise $\hat{\mathbf{q}}_1$
7. $\hat{\mathbf{u}}_1 = \mathbf{Y} \hat{\mathbf{q}}_1$
8. Compare the \mathbf{t} in (4) with the old one and if they are the same stop the algorithm, otherwise go to (2).

In starting with a column of \mathbf{Y} for the determination of the scores vector the influence of the reference lab values is kept high in determining the PLS factors.

However this method will not give orthogonal values for the scores matrix \mathbf{T} because the order of calculations that was used for PCA has changed (Geladi and Kowalski, 1986). In order to compensate for this the \mathbf{X} loadings, \mathbf{P}^t , are replaced by \mathbf{X} weights \mathbf{W}^t in the sequence above (as shown in the bracketed equations). The matrix \mathbf{W} is known as the loading weights matrix. The relation between \mathbf{W}^t , \mathbf{T} and \mathbf{X} is outlined below:

$$\mathbf{X} = \mathbf{T}\mathbf{W}^t + \mathbf{R} \tag{6.97}$$

Convergence is determined based on the change in the scores vector, i.e. $\|\hat{\mathbf{t}}_{\text{old}} - \hat{\mathbf{t}}_{\text{new}}\| / \|\hat{\mathbf{t}}_{\text{new}}\| < \varepsilon$, where ε is a small value such as 10^{-6} or 10^{-8} (Wold et al., 2001). Following convergence the following steps take place:

9. Determine $\hat{\mathbf{p}}_1$ as the loading vector that contains the coefficients of the regression of \mathbf{X} on $\hat{\mathbf{t}}_1$ and is equal to:

$$\hat{\mathbf{p}}_1 = \frac{\mathbf{X}^t \hat{\mathbf{t}}_1}{(\hat{\mathbf{t}}_1^t \hat{\mathbf{t}}_1)}$$

10. Given the inner relation, $\mathbf{U} = \mathbf{T}\mathbf{D} + \mathbf{H}$, a least squares estimate for \mathbf{D} will give the following constant:

$$d_1 = \frac{(\hat{\mathbf{u}}_1^t \hat{\mathbf{t}}_1)}{\hat{\mathbf{t}}_1^t \hat{\mathbf{t}}_1} \quad (6.98)$$

11. The loading weight of subsequent factors, $\hat{\mathbf{w}}_a$, and subsequent vectors of $\hat{\mathbf{P}}$ and $\hat{\mathbf{Q}}$, are defined through the same process except the \mathbf{X} matrix is deflated and replaced by its residual, $\hat{\mathbf{E}}_{a-1}$, the residual variability remaining after the previous $a-1$ factors have been extracted:

$$\hat{\mathbf{E}}_{a-1} = \mathbf{X} - \hat{\mathbf{t}}_1 \hat{\mathbf{p}}_1^T - \dots - \hat{\mathbf{t}}_{a-1} \hat{\mathbf{p}}_{a-1}^T \quad (6.99)$$

The deflation of \mathbf{X} will ensure that mutual orthogonality of all extracted score vectors in the algorithm

The \mathbf{Y} matrix can be replaced by its residual, $\hat{\mathbf{F}}_{a-1}$:

$$\hat{\mathbf{F}}_{a-1} = \mathbf{Y} - d_1 \hat{\mathbf{t}}_1 \hat{\mathbf{q}}_1^T - \dots - d_{a-1} \hat{\mathbf{t}}_{a-1} \hat{\mathbf{q}}_{a-1}^T \quad (6.100)$$

However, this is not necessary, the results are equivalent with or without \mathbf{Y} deflation (Wold et al., 2001).

12. $\hat{\mathbf{w}}_a$ will refer to the residual $\hat{\mathbf{E}}_{a-1}$ rather than the \mathbf{X} variables themselves; however, it is possible to transform these weights to $\hat{\mathbf{W}}^*$, which relate directly to \mathbf{X} , by (Wold et al., 2001):

$$\hat{\mathbf{W}}^* = \mathbf{W}(\mathbf{P}^t \mathbf{W})^{-1} \quad (6.101)$$

13. Repeat steps 1 to 12 until either \mathbf{X} has been fully depleted (i.e. A is equal to the rank of \mathbf{X} , in which case the regression is equivalent to multivariate regression on the original \mathbf{X} variables (Massart et al., 1998b)), or the desired number of factors/accuracy has been reached.

It can be said that PLS forms new \mathbf{X} variables, $\hat{\mathbf{t}}_a$, as linear combinations of the old \mathbf{X} variables, and uses these as predictors of \mathbf{Y} . In geometric interpretation, PLS involves the projection of \mathbf{X} onto an A -dimensional plane (where A is the number of factors) in such a way that the coordinates of the projection ($\hat{\mathbf{t}}_a$) are good predictors of \mathbf{Y} and also approximate \mathbf{X} well with the direction coefficients of the axes of this plane defined by $\hat{\mathbf{p}}_a$ (Massart et al., 1998b).

As with PCA, the number of latent variables (factors) that are used in the model is important and the variety of statistical tools (e.g. PRESS, SECV etc.) outlined in Section 6.8 can be used to determine the optimal number of components to allow good prediction for future unknown samples.

Once the desired number of factors have been determined it is possible to determine the \mathbf{y} values for an unknown sample, from its \mathbf{x} data, via two potential methods. The first involves determining the scores for the sample and includes the following steps (Massart et al., 1998b):

1. Firstly, the estimated \mathbf{y} values for the sample ($\hat{\mathbf{y}}_0$) are made equal to the mean values for the training data set.
2. The \mathbf{x} data of the unknown sample is mean centred (giving \mathbf{m}_0 for mean centred \mathbf{x}).
3. $t_1 = \mathbf{m}_0^t \mathbf{w}_1$
4. $\hat{\mathbf{y}}_1 = \hat{\mathbf{y}}_0 + d_1 t_1 \mathbf{q}_1^t$
5. $\mathbf{m}_1 = \mathbf{m}_0 - t_1 \mathbf{p}_1^t$
6. This is then repeated for the other dimensions with each step leading to an “improvement” in the estimate of $\hat{\mathbf{y}}$. For example, use of the second loading and loading weight vectors would involve:

$$t_2 = \mathbf{m}_1^t \mathbf{w}_2$$

$$\hat{\mathbf{y}}_2 = \hat{\mathbf{y}}_1 + d_2 t_2 \mathbf{q}_2^t$$

$$\mathbf{m}_2 = \mathbf{m}_1 - t_2 \mathbf{p}_2^t$$

7. This procedure is repeated for A cycles, where A is the total number of factors in the model.

The other method involves direct determination of the \mathbf{y} values via the \mathbf{x} data and the regressions coefficients matrix \mathbf{B}_{PLS} :

$$\hat{\mathbf{y}}_0 = \bar{\mathbf{y}} + (\mathbf{m}_0)^T \mathbf{B}_{PLS} \tag{6.102}$$

Where (Massart et al., 1998b):

$$\mathbf{B}_{PLS} = \hat{\mathbf{W}}(\hat{\mathbf{P}}^t \hat{\mathbf{W}})^{-1} \hat{\mathbf{Q}}^t \tag{6.103}$$

6.9.2 Other PLS Algorithms

The NIPALS method described in the preceding section is known as **PLS2** and is available in The Unscrambler X. It is used when there is more than one **Y** variable - a model is sought that optimises the covariance between linear combinations of the **X** data and linear combinations of the **Y** data. Hence a single model is derived for all dependent variables, e.g. cellulose, klason lignin, xylose etc. Since the **Y** data may have very different values and ranges it is typical for the columns to be autoscaled (see Section 8.3.2).

PLS1 is the model that is used when there is only one **y** variable, i.e. so **y** is now a vector instead of the matrix **Y**. Hence, in PLS1 the model optimises the covariance between **y** and linear functions of **X**. It is available to be used in The Unscrambler X and Vision. Exploratory PLS work can be done using PLS2 since it allows one model to be built for a matrix of **Y**-values and so will save time compared to examining specific models for each, and it may also provide a simpler model to interpret, particularly if the **Y** columns are correlated. However, in order to get the accurate final calibrations of uncorrelated **Y** variables it is usually better to use PLS1 regression.

It has been suggested (Wold et al., 2001) that, in order to decide whether to use PLS1 or PLS2, a PCA on the **Y** matrix should be carried out in order to determine the practical rank of **Y** and that, if this is small compared to the number of **Y** variables, PLS2 may be most suitable. However if the **Y** variables cluster into strong groups (as seen in the PCA loading plots) separate PLS models should be developed for these groups or for individual variables.

The NIPALS algorithm can handle missing data in the **X** and **Y** matrices (but to handle missing data in **Y** it must be multivariate) with approximately up to around 10-20% missing data being handled (Wold et al., 2001). This is possible because the missing values are substituted by predictions from the model. The Unscrambler X software also offers an older version of the NIPALS algorithm, referred to as Orthogonal scores PLS, however this algorithm does not handle missing values (CAMO, 2011).

An alternative way of looking at and understanding PLS is through the direct calculation of eigenvectors. Taking first the example of separate factorisation of **X** and **Y** (i.e. PCA of each), the following rules apply if the loading weights vectors are normalised:

For the **X**-block:

$$\mathbf{T} = \mathbf{X}\mathbf{W} \quad (6.104)$$

$$\mathbf{W} = \mathbf{X}^t\mathbf{T} \quad (6.105)$$

For the **Y**-block:

$$\mathbf{U} = \mathbf{Y}\mathbf{Q} \quad (6.106)$$

$$\mathbf{Q} = \mathbf{Y}^t\mathbf{U} \quad (6.107)$$

For each block the two equations can be linked giving that $\mathbf{w} = \mathbf{X}^t\mathbf{X}\mathbf{w}$ and $\mathbf{q} = \mathbf{Y}^t\mathbf{Y}\mathbf{q}$, meaning that \mathbf{w} is the first eigenvector of $\mathbf{X}^t\mathbf{X}$ and \mathbf{q} is the first eigenvector of $\mathbf{Y}^t\mathbf{Y}$. The equations can also be linked the other way implying that $\mathbf{t} = \mathbf{X}\mathbf{X}^t\mathbf{t}$ and $\mathbf{u} = \mathbf{Y}\mathbf{Y}^t\mathbf{u}$ meaning that \mathbf{t} is proportional to the eigenvector of $\mathbf{X}\mathbf{X}^t$ and \mathbf{u} is proportional to the eigenvector of $\mathbf{Y}\mathbf{Y}^t$. Therefore \mathbf{w} and \mathbf{t} form the first pair of singular vectors in the SVD of \mathbf{X} as do \mathbf{q} and \mathbf{u} for \mathbf{Y} . These scores are then fitted to \mathbf{X} or \mathbf{Y} and the residuals determined and from these residuals the dominant pairs of singular vectors are extracted as before. These correspond to the second pair of singular vectors of the starting matrices \mathbf{X} and \mathbf{Y} (Massart et al., 1998b).

The above process is clearly separate for \mathbf{X} and \mathbf{Y} . However utilising the same interchange of scores as was outlined for NIPALS (i.e. $\mathbf{w} = \mathbf{X}^t\mathbf{u}$ and $\mathbf{q} = \mathbf{Y}^t\mathbf{t}$) provides the following relationships (Wold et al., 2001):

$$\mathbf{w} = \mathbf{X}^t\mathbf{u} = \mathbf{X}^t\mathbf{Y}\mathbf{q} = \mathbf{X}^t\mathbf{Y}\mathbf{Y}^t\mathbf{t} = \mathbf{X}^t\mathbf{Y}\mathbf{Y}^t\mathbf{X}\mathbf{w} \quad (6.108)$$

$$\mathbf{q} = \mathbf{Y}^t\mathbf{t} = \mathbf{Y}^t\mathbf{X}\mathbf{w} = \mathbf{Y}^t\mathbf{X}\mathbf{X}^t\mathbf{u} = \mathbf{Y}^t\mathbf{X}\mathbf{X}^t\mathbf{Y}\mathbf{q} \quad (6.109)$$

From these equations it is clear that \mathbf{w} is an eigenvector of $\mathbf{X}^t\mathbf{Y}\mathbf{Y}^t\mathbf{X}$ and \mathbf{q} will be an eigenvector of $\mathbf{Y}^t\mathbf{X}\mathbf{X}^t\mathbf{Y}$. These two matrices represent the two symmetric products of $\mathbf{X}^t\mathbf{Y}$, i.e. if $\mathbf{X}^t\mathbf{Y}$ is \mathbf{B} then \mathbf{w} is an eigenvector of $\mathbf{B}\mathbf{B}^t$ and \mathbf{q} is an eigenvector of $\mathbf{B}^t\mathbf{B}$. Therefore \mathbf{w} is the first right singular vector of the covariance matrix $\mathbf{X}^t\mathbf{Y}$ and \mathbf{q} is the first left singular vector. It can also be demonstrated that the first scores vector \mathbf{t} is an eigenvector to $\mathbf{X}\mathbf{X}^t\mathbf{Y}\mathbf{Y}^t$ through the following sequence (Wold et al., 2001):

$$\mathbf{t} = \mathbf{X}\mathbf{w} = \mathbf{X}\mathbf{X}^t\mathbf{u} = \mathbf{X}\mathbf{X}^t\mathbf{Y}\mathbf{q} = \mathbf{X}\mathbf{X}^t\mathbf{Y}\mathbf{Y}^t\mathbf{t} \quad (6.110)$$

PLS aims to maximise the covariance between the \mathbf{X} and \mathbf{Y} scores vectors, i.e. $\max \text{Cov}(\mathbf{t}^t \mathbf{u})$. A covariance involves three terms as shown in Equation (6.111), or, through taking the squares, in Equation (6.112) (Massart et al., 1998b):

$$\text{Cov}(\mathbf{t}, \mathbf{u}) = \text{std_dev}_t \text{std_dev}_u r_{tu} \quad (6.111)$$

$$[\text{Cov}(\mathbf{t}, \mathbf{u})]^2 = \text{Var}(\mathbf{t}) \text{Var}(\mathbf{u}) r_{tu}^2 \quad (6.112)$$

Therefore, PLS regression, in maximising this covariance matrix through derivation of the scores for \mathbf{X} and \mathbf{Y} will develop latent factors that explain much of the variance in \mathbf{X} as well as much of the variance in \mathbf{Y} as well as keeping the correlation between each corresponding \mathbf{t} and \mathbf{u} factor high (Equation (6.112)). By replacing \mathbf{t} and \mathbf{u} with the relations outlined above, the importance of \mathbf{w}^t and \mathbf{q} become clear; they become the vectors that should be derived in order to maximise the following expression (Massart et al., 1998b):

$$\text{Cov}(\mathbf{t}, \mathbf{u}) = \mathbf{t}^t \mathbf{u} = (\mathbf{X}\mathbf{w})^t \mathbf{Y}\mathbf{q} = \mathbf{w}^t (\mathbf{X}^t \mathbf{Y}) \mathbf{q} \quad (6.113)$$

The **Kernel PLS** algorithm (Lindgren et al., 1993), which is also available in The Unscrambler X, follows the concepts outlined above and derives the loading weight matrix $\widehat{\mathbf{W}}$ through eigenvalue/eigenvector decomposition. This procedure calculates the “kernel matrices” $\mathbf{X}^t \mathbf{Y} \mathbf{Y}^t \mathbf{X}$ and $\mathbf{Y}^t \mathbf{X} \mathbf{X}^t \mathbf{Y}$ and uses eigenvalue decomposition to determine the weights, scores and loadings from these. The steps in this algorithm are summarised below:

1. Find \mathbf{w}_1 as the first eigenvector of $\mathbf{X}^t \mathbf{Y} \mathbf{Y}^t \mathbf{X}$.
2. Find \mathbf{q}_1 as the first eigenvector of $\mathbf{Y}^t \mathbf{X} \mathbf{X}^t \mathbf{Y}$.
3. Normalise both vectors so that $\|\mathbf{X}\mathbf{w}_1\| = \|\mathbf{Y}\mathbf{q}_1\| = 1$
4. Project the \mathbf{X} data on the \mathbf{X} weights to calculate the \mathbf{X} scores $\mathbf{t}_1 = \mathbf{X}\mathbf{w}_1$
5. Project the \mathbf{Y} data on the \mathbf{Y} weights to calculate the \mathbf{Y} scores $\mathbf{u}_1 = \mathbf{Y}\mathbf{q}_1$
6. Calculate \mathbf{X} loadings by regression $\mathbf{p}_1^t = (\mathbf{t}_1^t \mathbf{t}_1)^{-1} \mathbf{t}_1^t \mathbf{X}$
7. Deflate the data matrix $\mathbf{E}_1 = \mathbf{X} - \mathbf{t}_1 \mathbf{p}_1^T$
8. Repeat 1-7 using the deflated data until A factors have been found. No deflation of \mathbf{Y} is necessary

The Kernel PLS algorithm does not handle missing values and it is considered most appropriate for a “tall” data set, i.e. one with a large number of samples but much fewer variables (CAMO, 2011). The

Unscrambler X also offers **Wide-Kernel PLS** (Rännar et al., 1994) to be used. This process involves finding eigenvectors to the kernel matrix $\mathbf{XX}^t\mathbf{YY}^t$, which is a square non-symmetric matrix of size $(n \times n)$ where n is the number of samples. This algorithm can also not handle missing values and it is considered to be most suitable for a “wide” data set with many variables but much less samples (CAMO, 2011). This algorithm is therefore suitable for NIRS data since there will be many more variables (wavelengths) than samples in most cases.

6.9.3 PLS Analysis

Some properties of the different loading weights, scores and loading vectors of the factors are outlined below:

- The matrix of loading weights, $\widehat{\mathbf{W}}$, is orthonormal, i.e. $\widehat{\mathbf{W}}^T\widehat{\mathbf{W}} = \mathbf{I}$.
- The columns of the scores matrix are orthogonal ($\widehat{\mathbf{T}}^T\widehat{\mathbf{T}}$ is diagonal).
- The $\widehat{\mathbf{P}}$ loadings are not orthogonal, however.
- Also, the vectors of $\widehat{\mathbf{U}}$, the \mathbf{Y} scores matrix are not orthogonal to each other.
- The $\widehat{\mathbf{U}}$ and $\widehat{\mathbf{P}}$ vectors are orthogonal to those $\widehat{\mathbf{T}}$ and $\widehat{\mathbf{W}}$ vectors, respectively, that are one and more components earlier (Wold et al., 2001).
- Each factor describes as much as possible of the covariance between \mathbf{X} and \mathbf{Y} remaining after the contributions of the previous factors have been subtracted.
- Since \mathbf{X} and \mathbf{Y} are modelled, outliers in both can be spotted in the calibration set or for unknown objects.
- As with PCA, plots of loadings can help determine where noise is being included into the model. The loadings express how each of the \mathbf{X} and \mathbf{Y} variables are related to the model component summarised by the T scores (CAMO, 2011).
- The \mathbf{P} loadings express how much each \mathbf{X} -variable contributes to a factor.
- The \mathbf{Q} loadings express the direct relationship between the \mathbf{Y} -variables and the \mathbf{T} scores, hence projections of \mathbf{Y} -variables by a \mathbf{Q} vector can be used to interpret the meaning, in terms of sample variation in \mathbf{Y} , of the location of a projected data point on a \mathbf{T} -scores plot (CAMO, 2011).
- The \mathbf{W} loading weights express how the information in each \mathbf{X} -variable relates to the variation in \mathbf{Y} summarised by the \mathbf{U} scores. These loading weights express how the \mathbf{T} scores are to be

computed from the \mathbf{X} matrix to obtain an orthogonal decomposition. Variables that have large loading weight values are important in the prediction of \mathbf{Y} . The loading weights are normalised meaning that their direction and length can be interpreted (CAMO, 2011).

- As with PCA, plots of scores can be drawn, but PLS also allows the \mathbf{U} scores to be displayed. These \mathbf{U} scores summarize the part of the structure in \mathbf{Y} which is explained by \mathbf{X} along a given factor (CAMO, 2011). The \mathbf{T} scores therefore represent the coordinates of the data points in the \mathbf{X} space computed in such a way that they capture the part of the structure in \mathbf{X} that is most predictive for \mathbf{Y} . The \mathbf{U} scores are related to the \mathbf{T} scores by a constant as outlined in Equation (6.98).
- A plot of the \mathbf{T} Scores versus the \mathbf{U} Scores, known as an **X-Y Relationship Outliers** plot in in The Unscrambler X, is a helpful diagram for checking the relationship between the \mathbf{X} and \mathbf{Y} scores for each factor.
- As with PCR, sample/variable \mathbf{X} and \mathbf{Y} residuals can be calculated and plotted for each factor. These can be summed across samples to get an (\mathbf{X} or \mathbf{Y}) variance plot describing how the residual (or explained) variance of a variable changes with different numbers of factors, or it can be summed across variables to obtain a total variance curve describing the global fit of the model. Figure 6-9 shows an Explained \mathbf{Y} -Variance plot for cross-validation in a PLS2 model that calibrates for the arabinose, galactose, rhamnose, glucose, xylose, and mannose contents (all on an extractives-free basis) on a data set comprised of 52 wet peat samples (SG(1D) pretreatment). This plot helps to show the relative importance of each factor for each constituent.
- Influence, leverage, Hotelling T^2 , and Q-residuals plots are also possible, as with PCA.

The \mathbf{X} -loading weights (\mathbf{W}) and \mathbf{X} -loading (\mathbf{P}) vectors tend to be quite similar. Regions of discrepancy between $\hat{\mathbf{p}}$ and $\hat{\mathbf{w}}$ will exist in variables/factors where PLS is heavily influenced by the X-Y correlation more than in describing the variance of \mathbf{X} (Naes et al., 2007).

The performance of PLS in comparison to other methods will depend on the noise in \mathbf{X} , as well as on the relationship between \mathbf{X} and \mathbf{Y} and, unlike PCA, on the noise in \mathbf{Y} . Hence, if the reference laboratory data are not of a high quality the factorisation of \mathbf{X} will be poor and the model may overfit \mathbf{y} , by incorporating its noise in the model, but be a poor predictor for unknown samples.

The method can still be subject to non-linearity problems as is the case with PCA and various methods may be needed to solve for these.

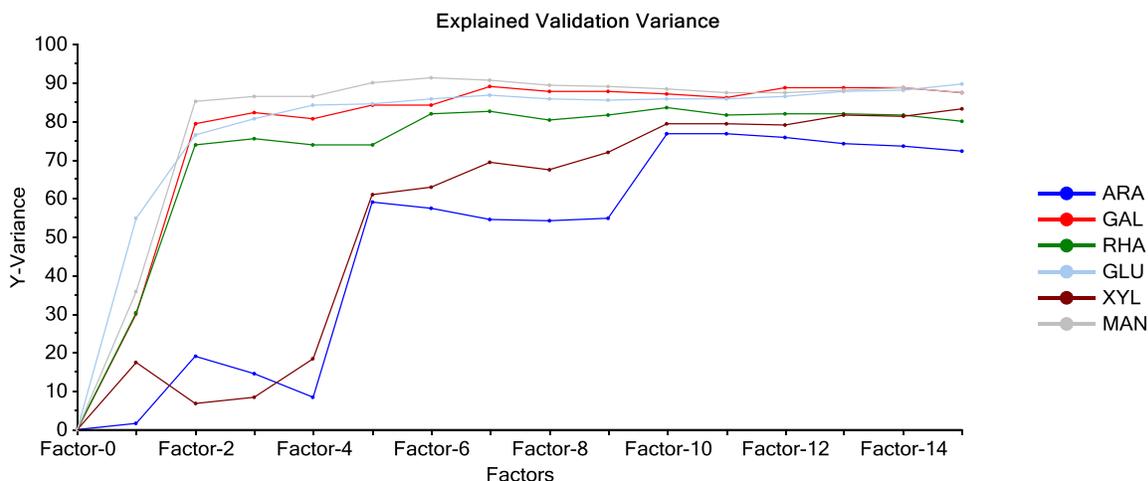


Figure 6-9: An Explained Y-Variance plot for (full) cross-validation of a PLS2 model for the 6 lignocellulosic sugars in 52 wet peat samples (a SG-1,1,10,10 (See Section 8.2.1) pretreatment was used on all spectra).

6.10 Non-Linearities in Calibration

Calibrations for the prediction of a range of y values (e.g. Klason lignin content) from a spectral data set (X) should ideally be accurate over the whole range of y . If instead the residual varies with increasing y then it is likely to be the case that there is a non-linear relationship between y and X . **Univariate nonlinearities** occur where modelling y from a single variable/factor does not give a linear relationship over y , but it is possible that moving to a multivariate calibration will allow a linear relationship. In contrast **multivariate non-linearities** mean that the regression surfaces can never be modelled as linear functions of the x -variables. A particularly problematic form of nonlinearity can occur when the absorption band of a constituent is lower than the bandwidth of the instrument, meaning that a smaller percentage of the total energy passing through or into the sample is attenuated by the absorbing substance (Mark, 2001a).

As mentioned previously, non-linearity in NIRS is partially addressed by the transform of spectral data to Absorbance = $\log(1/R)$ or (in some cases) the Kubelka-Munk transform (see Section 8.1). The various spectral pre-treatment techniques outlined in Chapter 8 (such as multiplicative scatter correction (MSC)) can also help to reduce non-linearities. Other potential methods to address non-linearity include:

1. **Add extra X terms to the model:** These can include squares or products of the variables. The application of PCA to the modified X will solve any collinearity problems that may result from this treatment.
2. **Use a limited wavelength range:** It has been said that the linearity between concentration and absorbance is reduced over around 1800 nm (Naes et al., 2007). Therefore calibrations could be attempted for subsections of the spectra or in situations where certain variables that may contribute to non-linearity have been removed.
3. **Eliminate Useless Variables in Prediction:** This differs from method (2) in that, rather than selecting a specific spectral range, the method eliminates specific wavelengths that are not relevant to the calibration.
4. **Transform the y values.** Naes *et al.* (2007) suggest a power model for the transformation of y , known as the Box-Cox transformation; however, the author has noted that no papers regarding the use of this technique in NIRS had been published.
5. **Locally Weighted Regression (LWR):** This is a method whereby it is assumed that linearity exists for spectrally similar samples. PCA is carried out on a calibration set and, for the prediction of an unknown sample, its spectrum is projected onto the first few PCs of this model and a number of the nearest neighbours in the factor space from the calibration set are selected and PCR is conducted on these. LWR can also involve a weighting of the samples in the new regression set according to how far they are from the prediction sample; for example, Cleveland and Devlin (1988) proposed to use a cubic weight function defined by:

$$W(d) = (1 - d^3)^3 \quad 0 \leq d \leq 1 \quad (6.114)$$

Where d is the distance from the prediction sample to a calibration sample. The distance will be scaled such that the distance from the furthest away calibration sample within the local set is equal to one.

There are complications brought forward by LWR, however, since it requires the optimisation of two parameters (number of local samples and number of factors) compared to just the number of factors for PLS/PCR. Cross-validation is typically used to determine these parameters. Also, the procedure is likely to require many more samples than PCR and PLS to build robust and relevant local models (Boysworth and Booksh, 2001).

6. **Development of linear subgroups in calibration:** This is different to LWR in that the development of linear subgroups occurs during the initial calibration and then an unknown sample is assigned to a certain subgroup (based on which subgroup is closest to it in the factor space, for example) and the regression coefficients from that subgroup used for prediction of that sample. This was a method proposed by Naes and Isaksson (1991) where the criterion for a definition of a subgroup involved a combination of the Mahalanobis distance in x -space and the y -residual meaning that samples within a subgroup are both spectrally similar and should exhibit a reasonably linear relationship between x and y . As with LWR, however, this method will require a relatively large sample set when compared with standard PCR/PLS.
7. **Range-splitting:** This technique, which is also known as bracketing (Workman, 2001), is similar to the subgroups method proposed by Naes and Isaksson (1991) except in this instance only the concentration range of the constituent is used as a basis for the different groups. For example, if it is noticed that there are non-linearities in calibrating for the cellulose content of *Miscanthus* over a concentration range of 10-40%, the range-splitting technique would develop several calibrations for split sections of this range (for example one calibration for 10-20%, another for 20-30%, etc.). If this method is used care should be taken to ensure that the X data in each subset are well distributed.

6.11 Important Considerations When Developing Calibration Models

Developing an accurate and robust calibration that can be suitable for the prediction of the composition of unknown samples requires planning and the strict adherence to well-thought out protocols throughout the whole analytical process so that errors are minimised. There can be three types of analytical errors (Mark, 2001a):

- a) Errors in the reference lab values (**Y**);
- b) Errors in the optical data (**X**);
- c) Systematic errors in the relationship between the two.

It is considered that the least squares regression method is only valid when (Mark, 2001a):

- i. There is minimal/no error in the independent **X** variables;
- ii. The error in the dependent **y** values is normally distributed;
- iii. The error in **Y** is constant for all values of the variable;
- iv. The regression coefficients are linear.

Regarding (iii), a plot of the lab error (e.g. standard deviation between duplicates) versus concentration could be useful in order to check for errors that are constant over **y**. NIR analysis assumes that there is minimal error in the x-variable (why it is called the independent variable) and the main source of error is in the y-variable (the dependent variable).

Errors can occur with spectral collection if:

- The method of sample presentation to the scanning cell of the spectrophotometer is not consistent.
- Samples are not homogeneous.
- There is drift in the spectrophotometer that is not regularly checked for and corrected for.
- The operating environment is unstable and this is not effectively corrected for with frequent scans of a reference.

Sample homogeneity will increase if a larger sample size can be scanned in one analysis and it will also increase with a reduction in the particle size of the material to be analysed. An alternative, or additional, technique for improving the **X** data can involve several repacks of the cell with the same sample and the

taking of multiple scans which can then be averaged for a final spectrum that can be used for calibration. This is a particularly important method if particle size reduction is not feasible or is not desired (see Section 11.1)

Regarding systematic errors in the relationship between lab reference values and spectral data (error (c) in the list above), these may occur where the reference method does not measure the same component as the spectroscopic method does. An example of this is the measurement of protein via elemental analysis or the Kjeldahl procedure. These reference methods will provide measurements of total nitrogen content whereas an NIR spectrum would include information on peptide bonds directly but not on total nitrogen (Workman, 2001).

The effect of outliers on calibration models is also important. The detection of spectral outliers has been discussed in Section 6.4.3.6. If a sample is determined to be a spectral outlier but is considered important to the model then more samples similar to it should be sought and included in the model. Alternatively the sample should be removed from the calibration/validation sets.

There can also be y outliers, where the residual is significantly greater than the SEP/SEC. Plots of the predicted y value versus the y residual can also be useful for spotting outliers and trends (e.g. non linearities) in the predictive abilities of the model. The first action on finding a y outlier should be to check the reference lab data for that sample has been inputted correctly and, if this does not help, a re-analysis of the sample should follow. If this still does not help then, as with the x-outlier, the relative importance of that sample should be considered before deciding whether to remove it as an outlier or whether more samples like it should be brought in to the calibration.

When an outlier has been removed the model development has to be repeated and the resulting model examined again for subsequent improvement.

As mentioned in Section 5.3.6, temperature can have an effect on the spectra of samples with even small changes affecting absorbance band sizes and positions. If a calibration is to be robust and suitable for the future prediction of unknowns, at an unknown later date, then it is important that the whole range of potential environmental conditions are covered when taking spectra for the calibration set. Hence, it is inadvisable to scan in one day all of the samples that will constitute the calibration set since environmental conditions in the future may differ from those experienced that day. The collection of spectra should therefore be spread out to cover the heterogeneity in conditions. Furthermore, the decision of which samples to scan on which day should be somewhat random since a more ordered

approach may result in a bias whereby the first type of samples (e.g. *Miscanthus* internodes) are all scanned in a period of time that may have different conditions from when another group of samples are analysed.

6.11.1 Sample Selection

This consideration also clearly extends to the samples themselves, as well as the environmental conditions. In order for regression methods to be effective in predicting a wide range of compositional values, it is important that this range is well represented in the calibration set. Furthermore, while in nature the distribution of, for example, the cellulose contents of a group of randomly selected samples of *Miscanthus* would be normal with a maximum at the mean and much fewer samples at the upper and lower ends of the concentration range, the distribution of samples in the calibration set should aim to be distributed reasonably evenly across that range. Hence, a pure random selection of samples for the calibration set would favour the normal distribution and would cause the mathematical model to most closely fit to the middle concentration samples while the few samples at very high or low levels would have extremely high leverage and therefore influence the slope and intercept inordinately (Workman, 2001). In contrast a more even concentration distribution among the calibration set will provide a more stable model that will not be highly influenced by individual samples and their potential errors, since the leverage of the samples towards either end of the calibration set will be decreased. Similarly, where calibration sets will include a variety of different sample types these should, as much as possible, be equally represented in the set in order to avoid excessive leverage.

It is also important to consider the variation of other properties of the sample, even when they may not be directly analysed for. This is mostly important for properties that will influence the spectra of samples and, hence, will play a role in the development of factorisation models. A clear example is moisture content. Section 5.3.5 details the effect that moisture can have on spectra and the position and shape of many absorbance bands in the NIR region. Given the large influence that moisture has on spectra and model development, it is important that samples covering a wide range of moisture contents be analysed spectrally, and subsequently by reference lab methods for the properties of interest. It will also be important that particle size variations are covered, in situations where particle size reduction prior to

NIR analysis is not desired. Also, where possible the sample set should avoid intercorrelation between the components of interest and moisture, particle size, or other spectrally important properties.

As mentioned previously, the validation and (if used) fitting sets should have a similar structure to the calibration set and so all of the points discussed above regarding the selection of samples for the calibration set also apply for these sets.

Therefore, there are two important areas regarding sample selection:

- 1) Selection of samples for NIR and reference analysis.
- 2) Selection of those samples analysed for inclusion in the calibration, (fitting) and test sets.

6.11.1.1 Selection for NIR and Reference Analysis

Point 1 shall be examined first. The first element concerns sample collection. There are, in many cases, an almost infinite different number of samples that can be obtained for a potential experiment/calibration. Regarding the energy crop *Miscanthus*, as of 2011 there are several thousand hectares of the crop under production and clearly sampling every hectare would not be practical. Sample collection should therefore follow reasonable hypotheses regarding how the physiochemical characteristics of the crop will change with various factors (such as stand age, time of year, crop variety, location etc., anatomical fraction) so that a more targeted strategy can be employed.

In many cases the sample collection methodology may follow principles devised for studies/experiments other than the development of quantitative NIR calibrations, in which case the NIR analysis can clearly only work with what samples these experiments make available. Alternatively, as in the case of the experiments on peat samples (Section 13) and Australian sugarcane bagasse samples (Section 12), the samples may be supplied by a third party, in which case the same principle applies.

Once a set of samples are available their spectra can be taken with NIRS, although this step may take a significant amount of time if some sample preparation is required. The resulting **X** matrix will contain the spectra of all the samples and can possibly be used as a basis for sample selection for reference analysis methods. Again, this principle will only apply if the basis for reference analysis is for the development of NIR calibration equations and not for other experiments. There are several techniques for sample selection at this stage:

1. **Random selection:** On deciding the number of samples that will be analysed these samples can be randomly selected from the global set. Alternatively, if an even distribution between different types/subgroups of the data-set (e.g. internodes, nodes, sheaths etc.) is desired then the samples can be subdivided into these groups and a chosen number selected randomly from each.

2. **Selection according to distance metrics used within the spectral/factor space:**
 - a. The Kennard-Stone algorithm. This is a stepwise procedure where new selections are taken in regions of space far from the samples already selected (Galvão et al., 2005). The Euclidean distance between the \mathbf{x} vectors of each pair of samples in the data-set is calculated and the pair of samples for which the distance is largest are selected. In each subsequent step the algorithm selects the sample that has the largest minimum distance to any of the samples that have already been selected. The algorithm stops when the desired number of samples have been selected.
 - b. A technique based on cluster analysis (see Section 7.1) was suggested by Naes (1987) and involved the following steps:
 - i. Carry out a PCA on the \mathbf{X} data and decide on the number of components that are relevant for the constituent of interest.
 - ii. Perform a cluster analysis on the standardised scores from this PCA, Naes (1987) used complete linkage, see Section 7.1, and stop the clustering when there are as many samples as possible for reference analysis. Standardisation (weighting to variance 1) was used to give all the PCs that were selected equal weight in the selection of samples.
 - iii. Take one sample from each cluster for reference analysis. The sample can be selected randomly or the sample within each cluster that is farthest away from the centre of the data as a whole can be used (Næs, 1987).
 - c. Distances based on the Mahalanobis distance, e.g. the standardised H statistic (known as Global H) and the neighbourhood H statistic, can be used as a metric by which to select samples for reference analysis (Shenk et al., 2008):

- d. Manual selection based on observation of spectra or 2D score plots: Samples can be manually selected from 2D scores plots involving the first few PCs with the target being that the samples selected are well distributed across this space. An alternative to manual selection is available in The Unscrambler X via the “Mark Evenly Distributed Samples” option. In this setting, for each PC the number of max/min samples to be selected can be specified by the user. Setting the number as 2 would select 2 (of each) of the extreme maximum and minimum samples along the PC axis, resulting in a total of four samples selected. Alternatively the number of classes into which each PC is divided can be specified, one pair of max/min samples will be selected from each class; hence, choosing 2 classes in this instance would also result in a total of four classes being selected.
3. **Selection based on the predicted Y values of the scanned samples:** This technique is only possible if a suitable calibration equation can be applied to the **X** data. There are two possible types of equation that could be used here:
- a. A previous equation that has been developed on the sample type (e.g. Miscanthus) and requires additional samples to improve its predictive ability. Figure 6-10 (a) shows a plot of reference lab values for the extractives-free glucose content of Miscanthus internode samples that were used in the development of an early calibration equation based on the spectra taken of these samples when they were in the wet-unground state. It can be seen that there are some regions of the concentration range that are not represented by the calibration samples (for example in the concentration range 42-44%). This early equation was applied to the 479 spectra of wet Miscanthus samples that had not, to date, been analysed via wet reference methods and the plot of Figure 6-10 (b) was produced. From this plot some samples that could fill in the missing areas in Figure 6-10 (a) were identified, processed, and analysed for their reference lab values.

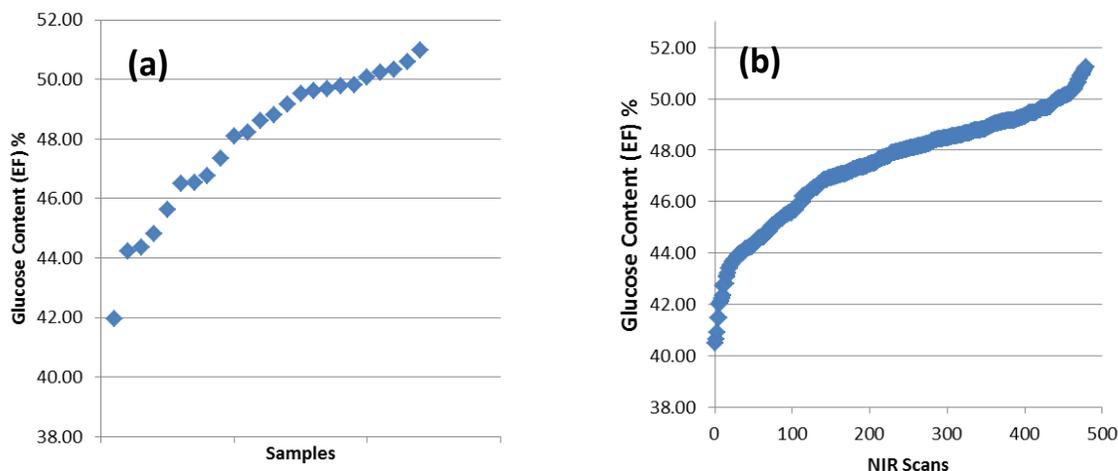


Figure 6-10: Glucose contents of samples in the calibration set and predicted glucose contents of samples not analysed by reference methods. (a) A plot of the reference lab values for the glucose content (extractives-free basis) for *Miscanthus* internode sections that were used in the development of an early PLS calibration model. This model was applied to 479 NIR scans of various *Miscanthus* samples that had not been analysed via wet chemical methods and the glucose content predicted from these scans, shown in (b).

- b. An equation developed for another feedstock. An example could be applying the wet-*Miscanthus* glucose content calibration equation to the spectra of wet sugarcane bagasse samples (see Section 18.2.1.2). Some of the methods discussed in Section 6.11.1.2 could then be applied in order to select a set of samples based on their (predicted) Y values. Caution should be taken when employing this method since accurate prediction may not be possible using the selected calibration equation and the reference lab values obtained from samples collected under this protocol should be checked, at an early stage and then at regular intervals, against the predicted values to examine the closeness/absence of correlation.

4. **Selection based on X and predicted Y:** A combination of (3) and (4) could be employed.

6.11.1.2 Selection for Calibration, Fitting and Validation Sets

Once a data-set of reference lab values and their associated spectra is compiled, sample selection within this group will be necessary for the calibration, fitting and validation sets. If the samples were selected under a different methodology than outlined in the preceding section then it may not be the best option to make all of the samples for which data exist available for calibration development/testing since this may bias the data-set towards certain subgroups of the set, or result in a normal distribution in the spread of the constituents of interest. There are six main techniques that can be employed in selecting samples:

- 1) Manual selection.
- 2) Random selection.
- 3) Selection according to **X**.
- 4) Selection according to **Y**.
- 5) Selection according to **X** and **Y**.
- 6) Selection according to other criteria (e.g. non-calibration **Y** values).

Technique 2 can occur for all sets simultaneously with the user specifying the proportion of samples that will go to each set (e.g. 50% calibration set, 25% fitting set, 25% test set). With techniques 3-6, there can be various orders by which samples are selected for each of the sets. These can be as follows (it is assumed that any outliers have been removed at this point):

- a) Select a global sub-set of the main data set according to techniques 3-6 (e.g. a total of 100 samples from the larger set of 200) and then distribute the samples within this set randomly (with selected proportions) to the calibration/fitting/test sets; or
- b) Select a specified number of samples to the calibration set via methods 3 -6 and then send the “redundant” samples to the fitting/test sets; or
- c) Retain any clusters/subgroups that may have been formed under techniques 3-6 and select a proportion of samples for each set from each cluster.
- d) Where possible, use an iterative process for methods 3-5 with (for example) the first two samples selected going to the calibration set, the third to the fitting set, and the fourth to the test set with this sequence repeated until techniques 3-5 have completed.

Selection according to X:

The same principles as outlined in Section 6.11.1.1 will apply.

Selection according to Y:

In describing the distribution of samples according to constituent values there are some useful statistics and plots that can be used. Figure 6-11 shows a Histogram plot for the (extractives-free) glucose content of 112 Miscanthus samples. A total of 20 bars were selected for this plot which covers a range of 31.9-51.2% glucose (i.e. 19.3%). The red line superimposed on the plot represents a normal distribution.

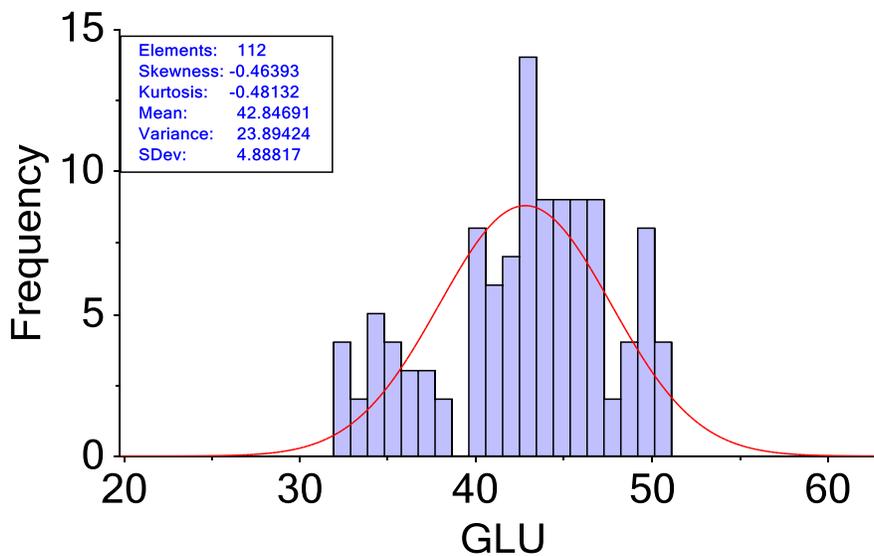


Figure 6-11: A histogram plot for the glucose content (extractives-free basis) of 112 Miscanthus samples. The plot includes the normal distribution (red line) and various statistics to describe the data.

Figure 6-11 also includes various statistics to describe the y values including the mean, standard deviation, and variance. The skewness and kurtosis are also provided. The **skewness** statistic describes the asymmetry of a histogram with distributions that have a skewness of 0 being symmetrical, those with a positive skewness having a longer tail to the right, whereas a negative skewness represent a longer tail to the left. The skewness is calculated as (CAMO, 2010):

$$Skewness = \frac{\frac{1}{N} \sum_{i=1}^N (y_i - \bar{y})^3}{\left(\sqrt{\frac{1}{N} \sum_{i=1}^N (y_i - \bar{y})^2} \right)^3} \quad (6.115)$$

The **kurtosis** statistic measures the flatness of a histogram with a value of 0 being the normal distribution. Distributions that have kurtosis values of more than 0 are more pointed in the middle whereas distributions with a kurtosis smaller than 0 are flatter or have thicker tails (this is also the case for symmetrical bi-modal distributions). The kurtosis is calculated as (CAMO, 2010):

$$Kurtosis = \frac{\frac{1}{N} \sum_{i=1}^N (x_i - \bar{x})^4}{\left(\sqrt{\frac{1}{N} \sum_{i=1}^N (x_i - \bar{x})^2} \right)^4} - 3 \quad (6.116)$$

A quantiles plot marks the median value (of the sorted vector) with the middle horizontal line, the lower (25%) quartile and upper quartile (75%) form the borders of the blue bar, and the minimum and maximum are represented by the horizontal lines at extreme ends of the distribution. The interquartile range (IQR) is defined as the value for the upper quartile minus that of the lower quartile

The most basic selection technique using the y values would be (where **y** is a vector rather than a matrix) to order the samples according to their constituent values and then select every x (e.g. 2nd) sample. This technique could incorporate subset method (d) for splitting samples between the sets. Such a process would in most cases not significantly change the structure (histogram) of the set, however.

Alternatively, many of the X-space distance metric tools, such as cluster methods, can be applied to the y space. Such tools can be used on a single **y** vector or on a **Y** matrix containing the values for a range of constituents. The different **Y** vectors can be normalised so that their absolute values do not dominate the distance calculations (i.e. constituents with significantly different average values, e.g. galactose and glucose, can have the same weight in the algorithms) meaning that their variance will be the deciding criterion.

Figure 6-12 shows the 21-bar histogram obtained for the (extractives-free) glucose content values for a data-set determined by the extraction of two samples per cluster whereby 21 clusters are formed from the data set of Figure 6-11 (hierarchical complete linkage clustering using Euclidean distance). It can be

seen that the histogram is much flatter than in Figure 6-11, representing a more even distribution of samples across the concentration range.

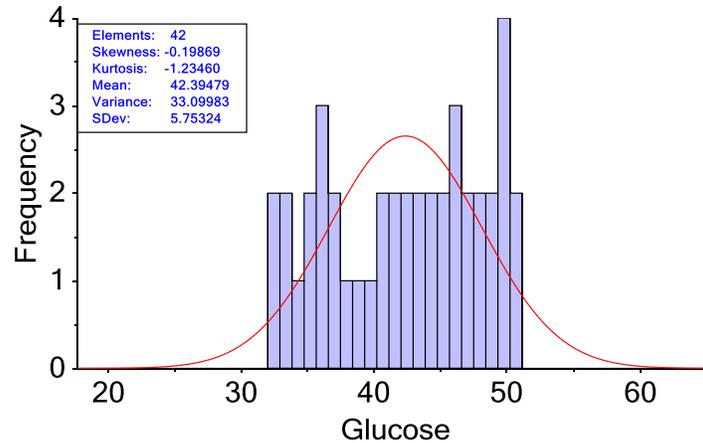


Figure 6-12: A histogram based on a cluster analysis of the data-set described in Figure 6-11 and the extraction of two samples per cluster. The constituent is extractives-free glucose (% of dry matter).

6.11.2 Quality Guidelines for Calibrations

This Chapter details a large number of statistics that can be used to assess the quality of a quantitative calibration model. Attempts have been made to define guidelines for these parameters so that quality thresholds for calibration models can be put forward. For instance, AACC Method 39-00.01 (AACC, 1999) lists quality thresholds and calibration guidelines for the quantitative analysis of cereals that include the following:

- Cross validation cannot be called independent validation.
- Independent validation should normally involve at least 30 samples.
- The following statistics should be reported: SEC, SEP, RER, RPD, Standard error of precision of the instrument, i.e. multiple tests on same sample.
- Recommended performance targets:
 - RER:
 - ≥ 4 – the calibration is acceptable for sample screening.
 - ≥ 10 – the calibration is acceptable for quality control.

- ≥ 15 – the calibration is good for quantification.
- Alternatively, RPD targets may be used:
 - ≥ 2.5 – the calibration is suitable for screening in breeding programs.
 - ≥ 5 – the calibration is acceptable for quality control.
 - ≥ 8 – the calibration is good for process control, development and applied research.

The RPD will vary according to the histogram of the concentrations of the calibrated-for constituent in the model and so can be subject to manipulation according to how that set is constructed. For example the standard deviation for the calibration set in Figure 6-12 is 5.75% while the standard deviation for the global-set, Figure 6-11, is 4.88%. An SEP of 0.7% would give an RPD of 8.2 in the first instance, representing a calibration of a suitable quality for applied research and process control, but only 6.97 in the second instance, indicating the calibration is only suitable for quality control.

It is considered by the Author that the RER value is a better test for the quality of the model, providing that there are no concentration outliers to inflate the value and that the concentration range of the constituent is well represented.

The calibration process can sometimes be a trade-off between robustness, in terms of the model being applicable to a wide variety of samples, and accuracy/precision, which will typically be high for calibration sets that have a smaller variance in sample type and constituent range but may fall if the set is broader. One alternative is to have both types of calibrations available. With the broad calibration being used for outlier detection or to determine which narrow-calibration model should be used for unknown samples. These are known as autocalibration models (Workman, 2001) and are similar to the subset and range-splitting concepts for non-linearities in calibrations (Section 6.10)

6.12 Summary

This chapter has described the latent variables that can be produced to remove collinearity problems from dataset matrices. These new variables allow for simpler models more relevant to the variation in the dataset. PLS was chosen as the main regression method for the analytical work in this Thesis since it finds variables that explain both the variance in X and that in Y. Numerous regression statistics are outlined and the importance of the root mean square error (RMSE) and standard error values explained in relation to the range or standard deviation of compositional values. These statistics are among the most important for assessing the accuracy of the chemometric models developed.

7 Qualitative Analysis and Sample Discrimination Techniques

As well as being able to quantify various physicochemical properties of samples, NIRS also has application in the field of sample identification/classification. Such a facility could be of great use in a biorefinery system. It is likely that many biorefineries, due to the seasonality of supply of many biomass resources, may receive a wide range of feedstocks over the course of a year. NIR-based classification could be used to identify the biomass-type prior to its processing. This would enable facility operators to modify the process conditions in order to encourage the maximal possible process yields from that feedstock. Such classification need not only be on a broad basis, for example discriminating between *Miscanthus* and sugarcane bagasse, but could discriminate over the range of potential varieties of a feedstock.

The result of a classification could be used to select the most appropriate quantitative calibration to be used for that feedstock – more targeted calibrations often involve fewer factors than global calibrations and can be more accurate in their predictions.

In this chapter various different methods for sample classification and discrimination will be discussed. The same general principles as with quantitative calibrations apply regarding qualitative analysis in that a model is trained on a set of samples in a calibration set. This model can then be validated via various methods and then applied to classify future samples based on their NIR spectra. The main different criterion in qualitative analysis is in the assignment of samples to classes. This can occur manually or through algorithms that group samples according to their spectral similarities via clustering methods.

7.1 Cluster Analysis

Clustering involves forming a series of subgroups of samples from a larger sample set whereby spectrally similar samples will occupy the same cluster/subgroup. In Unscrambler X there are two main methods for clustering (CAMO, 2011):

1. *K*-clustering methods or Partitioning-Optimisation techniques:
 - *K*-means clustering

- K -median clustering
2. Hierarchical clustering methods (hierarchical clustering analysis, or HCA) with different linkage measures:
 - Single-linkage
 - Complete-linkage
 - Average-linkage
 - Median-linkage
 - Ward's method.

7.1.1 K -Clustering

With K -clustering methods the target is to classify a data set into a user-defined number (K) of clusters. K -means (Faber, 1994, MacQueen, 1967) clustering involves the following steps:

1. K different seed points are chosen in the multidimensional space (in some algorithms it is a random sample that is chosen).
2. Each sample is then classified according to the cluster/seed-point to which it is nearest.
3. The vector describing the centroid position of each cluster is determined from the mean of each of the X -variables of the respective samples in the cluster.
4. These new centroids are used as the new data points for clustering development. Each of the samples is tested to see which centroid it is closest to:
 - a. If, upon examining the distance of a data-point from these new centroids, it is determined that this sample still belongs to the same cluster then the test moves to examine the next sample.
 - b. If it is found that the data-point is now closer to another centroid then it moves to that cluster and the centroids of these affected clusters are now recalculated.
5. Step 4 will loop until there is no change (over a loop containing all samples) in the position of all the centroids, at which point all samples will be assigned to their respective cluster.

The K -means algorithm developed by Lloyd (1982) differs from the one described above in that it only recalculates the centroids once all the samples have been assigned to clusters following the adjustment of the cluster centroids in the previous step.

All *K*-means methods can be susceptible to the initial choice for the seed points. Therefore Unscrambler X repeats the whole algorithm a number of times (the default is 50). The effectiveness of each clustering procedure is determined by examining the Sum of Distances (SOD) which is the sum of the distance values between each of the samples and its respective cluster summed up over all the clusters. A minimum SOD is sought.

K-median clustering uses the same principles as *K* means except the median value of the locations of all the samples in the cluster is used to locate the centroid. It is said to be more robust to outliers (CAMO, 2011).

7.1.2 Hierarchical Clustering (HCA)

Generally agglomerative hierarchical methods are used. These involve each sample being a separate cluster at the start with samples/clusters joining with each step in the algorithm. This method will require knowledge of the distances between all samples and a distance matrix can be used for this. This matrix will be a symmetric matrix with zeros along the diagonal. HCA results are presented as dendrogram plots which illustrate the clustering of samples with the distance between samples being on the x axis and the sample number on the y axis. A large distance between a cluster sub-division will be a sign of a good and distinct group structure. HCA agglomerative methods progress as follows for the different linkage methods:

HCA Single-Linkage: This is also known as the **nearest neighbour method**. The first step involves putting the two samples with the smallest distance value in the same cluster. The distance that a sample outside this cluster lies from that newly formed cluster is defined as the smallest of the distances between the sample and all of those samples in the cluster. Similarly, the distance between two clusters is equal to the smallest distance between the samples within each cluster. This allows a new distance matrix to be set up and, as before, the sample/cluster with the smallest distance between another sample/cluster is joined with that sample/cluster. This iterative process (resulting in the formation of a new distance matrix at each step) will continue until all the samples are in one cluster, or until the desired number of clusters have been reached.

It is considered that this method will make large clusters and will not provide a good distinction between groups that are different but not well separated (CAMO, 2011).

HCA Complete-Linkage: This is also known as the **farthest neighbour method**. This is the same as the nearest neighbour single linkage method described above except the criterion for deciding the distance between samples and clusters, or between clusters and clusters is based on the maximum distance between the samples instead of the minimum distance. These tend to be more compact and rounded clusters than in single linkage where the clusters can be elongated.

HCA Average-Linkage: This uses the average distance between samples in clusters as a criterion for the difference between clusters. It is said to be a compromise between the single and complete linkage methods.

HCA Median-Linkage: This instead uses the median value for the distance and is said to give similar results to average-linkage.

Ward's Method (Ward Jr., 1963): This method is somewhat different to the other agglomerative hierarchical approaches in that it uses analysis of variance as a criterion for cluster development. The method uses the following statistics, where x_{ijk} denotes the value for variable k of object j belonging to cluster i (Shalizi, 2009):

$$\text{Error Sum of Squares (ESS)} = \sum_i \sum_j \sum_k (x_{ijk} - \bar{x}_{ik})^2 \quad (7.1)$$

$$\text{Total Sum of Squares (TSS)} = \sum_i \sum_j \sum_k (x_{ijk} - \bar{x}_k)^2 \quad (7.2)$$

$$r^2 = \frac{TSS - ESS}{TSS} \quad (7.3)$$

Error sum of squares (ESS) compares the individual values for each variable against the cluster means for that variable, summed over all variables and all clusters. Total sum of squares (TSS) compares the individual values for each variable against the total mean for that variable, summed over all variables. Therefore the r^2 value will indicate the proportion of variance that is explained via a particular clustering set-up.

The Ward's algorithm starts with each sample as a cluster and forms the first cluster on the basis of which cluster forms the lowest ESS (or the highest r^2). This condition is repeated in each subsequent step of the agglomerative clustering method. Note, however, that this condition does not necessarily mean that the total ESS obtained with K clusters (where K is greater than $n-1$) will be the lowest ESS possible with that number of clusters. This is because the method is constrained in choosing which clusters to form based on the previous steps of cluster development (e.g. $K-1$, $K-2$ clusters etc.)

It is considered that average linkage and Ward's method generally give the best results (Massart et al., 1998b).

7.1.3 Distance Measures in Clustering

The Unscrambler X provides the following options for distance measures in HCA and K -means/medians cluster analysis:

- Euclidean distance (see Section 6.4.3.3).
- Squared Euclidean distance – This uses the same equation as the Euclidean distance metric, but does not take the square root. It is said that the result of K -means clustering will not be affected if this metric is used instead of the Euclidean distance, but the output of HCA may change.
- City-block distance – This is also known as the Manhattan distance. It is different from the Euclidean distance, which considers the distance as the shortest path between two samples, in that it refers to the sum of the distances in each dimension needed to reach the sample.
- Pearson correlation distance – This is equal to one minus the Pearson correlation coefficient between the two spectra (which can vary from 1 to -1), hence the Pearson correlation distance lies between 0 and 2. For this method each vector is centred by subtracting its mean and scaled by dividing by its standard deviation. The correlation coefficient can then be determined via the cross product of these vectors divided by, k , the dimension of these vectors.

- Uncentred correlation distance – This is the same as above except the sample means are set to zero.
- Absolute Pearson correlation distance – Here the absolute value of the Pearson correlation coefficient is used meaning that the distance can be between 0 and 1. This method means that negatively-correlated samples will get clustered together.
- Absolute uncentred correlation distance – the same as above but the means are set to zero.
- Spearman’s rank correlation distance – The Spearman’s rank correlation coefficient, ρ , can be thought of as the Pearson correlation coefficient between the ranked variables. Taking vectors \mathbf{x} and \mathbf{y} , each vector is ranked and the differences in rank d_i , at each position, are used to calculate ρ according to the formula below (Gauthier, 2001):

$$\rho = 1 - \frac{6 \sum_i d_i^2}{n(n^2 - 1)}$$

Where n is the dimension of the vector.

As with the Pearson correlation, the Spearman’s rank correlation value can range from -1 to +1. Because this method operates on ranks it is relatively insensitive to outliers (Gauthier, 2001). However the conversion to ranks does clearly result in a loss of information.

- The Kendall tau rank correlation coefficient distance – This technique says that any pair of observations (x_i, y_i) and (x_j, y_j) are concordant if the ranks for both elements agree, i.e. if both $x_i > x_j$ and $y_i > y_j$ or if both $x_i < x_j$ and $y_i < y_j$. They are said to be discordant if $x_i > x_j$ and $y_i < y_j$ or if $x_i < x_j$ and $y_i > y_j$. Another way of saying this is that the two points are concordant if $(y_j - y_i) / (x_j - x_i) > 0$ and disconcordant if $(y_j - y_i) / (x_j - x_i) < 0$. If $x_i = x_j$ or $y_i = y_j$ the pair is neither concordant nor disconcordant. If there are no ties then the Kendall correlation coefficient, τ , is calculated as (Davis and Chen, 2007):

$$\tau = \frac{N_c - N_d}{n(n - 1)/2} \tag{7.4}$$

Where N_c is the number of concordant rank pairs, N_d is the number of discordant pairs, n is the number of variables and the denominator $n(n - 1)/2$ represents the total possible number of pairs. In the case of ties, τ , can be defined as (Davis and Chen, 2007):

$$\tau = \frac{N_c - N_d}{N_c + N_d} \quad (7.5)$$

τ can vary between -1 and +1, a larger number of concordant pairs will tend the value towards +1 and a larger number of discordant pairs will tend the value towards -1.

- Chebyshev distance – This is the maximum distance between two vectors in any single dimension. It is said to be of most use when the difference between two points is best represented by individual dimension differences and not by considering all differences together. By its nature it is very sensitive to outliers.
- Bray Curtis Distance – This is also known as Bray Curtis dissimilarity, or the Sorenson distance. It is often used in ecology for quantifying the dissimilarity between populations. It can be calculated, for vectors \mathbf{x} and \mathbf{y} over k variables as:

$$d^{BCD} = \frac{\sum_k |x_k - y_k|}{\sum_k (x_k + y_k)} \quad (7.6)$$

7.2 Assigning Unknown Samples to Pre-Existing Groups

All of the subsequent methods to be described in this chapter require group association to have been defined prior to the development of models and/or classification of new samples. This group association can occur via clustering methods or manually.

7.2.1 Match Indexing

There are several methods that can be used to assign an unknown sample, based on its NIR spectrum, to a pre-existing subgroup. One of these methods involves the calculation of a **match index** between two spectra. This is considered to be the cosine of the angle between the multidimensional points representing the spectrum of the unknown sample (\mathbf{x}) and the group mean spectrum ($\bar{\mathbf{g}}$), and it is calculated as (Mark, 2001a):

$$MI = \frac{\sum_{i=1}^k x_i \bar{g}_i}{\sqrt{\sum_{i=1}^k x_i^2 \sum_{i=1}^k \bar{g}_i^2}} \quad (7.7)$$

In other words (if the spectra are mean centred) it is the correlation between these two spectra and it can take values between -1 and +1 with a sample identified as belonging to a group for which the MI is closest to +1.

Another method involves the **use of direct spectral matching** and it calculates the value of the discrimination criterion (Mark, 2001a):

$$\sum_{i=1}^k (x_i - y_i)^2 \quad (7.8)$$

Where \mathbf{x} and \mathbf{y} represent the unknown and known (library) spectra. This is equivalent to calculating the Euclidean distance between the unknown and each of the known samples, and then the subgroup of whichever sample the unknown sample is nearest to will be the subgroup to which this sample will be assigned. This method will be highly susceptible to physical variations in the spectra meaning that raw spectra should not be used and some spectral pretreatment will be necessary to remove particle size effects. Also, being based on Euclidean distances it has the disadvantages, in some situations, when compared with the Mahalanobis distance (MD) that were outlined in Section 6.4.3.3.

A similar method to direct spectral matching is known as **K-nearest neighbours (KNN)**. In this case the K nearest spectra (typical values of K can be three or five, and the optimal number can be determined through a test set) are selected and the unknown sample is classified according to the group which has the largest number of those neighbours. If there is a tie the class of the nearest neighbour can be used or the actual distances values can be incorporated into the calculation. The latter case is what is applied

in **weighted KNN**, whereby the ability of a sample to influence the classification of the unknown sample (i.e. its weight) is a function of its distance from the sample. (Balabin and Safieva, 2011). Both KNN and direct spectral matching can use PCA scores instead of the manifest variables.

It should also be noted that the chance of an unknown sample being classified to a group as a result of these spectral matching/comparison tests will depend partially on the number of samples that are in that group.

7.2.2 SIMCA

SIMCA (Soft Independent Modelling of Class Analogy (Wold, 1976)) is a supervised classification process whereby a PCA is conducted on each group. Each PCA model should describe its respective subgroup well, meaning that outliers should be checked for within each PCA model and the optimum number of components will most likely differ between different subgroups. This number should be determined with cross validation methods.

These cross validation methods can be used to determine, for differing number of components, the residuals for each sample that has been excluded and then these residuals summed (to give s) and compared between different models with that giving the lowest value for s being selected as the most appropriate model (as discussed concerning the PRESS statistic in Section 6.8). The sample residuals can be calculated from the sample scores on the non-retained eigenvectors. For example, if a PCA model for group G has a total of r eigenvectors but only r^* are selected to be used in the cross-validation step then the total residual, to be compared between different models, can be determined by (Massart et al., 1998b):

$$s = \sqrt{\sum_{i=1}^n \sum_{j=r^*+1}^r t_{ij}^2 / (r - r^*)(n - r^* - 1)} \quad (7.9)$$

Following calculation of s , a confidence limit for the classification of unknown samples based on their residuals is determined by (Massart et al., 1998b):

$$s_{crit} = \sqrt{F_{crit} s} \quad (7.10)$$

with F_{crit} being the tabulated one-side F -value for $(r - r^*)$ and $(r - r^*)(n - r - 1)$ degrees of freedom. The value of s_{crit} changes according to the value set for the confidence level, α . It is assumed that the data are normally distributed and that the setting of α (e.g. 5%) will mean that $\alpha\%$ of the objects belonging to the class will be considered as not belonging to it (Massart et al., 1998b).

So, for the classification of unknown samples the residuals, i.e. distance to the PC model, are determined for each group, and if $s_{sample} < s_{crit}$ for one or more groups then the sample will potentially be in those groups (depending also on the leverage value, if that is to be considered).

The leverage value (distance from the centre of the model for a model-projected sample) is considered because if only the residual were to be used as a boundary layer/plane in multidimensional space a boundary would exist in directions not covered by the axes of the PC but there would be no boundary along the PCs themselves. This is not an ideal situation, since unknown samples that are close to the model, but extremely far from the centre (high leverage) would be still assigned to the group even though they would be dissimilar from most of the samples in the calibration set. To solve for this a critical limit is also put on the leverage. This is done by treating each PC separately and setting maximum and minimum scores for these; hence (Massart et al., 1998b):

$$t_{max} = \max(t_k) + 0.5s_1 \quad (7.11)$$

$$t_{min} = \min(t_k) - 0.5s_1 \quad (7.12)$$

Where $\max(t_k)$ is the largest of the scores for the samples in the calibration set, $\min(t_k)$ is the smallest of the scores in the calibration set, and s_1 is the standard deviation of the scores along that PC.

SIMCA in The Unscrambler X:

The Unscrambler X is able to use SIMCA and has a variety of statistics and plots that can help understand how the unclassified samples are assigned to different groups.

The sample-to-model distance (**Si**) is calculated for the sample to be classified and compared to the overall variation of the class (i.e. s in Equation (7.10), although this statistic is called s_0 in Unscrambler). As in Equation (7.10), a critical limit is used in this comparison. It is calculated as follows (CAMO, 2010):

$$s_{MAX} = s_0 \sqrt{F_{crit(1,l,\alpha)}} \quad (7.13)$$

Where l is the total number of sample (either in the calibration set with cross validation or else in the validation set with test-set validation), and α can vary between 0.001 and 0.25

The sample leverage (**Hi**) is also determined, and a sample is considered as not part of the group under the following situation (CAMO, 2010):

$$\frac{H_{i_{sample}}}{\bar{h}_{model}} > Critical\ Value \quad (7.14)$$

Where \bar{h}_{model} is the average leverage value of samples in the calibration set and the critical value is a user defined number (set to 3.0 as default in the program). This means that the critical leverage limit for a group is three times the average leverage.

A class membership table can now be made and each sample is assigned to none, one, or more than one of the groups, depending on how the S_i and H_i values compare to the critical limits. The probability of assigning an unknown sample to more than one group is likely to fall as the quality of the PCA of each class improves, such as through the addition of more variables or more calibration samples.

Using the class membership table, statistics concerning the sensitivity and specificity of the classification model can be calculated. Sensitivity is the proportion of samples belonging to a certain category that were correctly identified by the model while specificity is the proportion of samples not belonging to a class that are correctly identified (as not belonging to that class) by the model (Sinelli et al., 2010). The percentage of Type 1 and Type 2 errors can also be calculated. A Type 1 error is where a model rejects a sample that is part of the group, and a Type 2 error is when a model accepts a sample as being a member of a group when in fact it is not (Adedipe et al., 2008). It can be seen that a Type 2 error is calculated as 100% - Sensitivity, and a Type 1 error is calculated as 100% - specificity. There can often be conflict between obtaining high values for both the sensitivity and specificity statistics.

7.2.3 Partial Least Squares Discriminant Analysis (PLS-DA)

PLS-Discriminant Analysis is simply conventional PLS that uses the spectral data and a new **Y** matrix containing indicator variables. The method assumes that a sample must belong to one of the classes. It uses binary discriminant analysis (BDA) to determine whether a sample belongs to a class or not. BDA involves the use of an indicator variable corresponding to each class. This variable can either have a value of 1 (meaning that the sample is part of the class) or 0 (the sample is a non-member of the class). Alternatively, assignment values of +1 (member) and -1 (non-member) can be used and this is often preferable since it enables easier visual identification and assignment of samples in predicted-Y plots. The resulting matrix of indicator variables (with *C* columns, corresponding to the number of classes) is used as the **Y** block and the spectral values as the **X** block (predictor variables) in PLS. This allows a model to be built and validated.

It can therefore be seen that, in contrast to SIMCA which only models the within-class variability, PLS-DA models the between class variability. This is an important difference because the variables that will receive high loadings in SIMCA may not necessarily be the best variables for class discrimination.

The rules for class assignment of a sample to a particular indicator variable is based on the predicted *Y* (Y_{pred}) and are as follows (note the critical value will be 0.5 if 0 and 1 are used as the binary values in the indicator variables):

- $Y_{pred} > 0$ the sample is considered to be a member.
- $Y_{pred} < 0$ the sample is considered to be a non-member.

PLS-DA should then be tested on unclassified samples. By enabling the full prediction option in Unscrambler X a range of diagnostic tools are available as discussed in Section 6.6. In particular the deviation in prediction can be used. Considering this deviation, the following rules can be applied in PLS-DA for the assignment of a sample to each indicator variable:

- Samples that have $Y_{pred} > 0$ and a deviation that does not cross the 0 line are predicted members.
- Samples that have $Y_{pred} < 0$ and a deviation that does not cross the 0 line are predicted non-members.

- Samples with a deviation that crosses the zero line cannot be safely classified.

An alternative to using the predicted Y values for classification in PLS-DA would be to project the samples to be predicted onto the PLS model developed on the calibration samples and observe the locations of these samples on Scores plots. If the prediction-samples lie in regions of factor space associated with one particular class, then they can be assigned to that class. Linear discriminant analysis (LDA) on these PLS-DS scores can be used for this purpose.

7.2.4 Bayes' Rule and Discriminant Analysis

The key theorem behind Bayes' rule is that the probability of an event A (e.g. a sample belonging to a particular group) given an event B (e.g. the provision of spectral data for a sample) depends not only on the relationship between A and B but also on the simple probability (**marginal or prior probability**) of each event (i.e. that a sample belongs to a particular subgroup when B is not considered).

The rule assumes that the probability distributions within all the groups are known and that prior probabilities (π_j) for the different groups are given, with the sum of these probabilities being 100%. If these prior probabilities are not known it is assumed that they are equal to $1/G$ (with G being the number of subgroups). Then Bayes formula is used to calculate the **posterior group probabilities** for each new sample. These probabilities also sum to 100% over all groups and are the chance that an unknown sample will belong to a particular group once that sample's spectral data have been considered (Naes et al., 2007).

A summary is that to get the posterior probability distribution you should multiply the prior probability distribution function by the likelihood function (which is the probability of event A given event B) and then normalise (so that the integral of the function would be one, making it a probability distribution).

Linear Discriminant Analysis (LDA) is the simplest method based on Bayes' formula. It assumes that the probability distributions within all the groups are distributed normally and that the covariances matrices of all the groups are equal, meaning that the only differences between groups will be in the position of their centres. The pooled covariance matrix across all groups will therefore be a weighted average of the individual covariance matrices (Naes et al., 2007).

$$\hat{\Sigma} = (N - G)^{-1} \sum_{j=1}^G (N_j - 1) \hat{\Sigma}_j \quad (7.15)$$

Where N_j is the number of samples in each sub group and G is the number of subgroups.

Naes *et al.* (2007) relate LDA to Bayes' rule with the following formula:

$$L_j = (\mathbf{x} - \bar{\mathbf{x}}_j)^t \hat{\Sigma}^{-1} (\mathbf{x} - \bar{\mathbf{x}}_j) - 2 \log_e \pi_j \quad (7.16)$$

Where $\bar{\mathbf{x}}_j$ is the empirical mean of each group.

The unknown sample with vector \mathbf{x} will be assigned to the group which has the smallest value of L_j . The difference $L_j - L_k$ (for groups j and k) can be reduced to a linear function of \mathbf{x} ; hence the name LDA (Naes *et al.*, 2007). If the prior probabilities are identical for all groups then the log term in Equation (7.16) will be the same for all groups, meaning that the criterion is reduced to the squared Mahalannobis distance (i.e. leverage, Section 6.4.3.5).

As can be seen in the formulas above, LDA uses the mean of each group as an important discriminatory criterion. Hence it will not perform well where the discriminatory information is not in the mean, but is in the variance of the data (CAMO, 2011, Naes *et al.*, 2007).

Quadratic Discriminant Analysis (QDA):

This method does not assume that the covariance matrices of all the groups are identical. Bayes' rule will then give the following allocation criterion - allocate an unknown sample with measurement vector \mathbf{x} to the group with the smallest value of (Naes *et al.*, 2007):

$$L_j = (\mathbf{x} - \bar{\mathbf{x}}_j)^t \hat{\Sigma}_j^{-1} (\mathbf{x} - \bar{\mathbf{x}}_j) + \log_e |\hat{\Sigma}_j| - 2 \log_e \pi_j \quad (7.17)$$

Where $|\hat{\Sigma}_j|$ is the determinant of $\hat{\Sigma}_j$.

In contrast to LDA, which separates groups according to lines, quadratic discriminant analysis uses curves as a means of separation.

Limitations with LDA for classification

If there are less training samples than variables (as will nearly always be the case in NIRS) it will not be possible to calculate the inverse of the covariance matrices due to the collinearity of the columns of the

matrix. Even if it would be possible to calculate the inverse it may be extremely unstable if standard methods are employed. This can be understood by considering that the inverse covariance matrix of the variability within the groups can be written as (Naes et al., 2007):

$$\hat{\Sigma}_j^{-1} = \sum_{a=1}^K \hat{\mathbf{p}}_{ja} \hat{\mathbf{p}}_{ja}^t / \hat{\lambda}_{ja} \quad (7.18)$$

Where the $\hat{\mathbf{p}}_{ja}$'s are the eigenvectors and the $\hat{\lambda}_{ja}$'s are the eigenvalues of $\hat{\Sigma}_j$. If some of the eigenvalues are very small then near collinearity will result and even a small inaccuracy can cause a dramatic effect on the inverse covariance matrix which may result in unstable classification rules.

Therefore it is typical for the dataset to be simplified and made orthogonal via the formation of a (limited component) PC model with the scores being used as input for the LDA. Clearly the number of PCs will be significantly less than the number of original variables but it may not always be the case that the discriminatory power of these PCs decreases with increasing PC number, since different rules are used to create PCs compared with those used to create canonical variates.

Use of LDA/QDA in Unscrambler X

This program has an option for the LDA/QDA of a data set. First a model needs to be built and this requires that each sample is already labelled as being within a subgroup. It gives options for the prior probabilities for each group (these can be set as equal or as equivalent to the relative size of the group in the whole training set). The program can also use either Linear or Quadratic discriminant analysis methods. The number of samples in each subgroup must be greater than the number of variables and the program allows a PCA to be done prior to the LDA in order to reduce the dataset to a (user specified) number of latent variables. The results of the LDA are a **Prediction Matrix**, which contains the classification scores for each class for each sample along with a predicted class for that sample (the class with the highest score will be selected), and a **confusion matrix**, which summarises the predicted classes as compared with the actual classes. In this matrix samples along the diagonal will have been placed in the right class whereas those off the diagonal are in the wrong class.

An alternative to carrying out an LDA on the PC scores of the first A components is to carry out PLS-DA (as described previously) and then use the scores from this in an LDA/QDA.

7.3 Comparisons Between Classification Methods

Not many scientific papers could be found regarding the comparison of two or more classification methods for feedstocks specifically related to biorefining. Most studies available in the literature concern the discrimination of consumer products, such as food and transport fuels. However, since the principles are the same these papers will be discussed below.

Balabin and Safieva (2011) compared, amongst other classification methods, the use of KNN, PLS-DA and regularised discriminant analysis (RDA, which the authors called a “compromise” between LDA and QDA) to discriminate, using NIR spectra, between 10 different biodiesels (each class being associated with a different vegetable oil, e.g. sunflower, coconut, palm etc.). Various forms of spectral pretreatments were tried and a second order Savitzky Golay derivative (see Section 8.2.1) followed by mean centring followed by Othogonal Signal Correction was used in most cases. The full set of 403 samples was subdivided into a calibration set (283 samples), a fitting set (50 samples), and a test set (70 samples). This was repeated five times and the average values of the five tests reported. The total error, on the test set, of the RDA model was 14.6%, while that for the PLS-DA model was 10.6% and that for KNN was 6.9%. The authors also carried out the same tests on different categories of gasoline and motor oils and the order of accuracy was in agreement in both cases.

Lau *et al.* (2009) compared LDA (using the PCA scores) and SIMCA in the discrimination between two species of *Radix puerariae* (*Pueraria lobata* (YG) and *Pueraria thomsonii* (FG)) using various spectral ranges of their NIR spectra. It was found that models based on the intensities from spectral regions among 1600–1800 nm and 2050–2450 nm were superior to the full spectral range (1100-2500nm) models and, in every instance, the successful prediction rate was superior for SIMCA compared against LDA, although both models performed well. The sensitivities and specificities in SIMCA were generally of a high quality, particularly for the FG model (sensitivity of 83.3% and a specificity of 100% in cross validation when using the 1600-1800 and 2050-2450 nm spectral range and the second derivative).

Cozzolino *et al.* (2006b) used spectra collected over the 400-2500 nm wavelength region to discriminate between the (minced) muscles of two breeds of pigs. PLS-DA, LDA (based on the PC scores) and SIMCA were used. PLS-DA correctly identified 87% of muscle type A and 78% of type B, for LDA the results were 87% and 67% and for SIMCA 100% and 57%.

McElhinney *et al.* (1999) attempted to discriminate, using spectra collected over the 400-2500nm wavelength range, between five types of meat (pork, lamb, beef, chicken and turkey) using, among other techniques, PLS-DA, SIMCA, and KNN. It was found that there was a particular difficulty distinguishing between chicken and turkey meats and two models were developed, one targeting the discrimination of five meats and the other the discrimination of four meats (with turkey and chicken classified under the same group). It was found that PLS-DA and KNN provided similar predictive accuracies and that these were mostly superior to SIMCA. In all cases the discriminations were better under a four-meat model than a five-meat model.

A later paper (Arnalds *et al.*, 2004) attempted to discriminate the same meats but using a hierarchical approach to conduct the classification whereby meat was first discriminated for by its colour (red or white), then by livestock type (pork and poultry for the white meat and beef and lamb for the red meat). There was a further subdivision for the poultry classification between chicken and turkey. Such an approach allows specific, targeted, conditions to be set for each level of the hierarchy (e.g. wavelength ranges, spectral pretreatments, classification algorithms) and also makes classification easier since it will only involve a binary discriminant analysis at each level. LDA (using PC scores) was compared against SIMCA and it was found that SIMCA was needed for the most difficult of the splits (chicken versus turkey). The performance of this technique was superior to that of the earlier paper.

From these examples, there does not appear to be a clear winner regarding the different classification methodologies and they should be applied and evaluated according to the particular application of interest. However it need not always be a case of one discrimination method or another. They can be used together, either in a hierarchical approach, as mentioned in the paper by Arnalds *et al.* (2004), or by the use of PLS-DA scores in LDA/QDA.

It should be noted that there are other tools available for qualitative analysis, for example there is the use of Artificial Neural Networks (ANN). Since these tools were not available in the software used (The Unscrambler X and Vision), they were not evaluated in this Thesis.

8 Spectral Pre-Processing Techniques

There are numerous treatments that can be made to spectra prior to the development of calibration models. These are typically done for the following reasons:

- 1) In order to remove or reduce the effects that physical phenomena (such as variations in particle size) will have on spectral variability.
- 2) Resolve as separate peaks those absorbances that may be hidden as shoulders and minor fluctuations on standard absorbance spectra.
- 3) Reduce the complexity of spectral variability so that the properties of interest can be calibrated for with simpler chemometric models that rely on latent variables.

This Section will present some of the spectral-treatment methods that have been used or tested in the software available during the course of this research (i.e. “Vision” by Foss and “The Unscrambler” by Camo).

8.1 Scatter Correction Methods

The scattering of light is said to be a function of two properties (Naes et al., 2007):

- The number of light and surface interactions which will depend partly on the size and shape of particles.
- The actual differences in refractive indices of these particles and of their surroundings.

The effect that particle size has on the scattering of light is of key importance. It clearly has no influence on the actual chemical properties of the sample (since one sample of biomass can be comminuted to a wide range of particle sizes) but it does have a significant effect on the amount of light that is scattered from, and therefore the amount of light that can be absorbed by, biomass. There will be more light scattering and reflection for smaller particles than for larger ones. Kubelka-Munk theory says that light going through a homogenous material will either be scattered or absorbed and the Kubelka-Munk (K-M) function relates the “true” absorbance coefficient (K), the scattering coefficient (S) and reflectance (R),

where R is defined as the intensity of reflected light divided by the intensity of incident light (Workman, 2001):

$$\frac{K}{S} = (1 - R^2)/2R \quad (8.1)$$

Reflectance data can be transformed to the K-M scale but mostly NIR results are presented in absorbance, which (see Section 6.1 for more details) is classified as:

$$A = \log(1/R) \quad (8.2)$$

Both of these transforms can result in the scatter effect being multiplicative, meaning that if two spectra from samples of the same material are subject to differing scattering effects (as a result of differences in particle size) the difference in the spectra can be compensated by multiplying the measurement at each wavelength of one of the samples by the same constant (Naes et al., 2007). NIR data can also have an additive scatter component (i.e. the absorbance at each wavelength is adjusted by a constant, resulting in a uniform baseline shift across the wavelengths). This effect is explained as follows – while the theoretical models such as K-M and Beer-Lambert assume that all or a constant part of the reflected light is detected, many NIR instruments will only detect a fraction (1/c) of this reflected light, meaning that (Naes et al., 2007):

$$I_{detected} = 1/c \times I_{reflected} \quad (8.3)$$

Hence:

$$\begin{aligned} A_{detected} &= -\log(R_{detected}) = -\log(I_{detected}/I_{incident}) \\ &= \log c + \log(I_0/I_{reflected}) = c' + A \end{aligned} \quad (8.4)$$

Therefore, if this constant c^j , which is equal to $\log(c)$, is specific for the sample it will cause an additive baseline difference.

8.1.1 Multiplicative Scatter Correction (MSC)

MSC, which is sometimes referred to as multiplicative *signal* correction since it has been found in some cases to remove varying background spectra with non-scattering origins (Boysworth and Booksh, 2001),

can assume that scattering is both additive and multiplicative, or that it incorporates just one of these effects. Ideally the scattering profile in a spectrum could be determined from a plot of an ideal scatterer (which has no NIR absorbance) versus a given spectrum at each wavelength. However given the difficulty in finding such a scatterer, typically the mean spectrum of a sample set of similar spectra is used (Boysworth and Booksh, 2001). This means that, unlike many of the other spectral pretreatments that operate on a per-spectrum basis, MSC is a set-dependent treatment that depends on all the spectra in the calibration set.

Figure 8-1 (a) shows the intensity for each wavelength for the mean of 41 Miscanthus DS internode/stem spectra (blue line) versus two of those spectra and Figure 8-1 (b) focuses in on the region between 1100-2500nm. There are clear scattering differences between the two, with sample 165 (red line) scattering more, meaning a lower absorbance, and sample 2003 (green line) scattering less. This can also be shown by plotting the sample absorbance against the average absorbance as shown in Figure 8-1 (c) for the 400-2500nm region and Figure 8-1 (d) for the 1100-2500nm region. In MSC it is considered that each sample plot lies about a regression line and the difference between each sample and that fit line can be interpreted as the chemical signal with the best fit line giving the spectrum of scattering of the sample.

The MSC model for each spectrum is therefore (Naes et al., 2007):

$$x_{ik} = a_i + b_i \bar{x}_k + e_{ik} \quad (8.5)$$

Where i is the sample number and k the wavelength number. This equation shows that there is an additive effect (a_i) and a multiplicative effect (b_i). These constants are unknown and must be estimated via least squares for each sample using all or part of the spectral range. The e_{ik} term incorporates additional effects that cannot be modelled by the additive or multiplicative constants. The \bar{x}_k is the average over samples at the k th wavelength (i.e. the blue lines in Figure 8-1).

Once the constants a_i and b_i have been estimated the MSC transform for the spectrum of sample i can take place (Naes et al., 2007):

$$x_{ik}^{MSC} = (x_{ik} - \hat{a}_i) / \hat{b}_i \quad (8.6)$$

This transform can be applied to all wavelengths or to specific regions. The model can be modified to only represent additive or multiplicative effects by removing the b_i or the a_i constant, respectively. Also, the calculation of the constants need not take place over the whole spectral region; for instance if there is a spectral region that is considered to be less reflective of chemical composition then this could be used for the constant calculation and these constants then applied across the spectrum. This may help to make the calculation of the MSC transform less dependent on chemical compositions. When the MSC method is used on future samples, for example where calibration techniques that employ MSC as a spectral pretreatment are used for prediction purposes, the mean of the calibration set is used as the scatter standard.

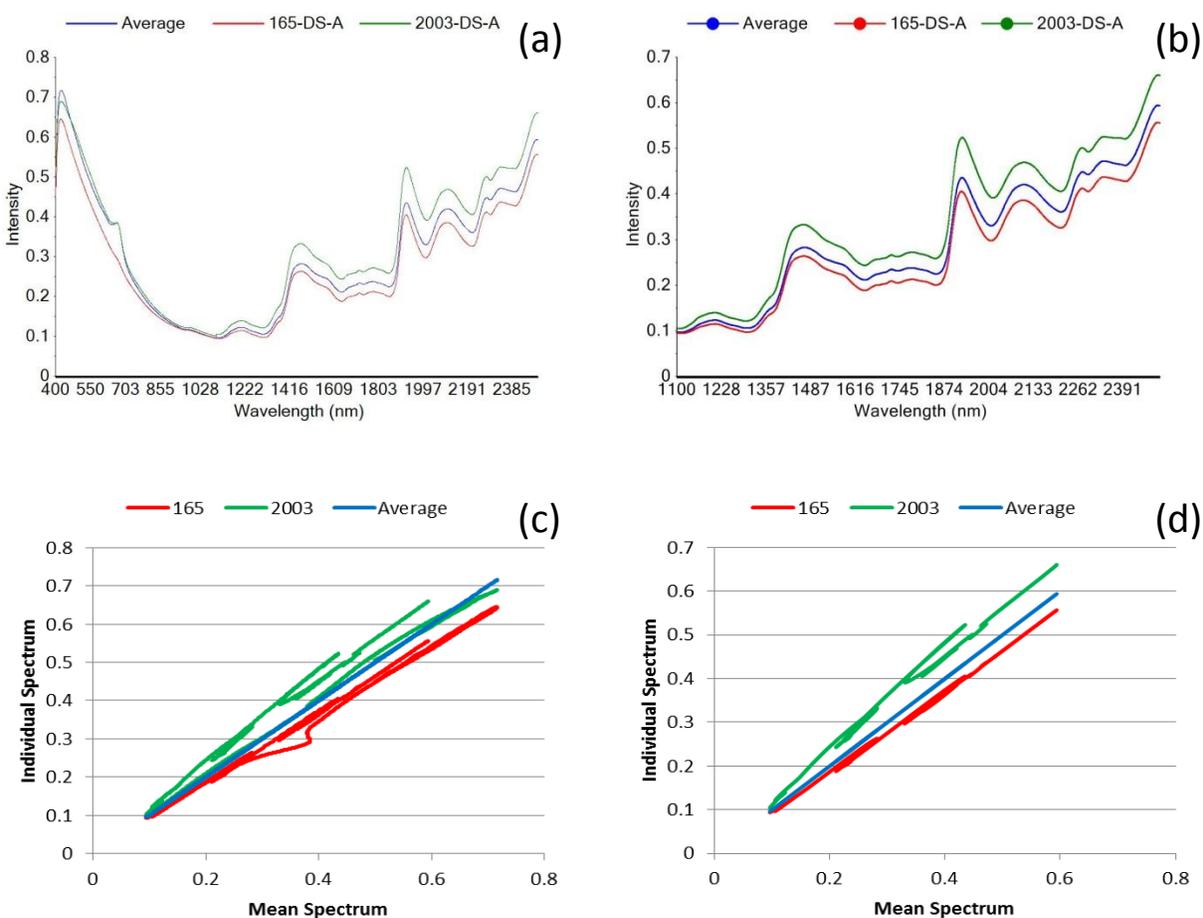


Figure 8-1: Illustrations of scattering effects. (a) A plot of the mean absorbance spectrum for 41 *Miscanthus* stem sections with plots of a higher than average scatterer (sample 165) and a lower than average scatterer (sample 2003) included; (b) the same plot but focussed on the 1100-2500 nm region; (c) a plot of the mean spectrum value versus the individual spectrum value for samples 165 and 2003 (the 1:1 line corresponding to the average is also included for comparison, blue line); (d) the same as (c) except only the values for the 1100-2500nm region is used.

It is considered that MSC is most effective and useful when the scattering effect is the dominating source of spectral variability. If, however, important chemical effects are more responsible for changes in the spectra then the e_{ik} term will have too much influence on the slope and intercept of the regression line and the MSC transform will therefore be too dependent on chemical information and will probably remove some of it, possibly to the detriment of the development of accurate calibrations. Hence, MSC may be of less use where very diverse spectra are all incorporated into the calibration set.

Also, given that the MSC uses the mean spectrum of a calibration set, it assumes that the spectra in this set are distributed normally and hence the mean spectrum is the most probable spectrum. The method attempts to correct spectra to behave like this mean spectrum as much as possible. Therefore the success of the MSC transform will depend on the actual distribution of samples within the calibration set and whether the calculated spectrum closely resembles the true mean spectrum (which will depend on the size of the calibration set). It should also ideally be the case that the sum of all the constituents that absorb light, in the spectral region used for the transform, should be constant (e.g. 100%). If, however this proportion varies, for instance in the case where there are significant changes in the (non-infrared-absorbing) ash composition of samples, then it may be impossible to differentiate between the sum of the constituents and the scatter since both may have a similar multiplicative effect on the spectrum (Naes et al., 2007).

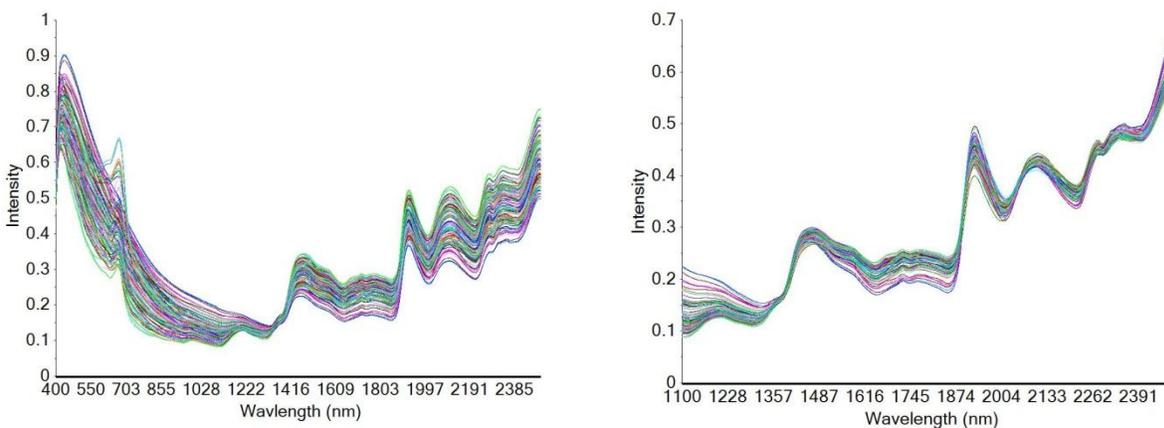


Figure 8-2: MSC-transformed spectra. (a) A plot of the MSC-transformed spectra of the 112 Miscanthus samples where the MSC transform used the full spectral range (400-2500 nm); (b) A plot of the MSC-transformed spectra of the 112 Miscanthus samples where the MSC transform used a smaller spectral range (1100-2500 nm).

Due to the complex and somewhat noisy spectra obtained in the visible regions, MSC is usually applied on the 1100-2500nm wavelength region in order to avoid these visible-region effects from influencing the calculation of the MSC constants in the NIR region. Figure 8-2 (a) shows the MSC transform (with the calculation of additive and multiplicative effects), which was calculated from and applied to the 400-2500 nm wavelength region, for 112 Miscanthus DS spectra, and Figure 8-2 (b) shows this transform where only the 1100-2500nm region is used for MSC calculation and transformation. It can be seen that Figure 8-2 (b) looks a lot clearer and more of the scatter effects appear to have been removed.

8.1.1.1 Extended MSC (EMSC)

The Unscrambler X software allows an extended version of MSC, known as E-MSC, to be used. This method is said to not only remove multiplicative and additive effects from spectra but also to allow a separation of physical light scattering effects from chemical light absorbance effects in spectra (Martens et al., 2003, CAMO, 2011). In this way it aims to improve on MSC in situations where strong chemical absorbances of the components of interest occur, since MSC, in simply removing additive/multiplicative effects, may in such instances confuse chemical interfering agents with physical light-scattering effects. Martens *et al.* (2003) explained the theory behind EMSC by relating it to the general principle of Beer's Law (Section 6.1). If Beer's law is obeyed then the (theoretical) chemical absorbance spectrum for sample i over a range of wavelength channels can be assumed to be a linear combination of the absorbance contributions of the various J constituents that make up the sample, i.e.:

$$abs_{i,chem} = \sum_{j=1}^J c_{ij} k_j^t \quad (8.7)$$

Where c_{ij} is the concentration of constituent j in sample i and k_j^t is the vector representing the absorbance (over K wavelengths) of that constituent. Under ideal conditions the measured absorbance spectrum for sample i , i.e. abs_i , should be equivalent to $abs_{i,chem}$. However there will also be physical scatter effects that will influence abs_i . The EMSC model attempts to model the influence this will have when compared against the ideal spectrum ($abs_{i,chem}$) using the following equation (Martens et al., 2003):

$$abs_i \sim a_i + b_i abs_{i,chem} + d_i \lambda + e_i \lambda^2 \quad (8.8)$$

Equation (8.8) includes the a_i and b_i constants from MSC, but it also includes two further constants, d_i and e_i , to compensate for the wavelength-dependent light scattering effects (linear and quadratic for $d_i \lambda$ and $e_i \lambda^2$, respectively) over the spectral region of interest.

If the coefficients were known theoretically, or could be estimated perfectly, then the EMSC correction shown in Equation (8.9) would remove the baseline variations as well as the wavelength dependent spectral effects that did not relate to the absorption constituents of interest, meaning that $abs_{i,corrected} \sim abs_{i,chem}$ (Martens et al., 2003).

$$abs_{i,corrected} = (abs_i - a_i - d_i \lambda - e_i \lambda^2) / b_i \quad (8.9)$$

In most cases, however, the coefficients are not known and they need to be estimated from the abs_i spectrum.

There are various options in the software for how the EMSC parameters are modelled and used. As with MSC, a_i and b_i can be used in the transform, or only one of them can. EMSC gives further options for parameters d_i and e_i ; these can be either modelled and subtracted in the transform, modelled only (the parameters will be calculated), or not used. A further option is whether the “squared spectrum” is modelled and subtracted, modelled only, or not used. Chemical effects are included in the squared spectrum (CAMO, 2011).

8.1.2 Standard Normal Variate (SNV)

SNV is similar to MSC except that it standardises each spectrum based only on the data for that spectrum, whereas MSC uses the mean spectrum of a set (or a reference spectrum). The SNV transform is shown below (Naes et al., 2007):

$$x_{ik}^{SNV} = (x_{ik} - m_i) / s_i \quad (8.10)$$

Where m_i is the mean of the K spectral measurements (wavelengths) for sample i , and s_i is the standard deviation of the same K measurements. The effect of SNV is that each spectrum is centred on zero on the vertical scale and varies roughly from -2 to +2. It is considered that this treatment can often be more suitable than MSC when the sample set covers diverse spectra or when the size of the set is small. Dhanoa *et al.* (1994) showed that MSC and SNV are linearly related and that the mean and standard deviation of the set mean spectrum, together with the correlation coefficient between each individual spectrum and the set-mean spectrum, are required to link the two transformations. SNV is often used in combination with Detrend (Section 8.3.1).

8.2 Derivative Spectra

The concept of a **first derivative** (1D) is that the line of the derivative would represent the slope at each point of the original spectrum, meaning that the 1D peaks would correspond to the points in the untreated spectrum with maximum slope and the 1D line would cross zero at a point of zero slope on the original (for example at a peak/trough). This means that the 1D will remove an additive baseline effect since an absorbance spectrum that is a shifted (upwards or downwards) version of another spectrum would still have the same slope and hence would be the same line on a 1D plot. The first derivative will not remove a multiplicative scatter effect, however.

The 1D can be calculated in several ways. The easiest way would be to take the derivative as the difference between the absorbance values at two points, so the 1D at wavelength w would be calculated as $y_w - y_{w-1}$ where y_w is the measured spectrum at wavelength w . However, this method will reduce the signal while also increasing the noise and so will result in a very noisy spectrum. This effect can be reduced by incorporating smoothing into the derivative. This usually takes the form that the absorbance values are taken as the average over several points. The region over which the average is taken is usually known as the **segment**. The distance between the two segments is known as the **gap**. The average absorbance values for the first segment (A) and the second segment (B) are calculated and the derivative is calculated as B minus A and is assigned to the data point in the middle of the gap. The process then shifts to the next wavelength in the spectrum and the process is repeated. The existence of segments and gaps means that derivatives cannot be calculated for a given number of data points at

the beginning and end of a spectrum; the size of this cut off will be equal at both sides and increase with the size of the segment and gap and the order of the derivative.

The **second derivative** (2D) will result from taking the first derivative of the 1D plot, i.e. it is the slope of the 1D plot. It can also be considered to represent a measure of the curvature in the original spectrum at each point (Naes et al., 2007). The 2D plot will maintain the band intensity and peak location from the original absorbance spectrum (although the absorption peaks will point down rather than up) and so can be easier to visually interpret than the 1D plot. It also helps to remove or reduce the overlap of absorption bands that is such a common effect in the NIR region and results in absorbance peaks appearing as relatively sharp valleys with lobes on either side. It is said that using a small gap and segment size (such as 2 nm for both) will produce 35-40 peaks in the spectrum of most agricultural products (Shenk et al., 2008).

While the 1D removes a baseline offset effect, the 2D will also remove a linear baseline. This can be understood by considering that a linear baseline can be represented by a first order equation, $(a \times w + c)$, where a is the slope, w is the wavelength, and c is the offset, and this effect will add to the function $f(w)$ (i.e. the spectrum). Taking the first derivative with respect to w will eliminate c but the slope would remain as a constant (a). However, taking the 2D will remove this effect (FOSS, 2006a).

A third order derivative (3D) is not often used to interpret spectra in NIRS since it will exhibit the same qualities as a 1D but will be even more difficult to interpret meaningfully (Shenk et al., 2008). Conversely, a fourth order derivative can be of more use for interpretation since peak locations will be as in the original spectrum. It is said that a 4D with a narrow gap (such as 4 nm) will generate between 60 and 70 apparent absorption peaks, that point upwards, for agricultural products (Shenk et al., 2008). Peak location may become clearer in this mode with little peak overlaps. It is considered that the treatment is more useful for resolving narrow absorption bands whereas broad peaks that have widths of 80 nm or more may be difficult to see (Shenk et al., 2008). There is also the matter of the formation of side lobe bands and other mathematical artefacts such as false valleys, and these events need to be considered when trying to interpret spectra and look for absorption bands. It should also be considered that 3D and 4D spectra will significantly reduce the signal in the transformed data.

The derivative treatments can be summarised in numerical form with each term representing a component of the treatment. They typically follow the order D, G, S1, S2 where D is the order of the derivative, G is the gap between points that is used to calculate the difference, and S1 and S2 are the

number of data points that are used to smooth the data, with S2 representing secondary smoothing. S2 is mostly not used, in which case it is set to 1 in the notation (Shenk et al., 2008). This Thesis will include a D in front of this notation to differentiate the Segment-Gap derivative from other derivative treatments such as Savitzky-Golay. Note that in some studies and software programs the order in the notation is instead D, S1, G, S2.

If even values were to be provided for the number of data points in the segment and gap then the averages and first derivative value would correspond to midpoints between actual data points. Therefore the number of data points for segments or gaps should be odd and they can be calculated from the following (FOSS, 2006a):

$$n = ODD \left[INT \left(\frac{x + 3}{2} \right) - 1 \right] \quad (8.11)$$

Where x is the declared size for the segment or gap, INT is a function that gets the nearest integer value, and ODD is a function that rounds an argument up to the nearest odd integer. Note that even numbers (in nm) can be provided for the segment and gap if these correspond to an odd number of data points (for example, a segment of 2 nm is equal to 1 data point for a spectrophotometer that has a resolution of 2 nm but it is equal to 4 data points for a spectrophotometer that has a resolution of 0.5nm, the resolution of the FOSS XDS used in this study).

The size of the gap and segment will affect the number of apparent absorption peaks and their resolution. There is a trade off in the choice of window size between the desire for noise reduction (smoothing) and signal enhancement, which will all occur with a larger segment and gap size, and the desire to avoid distortion of the curve which will occur if the segment is too large (Hopkins, 2001). Section 8.4 mentions some research papers where investigations were made regarding the most appropriate derivatives (in terms of their order, segment and gap size) with regard to developing calibration models for lignocellulosic materials. There does not appear, however, to be a universal rule regarding this and it appears that a degree of trial and error is necessary in order to achieve the best model.

The **Norris Gap** derivative is different from the Gap-Segment derivative that is described above in that it always has a segment size of one data point and therefore does not smooth the data prior to taking the derivative.

8.2.1 Savitzky-Golay Derivative (SG)

A different way to calculate derivative spectra is to use the Savitzky Golay (SG) treatment. In this process a segment size is chosen and a polynomial is fitted to these data points using least squares. For example, a seven point window chosen for the absorption at 1800 nm will have that wavelength as the centre data point and 3 data points on either side. If a quadratic curve is fitted to these points the equation would be (Naes et al., 2007):

$$y = \hat{a} + \hat{b}x + \hat{c}x^2 \quad (8.12)$$

Where x is wavelength, y is the spectral absorbance and the hats (\hat{a} , \hat{b} , \hat{c}) are the parameters estimated by least squares fit. The derivatives of this curve are then used to estimate the derivatives of the underlying spectrum. Therefore, for the quadratic above the 1D can then be calculated for any x and the 2D will be a constant for all x . If a quadratic curve is fitted to x and y data the coefficients will be linear combinations of the y s (absorbances) with weights that are functions of the x s (wavelengths) and the derivatives will be functions of the coefficients. Naes *et al.* (2007) provide an example of the estimation of the 1D at 928 nm for a seven point window:

$$(-3y_{922} - 2y_{924} - y_{926} + 0y_{928} + y_{930} + 2y_{932} + 3y_{934})/28 \quad (8.13)$$

Note that the absorbance, for instance y_{930} , will be the absorbance approximated by the fitting function at that wavelength and not the absorbance in the original spectrum. The estimated second derivative, with a seven point window, at 928 nm would be (Naes et al., 2007):

$$(5y_{922} + 0y_{924} - 3y_{926} - 4y_{928} - 3y_{930} + 0y_{932} + 5y_{934})/42 \quad (8.14)$$

This means that the weights (e.g. (-3 -2 -1 0 1 2 3)/28 for 1D) would only need to be computed once and could then be run along the spectrum to get the derivative at each wavelength (corresponding to the centre of the window). Therefore, for each point in the spectrum a curve is fitted and the derivative value would be a linear combination of the curve-approximated absorbances over the convolution window.

Higher order polynomials (Vision allows cubics, quartic and quintics whereas The Unscrambler allows up to a 12 order polynomial) can be used and these may allow a closer fit to the data points but they will need to have larger windows to achieve the same amount of noise correction. Higher order derivatives can also be used (up to 3D in Vision and up to 4D in the Unscrambler) but the degree of the smoothing polynomial must be greater than or equal to the order of the derivative. Vision allows the user to set the total number of points in the convolution window while The Unscrambler allows the user to specify the number of left points and the number of right points (if the symmetric kernel option is not employed). The polynomial order must be less than or equal to the sum of the left and right side points.

The polynomials can also be used just for smoothing which is done by applying a different filter with different weights to the absorbances over the convolution window.

It is said that SG will give a less distorted estimate of the “true” derivative than the Gap Segment derivative (Naes et al., 2007); however, that does not necessarily mean that it will provide a better spectral pre-treatment for subsequent calibration equation development. As with the Gap-Segment method the size of the window will be important, if it is too large it may smooth out too much of the important information, if it is too low the advantage of using a polynomial in smoothing/derivatising is lost. Again, trial and error may be necessary in order to find the best combination of convolution window size, polynomial order, and derivative order.

The notation used in this thesis for SG will be the following: $SG-D,P,SL,SR$, where D is the order of the derivative, P is the order of the fitting polynomial, SL are the number of datapoints used for smoothing on the left side and SR are the number of datapoints used for smoothing on the right side.

8.3 Other Transformations

8.3.1 Detrend (DT)

This is a technique that can be used to remove nonlinear trends in spectra. In particular it can be used to remove baseline offset, slope, and/or curvature from a spectrum. It does this by trying to derive a baseline function that is a least-squares fit of a polynomial to the spectrum of a sample. Hence, it works in a similar manner to SG, except the polynomial is derived for the whole spectrum and not just a small

window covering a limited number of data points. Vision allows a zero, first or second order polynomial to be used while the The Unscrambler program allows up to a quartic polynomial to be used.

A zero order polynomial will remove the baseline offset while a first order polynomial will remove the offset and slope and a second order polynomial will remove the offset, slope and parabolic curvature of the baseline. The procedure is often used together with SNV (and many research papers simply list SNV-DT as a spectral pretreatment) since SNV corrected data may still be affected by baseline curvature. Hence, SNV-DT will involve the use of a second order, or higher, polynomial in regression analysis where spectral values will be the y -variable, and the x -variable will be given by the corresponding wavelengths; hence (CAMO, 2011):

$$\hat{y}_{SNV} = a_{SNV} + b_{SNV}x + c_{SNV}x^2 + [d_{SNV}x^3 + e_{SNV}x^4] \quad (8.15)$$

In Equation (8.15) a , b , c (and d and e where third order and fourth order polynomials are used, respectively) are the regression coefficients. These form the basis for the baseline over all wavelengths. It is important to note that this process will remove baseline shift and curvature without overly changing the shape of the spectra, which can be an advantage over derivatives when trying to interpret the resulting spectra.

8.3.2 Mean Centring and Variance Scaling

Mean centring involves subtracting the mean spectrum of a data set from every sample in that set. In other words, the average absorbance value at each wavelength is subtracted from the spectrum to be centred. The effect of this treatment is that the data set is translated to the origin of multivariate space which will often allow a simpler and easier to interpret regression model (Boysworth and Booksh, 2001). It will help to reduce the number of factors needed in factorisation models (e.g. principal components analysis) since otherwise the first factor will usually be needed to describe the distance from the origin. It is selected as an automatic pretreatment in Vision and The Unscrambler for PCA/PCR and PLS model development methods. In the Unscrambler centring can also occur using the median or minimum for each variable.

Variance scaling is a method whereby the absorbance value at each wavelength for a spectrum is divided by the standard deviation of the absorbance values for that wavelength over all the spectra in

the calibration set. This means that the relative impact that each wavelength has in determining the parameters of the calibration model is equalised (Boysworth and Booksh, 2001). Thus the same weight is played to a wavelength that may have no useful data (i.e. whose variance is based primarily on noise) as to one which may be of vital importance for calibration development (the absorbance band of a vibrational overtone, for example). It therefore has limited use in NIR calibration but could be beneficial in situations where the signal at the important wavelength is very weak compared to the signal at other wavelengths (Boysworth and Booksh, 2001). The combination of mean centring with variance scaling is known as **autoscaling**. The Unscrambler software also allows the interquartile range and range to be used as scaling options.

8.4 Comparisons of Pretreatments in The Literature

Section 9 outlines the literature review on quantitative calibrations for lignocellulosic components of relevant feedstocks. However, this Section will describe any studies that involved detailed comparisons of a range of spectral pretreatments.

The most comprehensive study that was found in this review was by Sinnaeve *et al.* (1994). They undertook experiments to find the best combinations of spectral pretreatments, as well as to find whether local regression methods offered improved predictive ability over global regression methods. These experiments used three large forage databases (temperate grass-hay, tropical forages, and maize whole plant) that each had NIRS spectra (1300-2400nm, 1 mm particle size) and reference data (total protein, cellulose content, and enzymatic digestibility) for around 800 samples. These databases covered many different years, species, plant fractions, etc.

In total 540 different global calibrations were tested and compared. These varied according to the pre-treatment method employed (none, SNV, DT, SNV-DT, MSC, weighted MSC (WMSC)) and the derivatisation/smoothing applied to these pretreated spectra. It was found that large variations in SECV values were observed according to the spectral pre-treatment or to the use of derivatives. The relative differences (max–min/average) in SECV ranged from 9% for the cellulose in maize to 28% for the digestibility in grass-hay. It was generally found that the best pretreatments were SNV-DT, MSC or WMSC. Usually the worst results were obtained when no spectral pre-treatment was applied or when

DT was applied alone. Regarding derivatives, the two best treatments were D-1,5,5,1 or D-2,5,5,1. Increasing the gap over 5 points was found to increase the SECV.

For their local calibrations they assumed that the best treatment in the global calibration step would be the best in the local procedure. In these models the two unknown variables were the number of calibration samples to select (100, 200, 300, or 400) and the number of PLS factors (6, 9, 12, or 16). Hence, for each constituent and sample type 16 runs of the local regression were performed. It was found that the use of local calibration can improve the results compared with the best global equation with a relative gain in SEP being 5–11% of the SECV obtained with the best global calibration. It was also found that the SEP decreases with the number of PLS factors through a minimum and then tended to increase. This overfitting of the calibration equation occurred with a fewer number of factors for local regressions with the smaller sample sizes (e.g. with 100 samples, over 9 factors led to an increase in the SEP).

The authors concluded that trial and error remained the best technique to optimise universal calibrations. Also they noted that for samples that have low Neighbourhood H values (many neighbours), the accuracy is better for local than for global calibrations. The best results of the study are provided in the relevant Table in Appendix B for cellulose content.

Liu *et al.* (2010a) compared the RMSEP (5 samples in each prediction set) and RMSECV obtained, using PLS regression, for the glucose, xylose, lignin, and ash contents obtained via an analytical methodology similar to that in Section 11 (although extractives were not removed prior to hydrolysis), for 2 datasets (36 switchgrass samples and 35 corn stover samples), under scenarios where the following range of pretreatments were employed: MSC, SNV, 1st derivative, 2nd derivative, EMSC, the 1st derivative followed by SNV (1D+SNV), SNV followed by the 1st derivative (SNV+1D), the 2nd derivative followed by SNV (2D+SNV) and SNV followed by the 2nd derivative (SNV+2D). An FT-NIR device was used, scanning dry samples that had been ground down to a particle size of 420 microns or less.

The r^2 , RMSECVs and RMSEP's obtained are provided in Appendix B. It was found that models using some form of pretreatment were more accurate and required fewer latent variables than models developed on the raw spectra and that EMSC have the best predictive ability for most constituents in the two data sets.

It should be noted that the size of the calibration set was low (35 and 36 samples for each set) and size of the test set was very low (only 5 samples for each set). The authors also examined how spectral

pretreatments influenced the hierarchical clustering (Section 7.1.2) of samples whose spectra had been taken at various particle sizes. They used three samples (two different husks of corn cultivars and the internode of a switchgrass cultivar) and took their spectra in triplicate over three different particle sizes (0.841 mm, 0.420 mm, 0.250 mm). The resulting HCA (using Ward's distance) of the combined spectral data-set for all particle sizes using various spectral pretreatments combinations, or none, were compared with the HCA of the chemical composition data (which showed clustering together of the replicates of the three sample types). It was found that the cluster merging sequences for the spectral data set with no pretreatment were different from that of the compositional data and that particle size variation was a more important clustering criterion than chemical composition. All the spectral pretreatments improved the influence of the chemical composition on the clustering, and ANOVA tests comparing these pretreatments found that MSC successfully exposed the largest dissimilarity among the three biomass varieties (i.e. it gave the closest pattern to the ideal configuration), followed by [SNV+1D, 1D +SNV, MSC, SNV] then 1D and then [2D, 2D +SNV, SNV+2D].

Dardenne *et al.* (2000) made use of extensive NIR spectra and reference data-sets in order to evaluate various calibration methods (e.g. PLS and MLR) and spectral pretreatment methods. These sets included: fresh grass silage (1000 samples), wheat (2400 samples), whole plant maize (2250 samples), meats (650 samples), and two apple sets (775 and 380 samples). The calibrations were mostly not for lignocellulosic polymers but constituents such as protein and fat. They found that MLR gave the least accurate predictions, followed by PLS. Most importantly, it was found that good calibrations for high moisture products will not be possible if scatter correction pre-treatments (such as MSC, SNV and DT) are employed. It was considered that the influence of the water peak is too important to allow for a good correction to be made in the parts of the spectrum which are less influenced by water and where the most chemical information will be. The authors did find, however, that the scatter correction methods were appropriate for powders with variable particle size.

9 Literature Review of Quantitative NIRS Calibrations for Relevant Lignocellulosic Components

A wide literature review was conducted by the Author in order to determine the current state of the art for the determination, using NIRS calibrations, of the lignocellulosic components of interest in this study. This was done so that the quality of the calibrations developed in this research could be compared with those of other researchers and also so that an understanding could be reached concerning the best practices for deriving these.

As mentioned in Section 5.3, variations in moisture and sample heterogeneity can introduce complications in the development of calibrations. By far the vast majority of the papers encountered by the Author in this literature review aimed at minimising these variances as much as possible by preparing samples with small particle sizes and low moisture contents. As shall be seen in Section 11, the abbreviation DG is used for such samples. This stands for dry and ground meaning that the samples have been dried and the particle size distribution made to be relatively homogeneous via grinding/chipping methods. The methods for drying the samples (e.g. air-, oven-, freeze-drying etc.) varied between the different studies, but residual moisture contents were typically less than 10%. Indeed in some cases the samples were oven dried and then placed in a desiccator to cool prior to NIR analysis. This would reduce the spectral influence from water even further. Such a procedure may facilitate simpler calibration equation development but the speed of analysis of unknown samples, the great advantage of NIR as a primary analytical tool, would be lost under practices whereby oven drying would be necessary any time a sample is exposed to the humidity of the air.

As well as the drying process, another laborious element of sample preparation is the comminution of samples to a homogeneous particle size. As can be seen in Section 11, this is a lengthy procedure that can consist of several stages and take up much of the analysts' time. Such steps are unavoidable where reference analytical methods are to be employed but, once suitable NIRS calibrations are developed, it would be preferable for these to be simplified or even eliminated if possible. Hence, another class of NIR calibrations is termed DU, standing for "dry and unground" samples. In this case samples are still dry but are of a heterogeneous particle size. These "particles" could be the samples as collected or after some simple comminution process (e.g. chipping) necessary to prepare the sample in a form that could be

presented to the cell of the NIR instrument. In all cases, however, the mean particle size of the sample is significantly greater than in the DG category and the particle size distribution is also much wider.

The final category to be discussed with relevance to this literature review is WU, standing for “wet and unground” samples. The simple difference between this and the DU category is that WU samples have not been dried prior to the collection of their spectra by the NIR system. Based on all of the theory presented so far, this category should be the most difficult for which to develop accurate NIRS quantitative calibrations.

Discussions of previous studies that fall under each of these categories will now be presented. Throughout this discussion references will be made to Tables containing performance indicators (e.g. RMSEP, r^2 , RER, RPD etc.) of these studies. These Tables are presented in Appendix B, and group results from various papers according to the component and category (e.g. DG, DU, WU) of interest. For example, there is a Table with performance indicators of calibrations for cellulose content and within that Table the DG, DU, and WU categories are separated. This allows for easy comparison with the results presented by the Author in subsequent chapters.

9.1 Studies on DG Samples – Dry and of a Homogeneous Particle Size

One of the earliest and most comprehensive studies on the application of NIRS to the analysis of the lignocellulosic components of (dry and ground) biomass feedstocks for combustion/biorefining was by Sanderson *et al.* (1996). That paper included contributions from researchers at the National Renewable Energy Laboratory (NREL) at Boulder, Colorado. NREL has been very active in the use of NIRS for biomass analysis since the paper was published. The study involved the analysis of 121 samples of various biomass feedstocks including switchgrass (56 samples), corn stover (13 samples), poplar (18 samples) and lespedeza (18 samples). Some biomass feedstocks had only a handful of samples in the experiment, e.g. sugarcane bagasse (4 samples) and eucalyptus (3 samples). The sample set was therefore quite diverse, covering woody and grassy materials. The samples were air dried and milled to pass a 1 mm screen prior to NIRS analysis that was carried out in the region 1100-2500 nm. A total of 101 samples were used for the calibration set with the remaining 20 forming the validation set. Partial least squares was used for regression (with cross validation) and the data were scatter-corrected with SNV-DT.

Calibrations were attempted for ethanol soluble extractives, ash, lignin, uronic acids, arabinose, xylose, mannose, galactose, glucose, carbon, hydrogen, nitrogen and oxygen. The results are presented in the corresponding Tables in Appendix B. It was found that some samples (those of lespedexa, corn stover and bagasse) had a large baseline shift in the 1100-1400nm region, which could have been indicative of particle size variations. Calibrations were attempted where this wavelength range was excluded. These calibrations resulted in improved predictive abilities for extractives, lignin, xylose, glucose, and hydrogen, but calibration errors increased for ash, uronic acids, mannose, carbon, and oxygen.

A more recent paper from NREL (Hames et al., 2003), gives data for the predictive ability (SECV) of a model focused on corn stover samples. The calibration set of 47 samples included aged stover samples and hand-separated anatomic fractions. Some of the SECVs listed in Appendix B include 1.45%, 0.95%, 1.1%, and 0.19%, for glucan, xylan, lignin, and uronic acids, respectively.

The Author noticed in the literature review that studies on woods were particularly abundant. Schimleck *at al.* (1997) looked to examine the possibilities for quantitative NIRS measurement of the polysaccharides of the woods of plantation-grown *Eucalyptus (E.) globulus* and *E. nitens* trees. Wood meals were obtained by milling solid wood samples and NIR spectra were collected over the range 1100-2500 nm. These spectra were then transformed to the second derivative before PLS calibration. The *E. globulus* samples were analysed for a number of constituents including acetyl content, glucan, xylan, lignin and hot water extractives. For *E. nitens*, a reduced number of analyses were carried out including analyses for glucan and xylan. PLS models were developed for these constituents for each tree type. However, it should be noted that the data sets were extremely small (only 11 samples of *E. globulus* and 21 samples of *E. nitens*). The number of samples of *E. globulus* were so small that no validation was carried out and all 11 samples were used in the calibration set. For the *E. nitens* samples, 16 were used in the calibration set and 5 in the validation set. Up to 4 factors were used in PLS regression. The results are included in the relevant Tables in Appendix B. There are relatively few papers where acetyl content has been measured. The results from the Schimleck *at al.* (1997) study for this component ($r^2=0.99$, SEC=0.04% using 4 PLS factors) initially appear promising, however, given that there are 4 PLS factors and only 11 samples and no test set, it is hard to trust the results. For the *E. nitens* data, which did include a test set, there were significant differences between the SECs and SEPs, again reducing the confidence in the calibrations that were developed.

This paper was followed up by Raymond and Schimleck (2002). Here a larger number of samples were used to develop calibration equations for the cellulose content of *Eucalyptus globulus* Labill sampled

over three sites. The NIR spectra were measured in diffuse reflectance mode over the spectral range 1110-2500 nm. The spectra were converted to the second derivative using a segment width of 10 nm and a gap width of 20 nm. For cellulose content analysis 40 samples from each of the three sites were selected, based on their NIR predicted pulp yields (from a previous study) and this sample set was split into 30 samples for cellulose calibration and 10 samples for validation. The calibration data sets for all three sites were also combined to give 90 samples for the development of an across-site calibration, which was evaluated using the combined prediction set of 30 samples. Calibrations for each site were reasonable with r_{calib}^2 ranging from 0.82 to 0.94. SECs varied from 0.70 to 1.07%, whilst SEPs were slightly higher. The combined site calibration was also reasonable with an r_{calib}^2 value of 0.88 and both SEC and SEP less than 1%.

Another paper concerning the NIRS analysis of Eucalyptus was published by Downes *at al.* (2010b). They attempted an NIRS calibration for the cellulose content of the wood meal of 1212 samples of Eucalyptus, comprising numerous varieties from different countries (Australia, Uruguay and China). The initial calibration set comprised 1089 samples and the validation set 123 samples. An r_{pred}^2 of 0.72 and SEP of 0.92% were obtained from a calibration that used 11 PLS factors. Following this all calibration and prediction samples were combined into a single calibration data set that gave an r_{CV}^2 of 0.86 and an RMSECV of 0.80 with 9 PLS factors.

Another paper involving the derivation of calibration equations for wood meal was published by Hodge and Woodbridge (2010). They examined wood samples from seven pine species from five different countries (Brazil, Chile, Colombia, South Africa and the USA). The samples were ground to a 1 mm particle size and then oven dried at 50°C for 12 hours prior to NIR analysis over the wavelength range 400-2500 nm (with the range 1100-2500 nm used for PLS calibration). Wet chemical analysis was carried out for Klason lignin (KL) and cellulose. The authors examined the use of MSC and SNV as spectral pretreatments and found that MSC gave slightly better results for lignin models and SNV for cellulose models and, hence, these respective pretreatments were used in all subsequent models involving these parameters. These pretreated spectra were then all transformed to the second derivative of SG smoothed spectra (using seven points in the convolution window and a quadratic polynomial). Firstly, a global calibration model was developed with all samples included (457 for cellulose and 517 for lignin), with 67% of these in the calibration set and 33% in the validation set. For KL, an 11 factor calibration gave an r_{pred}^2 of 0.95 and an SEP of 0.44%. For cellulose a 10 factor calibration gave an r_{pred}^2 of 0.72 and a SEP of 1.1%. The authors also attempted calibrations for two subsets of the samples, one comprising

temperate species of pine and the other the tropical species of pine. These calibrations were then used to predict the opposite data set. The Temperate calibration models were based on only two species and, when extrapolated to the Tropical dataset demonstrated very poor predictive ability (e.g. $r_{pred}^2 = 0.44$ for lignin). For the extrapolation of the Tropical calibrations to the Temperate dataset, the fit statistics were much better (e.g. $r_{pred}^2 = 0.88$ for lignin). These values for the tropical calibrations compared well to the validation statistics for the global model. The authors also tried to develop species-specific models for two of the species (*Pinus (P.) taeda* and *P. tecunumanii*) with two thirds of the samples being used in the calibration set and the remainder in the validation set. It was found, however, that the global models gave predictions essentially as good as the species-specific models. The authors claimed that there was therefore no evidence to support the idea that predictions from species-specific calibration models will always be better than those from a robust global calibration model. This is an important note to take from this paper.

Towards the latter phase of his analytical work, the Author came across a series of papers that mirrored, to some degree, the methodology that he was using (for the development of calibrations for *Miscanthus*, see Section 15). However, these papers focussed on switchgrass and corn stover. The first of these papers, Ye *et al.* (2008), outlined the attempt to calibrate a FT-NIR device for glucose, xylose, galactose, arabinose, mannose, lignin (the sum of KL and acid soluble lignin (ASL)), and ash for corn stover samples. In a similar strategy to that outlined in Section 14.1, the plant samples had been separated according to their botanic fractions with 5 samples each of nodes, leaves, internodal piths, internodal rinds, sheaths and husks, giving (with 5 whole plant samples) a total of 35 samples. These had been air dried and milled to less than 420 microns and cooled to room temperature in a desiccator prior to NIR analysis over the spectral region 1000-2500 nm. The NREL protocol for lignocellulosic analysis was followed but, in what the Author considers to be a critical fault in the paper given the high extractives contents of corn-stover and the known interferences that such extractives can make to the analysis of lignocellulosics (see Section 3.4.2), the samples were not extracted prior to the hydrolysis stage. Various spectral pretreatments (multiplicative scatter correction (MSC), extended-MSC (EMSC), standard normal variate (SNV), first derivative (1D), 2D) were tried and it was found that EMSC gave the best predictive models. Despite the limitations, the results, as can be seen in the relevant Tables in Appendix B, were generally good with an RER (using the SECV) of 13.71 for lignin, for example. The authors also attempted to classify the 6 different anatomical fractions. It was seen from the PCA score plot of PC1 versus PC2 that the different botanical fractions could be differentiated and SIMCA was tested as an automatic

means for doing this. Sixty new FT-NIR spectra were collected, with 10 from each of the six botanical fractions, to validate the method and it was found that, even without data pretreatment, SIMCA correctly classified all of the new samples.

The authors followed this paper up two years later with a subsequent paper (Liu et al., 2010b) that included data for the analysis (by the same spectroscopic and wet-chemical methods) of the various anatomical fractions of six switchgrass cultivars (a total of 36 samples). Again, cross validation was used to test the model. Smaller relative errors were seen for the switchgrass model than for the corn stover model. Furthermore, the authors observed that, for a combined matrix containing the spectral data for switchgrass and corn stover, it was hard to distinguish between the switchgrass and corn-stover samples in a PCA plot of PC1 against PC2. This was also the case for a PCA conducted solely on the wet-chemical data. It was considered that a broad model containing both samples could be developed. From this model it was found that the correlations decreased slightly for some constituents compared to the two individual models but, regarding the prediction errors (RMSEPs), this broad model showed an improvement over the corn stover model. Furthermore, the RER statistic (using the SECV) improved for many constituents (as a result of the increased concentration range seen in this model), and rose to over the research quantification level (i.e. greater than 15, see Section 6.11.2 for more on this) for glucose and xylose. In order to further test this combined model, an independent validation set was used comprising 5 stover samples, 5 switchgrass samples, and 5 wheat samples. These results are also included in Appendix B. It was found that, for all constituents, the SEP was less than the SECV for the combined data set. Hence, an important lesson can be learned from this, and from the Hodge and Woodbridge (2010) study – models need not be focussed on too narrow a data-set (e.g. one species or one variety). Indeed, the predictive ability of expanded models can often be superior.

NIRS was also used for the analysis and discrimination of switchgrass samples by Labbé *et al.* (2008). They used NIR and FTIR equipment to attempt to differentiate between switchgrass ecotypes and varieties, and between varying nitrogen application rates. They also looked to use these instruments with PLS regression to predict the concentrations of non-structural carbohydrates (85% ethanol-soluble sugars and starch, liberated upon enzymatic hydrolysis). A total of 72 samples were used and these were ground to a particle size of less than 5 mm and dried at 105°C for 12 hours prior to spectroscopic analysis. The data were mean-centred and MSC was employed as a spectral pretreatment. Non-structural carbohydrates did not differ significantly between fertility treatments but did differ between cultivars. The best calibration was achieved with the NIR spectra with r_{CV}^2 (RMSECV) being 0.91 (0.683%)

for starch, 0.82 (0.311%) for sugar, 0.92 (0.895%) for total non-structural carbohydrates, and 0.95 (27.11 kg/ha) for nitrogen content.

Straw is another feedstock for which there are a number of NIRS studies available. Kong *et al.* (2005) attempted to develop calibrations for the ADF, NDF and ADL content of 207 rice stem samples. These samples had been oven dried at 65°C and ground to pass a 1 mm sieve prior to spectroscopic (1100-2498 nm) and reference analysis. Of the samples 136 were used in the calibration set and the remaining 71 in the validation set. Different derivative mathematical treatments (with SNVD) were applied for calibration. For ADF, a D-2,8,6,1 treatment gave the best results with r_{pred}^2 (SEP) of 0.959 (0.93%). For NDF a D-2,4,4,1 treatment gave the best results with r_{pred}^2 (SEP) of 0.775 (2.23%). For ADL a D-1,4,4,1 treatment gave the best results with r_{pred}^2 (SEP) of 0.847 (0.616%) although the 2,8,6,1 treatment gave very similar results with r_{pred}^2 (SEP) of 0.846 (0.617). The best results are included in the Tables in Appendix B.

Jin and Chen (2007) also analysed rice straw in a study that covered more chemical components and various anatomical fractions of the plant. They used a FT-NIR device to calibrate for the cellulose, hemicellulose, KL (all determined gravimetrically), total ash and acid insoluble ash contents of rice straw. The samples were divided up into leaves, sheaths, nodes, internodes, as well as kept as whole straw samples. The samples were air dried and with a particle size of between 425 and 250 microns. The calibration set comprised 34 samples and the validation set 9 samples. The first derivative, with a segment of 10 and a gap of 5, was used as a spectral pretreatment for all components. These results are also presented in Appendix B.

Huang *et al.* (2009) used NIRS to develop calibration equations for the C, H, and N contents (as well as the heating value) of a combined data set of Chinese rice and wheat straws. A total of 222 samples were used with one quarter in the validation set and the rest comprising the calibration set. The samples were all oven dried at 70°C for 24 hours and ground to pass a 1mm sieve. Various combinations of scatter correction methods and derivative treatments were employed but the authors claimed that the performance of all calibrations were similar. The results are presented in Appendix B and are generally of a high quality, with an RPD of 5.64 for carbon content, for example.

NIRS has also found applications in the analysis of decaying biomass; for example in the development of calibration equations for the composition of leaf litter. McTiernan *et al.* (2003) looked to study the changes in chemical composition that occurred during the decomposition (over a period of two years) of

Pinus sylvestris needles in coniferous forests, with 25 replicates per site. Once mass loss had been determined, the 25 replicates with the same duration of field exposure were pooled, and ground to a 1 mm particle size prior to NIRS analysis. PCA was used to select samples that appeared spectrally different (Mahalanobis distance greater than 0.6). These selected samples were put forward for reference analysis for C, N, NDF, ADF and ADL, and ash. It is not clear from the paper what number of samples were in the calibration and validation sets; however, a total of 117, 110, 112, 121 and 115 samples were analysed for NDF, ADF, ADL, N, and ash, respectively. It is also not clear from the paper what spectral ranges were used in NIRS analysis or what spectral pretreatments (if any) were used. The data that resulted from this study are presented in Appendix B.

Vavrova *et al.* (2008) also looked to use NIRS for the characterisation of plant litter. Litter from nine plant species, representing five groups of plant litter (monocotyledons, deciduous foliage, conifer foliage, wood, and moss) were collected at 8 peat-land sites to give a total of 78 samples. The litters were air-dried and ground prior to NIRS and reference analysis. Analyses were carried out for: extractives content (dichloromethane, acetone, ethanol, hot water), holocellulose (from extractive-free samples using the sodium chlorite method (Quaramby and Allen, 1989), C, N, KL, and copper oxidation products of lignin (e.g. vanillin, vanillic acid). Spectra were taken in the region 350-2500 nm but the 350-400nm part of the spectra was removed since it contained a lot of noise. PLS1 regression was used and various spectral pretreatment methods were evaluated. Models were tested for the entire spectrum (400-2500 nm) and for the NIR region only (780-2500 nm). It was found that models based only on the NIR part performed better. Generally, NIR spectra transformed by the second derivative combined with SG smoothing over 8 data points showed the best method for most parameters, although SNV-DT was best for holocellulose.

Ono *et al.* (2003) also examined the NIR spectra of leaves in various states of decay. They analysed a total of 132 samples of leaves, covering a variety of species, that were either fresh, fallen leaves, or organic material on the forest floor. The samples had been air dried and milled to pass through a 74 micron mesh prior to their spectroscopic (400-2500 nm) and reference analysis. A total of 88 samples were used in the calibration set and 44 in the validation set for total lignin, holocellulose and extractives (benzene:ethanol (2:1) solution) content. The resulting data, based on calibrations involving the second derivatives of the spectra, are provided in the respective Tables in Appendix B. The SEPs were relatively high, (5.0%, 3.5 % and 3.4% for lignin, holocellulose, and extractives, respectively) but the ranges in constituent values were also high, resulting in reasonable RERs and r^2 values. Interestingly, MLR was

used instead of latent variable methods, with either three or four wavelengths chosen, depending on the constituent. The authors also noticed that the absorbance in the region 1100 – 1400 nm increased continuously as the leaves decayed.

An interesting paper by Kelley *et al.* (2004b) compared the efficacy of NIRS and pyrolysis-molecular beam mass spectrometry (py-MBMS) to predict the concentrations of lignin, glucose, xylose, mannose, galactose, arabinose and rhamnose. This was carried out on a diverse array of agricultural samples including sugarcane bagasse, kenaf core, hemp, coconut core and palm trees. Some of the samples were subject to various extractions (e.g. NaOH, phenol) or chemical modifications (e.g. nitric acid) prior to NIRS/py-MBMS analysis. This provided a diverse range in the compositions of these materials. The NIRS analysis was carried out on dried and ground samples of these; however, only 23 samples were used due to a lack of material. Hence only cross validation statistics were calculated. In contrast 41 samples were analysed by py-MBMS. As can be seen in the relevant Tables in Appendix B, the RMSECVs are quite high for all components (e.g. 5.8% for xylose) and are significantly greater than the RMSECs. These poor results are probably due to the heterogeneous mix of materials and limited number of samples. The results for NIRS and py-MBMS were found to be similar

NIRS was also used for the quantitative (and qualitative) analysis of pretreated biomass by Krongtaew *et al.* with the results presented in two papers (Krongtaew *et al.*, 2010b, Krongtaew *et al.*, 2010a). They pretreated straws (wheat and oat) by means of mild acidic and alkaline pre-treatments in the presence and absence of hydrogen peroxide. FT-NIR analysis of milled samples (with a very small particle size of 80 μm) was carried out on these samples and calibrations developed for several parameters, including the residual KL content, with half of the samples used in the calibration set and the remainder in the validation set. Separate calibrations were developed for wheat and oat straws. The spectral range of 1449-1815 nm was selected for the calibration and the spectra were pre-treated with the SG second derivative. The results of the analysis, presented Appendix B, are impressive for samples that had undergone such extensive modifications, with a r_{CV}^2 (RMSEP) [# of PLS factors] of 0.94 (0.9%) [3] for wheat straw, and 0.96 (0.8%) (1) for oat straw. The low number of PLS factors used in these calibrations is interesting, it is possible that the pretreatment methods employed resulted in a reduction in the complexity of the lignocellulosic matrix, and hence fewer latent variables were required. It should be pointed out, however, that the particle size used for NIR analysis was extremely low meaning that the sample presented to the instrument was likely of an extremely high homogeneity. Nevertheless, the

paper demonstrates that effective NIR calibration for KL content of dry ground samples can be possible with a limited spectral range.

There has also been work at NREL on the NIR prediction of the composition of pretreated biomass samples (Hames et al., 2003). A calibration set comprised 96 samples produced using three bulk feedstocks of corn stover and various methods of dilute-acid pre-treatment. The SECVs, as shown in Appendix B, were 1.55%, 1.46%, and 1.51% for glucan, xylan, and total lignin, respectively.

9.2 Studies on DU Samples – Dry and of a Heterogeneous Particle Size

Much of the research involving the development of quantitative NIRS calibrations for dry samples of a heterogeneous particle size has focused on wood; usually either wood chips or wood cores/strips. For instance, Pole (2006) attempted to use NIRS to predict the extractives, KL, ASL, and cellulose content of (air dried) solid strips of *Eucalyptus globulus*. This was attempted in two ways; firstly it was seen if the calibration equations developed for ground wood of the same species in a previous study (Poke et al., 2004) could be used on the spectra obtained from these solid samples (9 validation samples used). Secondly, calibration equations were derived from the spectra of the solid wood samples with approximately 40 samples used for calibration development and 9 used for validation. In the first instance it was found that the ground wood calibrations were unsuitable for predicting extractives and acid soluble lignin and that the correlations for cellulose and total lignin were poor (0.63 and 0.54, respectively). It was considered that the reason for ground wood calibrations not working as well on solid wood was that the spectra from ground wood result from the cell wall polymers being analysed at varying angles relative to the incident radiation whereas in solid samples the polymers would be at a consistent incident angle within the same sample. In the second experiment reasonable calibration set statistics were obtained but the predictive ability of these equations, when tested on the validation set, were generally poorer. For example the r_{pred}^2 values were 0.87 for extractives, 0.12 for ASL, 0.79 for KL and 0.69 for cellulose.

Another DU paper involving wood samples by Jones *et al.* (2006) attempted to develop NIR calibrations for arabinan, galactan, glucan, mannan, xylan, cellulose, and hemicellulose (with the polysaccharides calculated from the monosaccharide compositions), ASL, and KL for 12.55 mm sections of *Pinus taeda*.

These sections were obtained by collecting breast height discs from the trees with radial strips being cut from these discs. Two or three 12.5 mm sections were selected on each strip and scanned with an NIRS instrument with a wavelength range of 1100-2500 nm. The strips were dried overnight before NIRS analysis so that their approximate moisture content was 7%. A total of 46 samples were used in an initial calibration that gave reasonable SECV values for all constituents except ASL, galactan and hemicellulose. The scans were then split into calibration (28 samples) and prediction sets (12 samples). The results of these analyses are outlined in Appendix B where it can be seen that the r_{pred}^2 statistics for all of the constituents were poor (the highest being 0.68). The poor results in this validation stage could be attributable to the small number of samples used for calibration/prediction. The authors also claimed that the prediction errors may be a consequence of the diverse origins of the samples in the test set and that further research with a greater number of samples would be needed.

Another wood DU paper (Kelley et al., 2004a) attempted to develop NIR calibration equations for KL, glucose, xylose, mannose, galactose and extractives for solid wood samples of loblolly pine trees. These samples were obtained from slices taken from tree disks and oven dried at 105°C prior to analysis (spectroscopic and reference). Spectra were taken over the full range of 500-2400 nm as well as a limited range of 650-1150 nm. No pre-processing techniques were used on the spectra. The calibration set comprised 45 samples and the test set 27 samples. PLS2 models were used to predict all the six wood components. Reasonable predictions for the test set, under the full spectral range (and using four PLS factors), were obtained for galactose (RMSEP = 1.0%) and extractives (RMSEP = 2.3%) but the other components all had correlation coefficients of less than 0.8. Interestingly, there was very little change in r_{pred}^2 , RMSEC, and in RMSEP with the reduced spectral range (using five PLS factors) with the exception of Klason lignin which saw the RMSEP increase from 1.0% to 1.4%. The poorer predictive ability of the limited spectral range model for KL was attributed to the very low intensity of CH overtone vibrations in this range.

Montes *et al.* (2009) attempted to design a new sample presentation system that would allow the NIRS analysis of large amounts of material of large particle sizes. It comprised of a box unit of dimensions 1 m long and 18 cm wide. Into this was placed the sample for analysis and a pressure system regulated the level of contact between the sample and a neutral glass interface. The box was then moved at a speed of 20 cm/s meaning that the full length was covered in 5 seconds. During this period the NIR reflectance spectra were taken from above, with a total of 30 spectra collected. Experiments were made with and without a glass interface, and it was found that such an interface was necessary to reduce spectral

variability. A total of 80 maize stover samples were collected from plots and manually cut into three particle sizes (8 cm, 6 cm and 4 cm) which were scanned in the box. After the first measurement the 4 cm particle size samples were cut to an average particle size of 0.5 cm and scanned again. These samples were then ground to a 1 mm particle size for laboratory NIR (using a commercial device and standard sample presentation system) and wet chemical analysis. Calibrations were attempted for N, ash, and NDF. These results are presented in Appendix B. It can be seen that the only reasonable results on the custom unit were obtained for ADF. For N and ash, the r_{CV}^2 were less than 0.5, even for the 0.5 cm particle sizes, in comparison to the standard NIR device where they were 0.92 and 0.74, respectively. This demonstrates underlying problems with the sample presentation unit under the current design. For NDF an r_{CV}^2 (SECV) of 0.79 (3.32) was obtained for the custom unit compared with 0.95 (1.68) for the laboratory unit. Regarding the analysis of particle sizes larger than 0.5 cm, in the custom unit, r_{CV}^2 decreased from 0.77 (0.5 cm) to 0.69 (4 cm) to 0.66 (6 cm) to 0.56 (8 cm). Few conclusions concerning the ability of NIRS to analyse samples with large particle sizes can be drawn from these results, however, given that the performance of the new unit was substantially less than a lab unit for 0.5 cm particle size samples. It should also be noted that the reference analysis was only carried out on subsamples of the 4 cm particle size fractions and that the coefficient of variation was generally quite poor (23.08% for nitrogen, 6.69% for NDF, and 18.78% for ash).

Another study involving a custom-built NIR unit is presented in a paper by Downes *et al.* (2010a). The Authors were interested to determine if Kraft pulp yield and cellulose content could be predicted from solid wood surfaces. The reason for doing this was that the authors wanted to investigate the variations in these properties along the radius of a tree section at a fine sampling scale of 1 mm. This scale would provide insufficient ground wood sample for chemical analysis or conventional NIR spectral analysis. Breast height discs were obtained from 12 trees and full diameter plinths were prepared from these discs. Each radius was scanned over 5-mm intervals with a Bruker MPA FT-NIR instrument fitted with a fibre-optic scanning attachment coupled to a linear sample transport system (a custom-built attachment that collected spectra at 1 mm intervals along the radial direction of the cores and provided spectra averaged over larger increments). The spectra (1000-2500 nm) were then averaged to 10 mm intervals. These 10 mm segments were then physically separated, providing a total of 91 sub-samples from the 12 cores. NIR spectra were also obtained from the radial longitudinal surface of each sub-sample with a hand-held device (939-1796 nm). Following this, each sub-sample was ground to woodmeal and spectra were collected with a laboratory NIR instrument and Kraft pulp yield and cellulose content were

predicted with existing calibration models. The resulting predictions were used as calibration data for the solid-wood spectra (10 mm averaged scans). It was found that the r_{CV}^2 (RMSECV) [#PLS factors used] for the cellulose content were 0.84 (0.97%) [4] and 0.87 (0.94%) [7] for the Bruker and handheld device, respectively, as outlined in Appendix B. The resultant calibrations were applied to spectra obtained from a single radius at 1 mm intervals to explore the potential to resolve sub-annual patterns of variation.

Nkansah *et al.* (2010) used NIRS to develop calibration equations for ash, extractives, total lignin, KL, holocellulose (estimated as the difference between the oven dried weight of the wood and the sum of the other chemical components), and bulk density, for yellow poplar samples. These were obtained from blocks (of dimensions 19 x 19 x 50mm) cut from several discs collected from one yellow-poplar tree. In total 60 samples were collected; these were oven dried at 103°C for 24 hours and then stored in a vacuum desiccator for 24 hours prior to NIR analysis with a FT-NIR device using a fibre optic sampling probe. Each wood specimen was scanned 10 times and averaged to a single spectrum. The authors compared the performance of prediction models developed using a limited spectral range (1300-1800nm) with those developed using the full spectral range (800-2500 nm). A total of 40 samples were randomly assigned to the calibration set with the remainder going to the prediction set. PCA, however, found 6 outliers that were removed from the models; hence the calibration set decreased in number to 37, with the number of samples in the prediction set dropping to 17. PLS1 regression was used for each constituent, with models developed for the raw spectral data and the SG first derivative (30 symmetrical points of smoothing with a second order polynomial). It was found that the raw spectral models tended to have lower r^2 and require more PCs (6 to 10 compared with 3 to 6) than the models utilising the first derivative. Furthermore, while the raw spectral models utilising the full spectral region gave higher model performance than using the reduced spectral region, the opposite was the case when the first derivatives were used (with the exception of KL and holocellulose contents).

9.3 Studies on WU Samples - Wet and of a Heterogeneous Particle Size

Studies involving the NIR analysis of lignocellulosic WU samples for constituents of relevance to this thesis are by the far the least abundant in the literature. One of the papers found (Cozzolino *et al.*, 2006a) involved an evaluation as to whether NIRS could be used to predict the chemical composition (dry matter, crude protein, *in vitro* organic matter digestibility, ADF, NDF, and pH) of wet, whole maize

silage. A total of 90 samples were used with moisture contents ranging from 47.2% to 77.2%. These samples were collected from commercial farms and placed in plastic bags. The samples were wrapped in PVC bags and scanned in a coarse rectangular cell (of dimensions 200 mm x 30 mm x 20 mm) using the FOSS 6500 spectrometer. Two scatter correction methods were assessed, SNV-DT and MSC. The mathematical treatments used in the transformation of the spectra were D-1,4,4,1 and D-2,4,4,1. PLS was used for calibration and full cross validation was employed in order to determine the appropriate number of PLS factors. The calibrations were not tested on an independent validation set but instead with cross validation. The results for the ADF, NDF, and dry matter for the full spectral range (400-2500 nm) are included in Appendix B. It can be seen that reasonable calibrations existed except for NDF. It was found that the use of MSC did not improve the NIR calibrations and appeared to have an inconsistent effect on the SECV. In looking to reduce the influence of moisture in the model, the Authors attempted calibrations for the Herschel infrared region (500-1100 nm). The resulting statistics generally showed improvements over those for the full spectral region.

An interesting application of WU NIR analysis to the online, real-time, characterisation of samples is discussed in a paper by Axrup *et al.* (2000). They attempted to predict the monosaccharide (glucose, galactose, mannose, arabinose, xylose), KL, and (acetone) extractive contents for wet (moisture content between 19 and 56% on a wet weight basis) wood chips that were travelling on a moving conveyor at a speed of approximately 1 m/s. These chips were to be used for the Kraft pulping process. Since the samples needed to be analysed while they were moving it was considered that scanning spectrophotometers would not be suitable since different parts of the spectrum would describe different parts of the sample. Instead a diode array detector consisting of silicon detectors capable of measuring in the 800-1100 nm region were used. These detectors allowed 240 scans to be collected per second. For the wet chemical constituents a total of 118 wood chip samples were used and analysed with primary methods. PLS2 regression was used on either uncorrected or MSC-corrected spectra and 35 samples were set aside as a test set to validate the calibrations developed. Reasonable calibrations were developed for extractives and for KL. With MSC used as a spectral pretreatment and five PLS factors, the RMSEP was 1.2% for extractives and 1.8% for KL. However, for the monosaccharides the predictive ability was extremely poor, as shown in the relevant Tables in Appendix B. The poor predictive ability here could be attributed to the limited spectral range used by the spectrophotometer. However, when monosaccharides were grouped according to their number of carbon units, the predictive abilities improved substantially and were close to the repeatability of the reference method.

The authors also calibrated for moisture content for wood chips and bark samples. For wood chips (262 samples in the calibration set and 110 samples in the validation set) an RER of 43 was obtained with MSC pretreated spectra and 3 PLS factors. For bark (359 samples in the calibration set and 180 samples in the validation set) an RER of 19.95 was obtained with MSC pretreated spectra and 5 PLS factors.

The remaining papers relating to WU calibrations for lignocellulosic components tend to involve composts/manures. Vergnoux *et al.* (2009) used an FT-NIR device to calibrate for a variety of physical and chemical parameters for wet samples of compost from a sewage sludge composting plant. These constituents included cellulose, hemicellulose, lignin (all measured using detergent methods), total organic carbon, and total nitrogen. There were a total of 31 samples that had been collected over a period of about 6 months. For each of these any stones and large vegetal components were manually removed and NIR spectra recorded on 5 sub-samples of about 30 mL of homogenised material in an 80 mm diameter petridish. For each petridish six spectra were taken at different points of the dish. This gave a total of 30 spectra per sample. From these a total of 3 averages from 10 different spectra were computed giving 3 spectra to be used for PCA and PLS (i.e. a grand total of 93 spectra). The moisture content of the samples ranged from 19.3% to 67.8%. For the PLS predictions of cellulose, hemicellulose, and lignin the number of samples was smaller than for the calibrations for the other components and there were only 9 samples (27 spectra) in the calibration set and 4 samples (12 spectra) in the test set. The r_{pred}^2 (RMSEP) [number of factors] were 0.99 (1.35%) [13] for hemicellulose, 0.92 (5.23%) [13] for cellulose, and 0.80 (3.16%) [13] for lignin. Considering only the r^2 here can be misleading, the SEP and relative error of prediction were quite high and also the limited number of samples and large number of PLS factors used makes these results hard to trust.

Huang *et al.* (2007) carried out experiments to see whether NIRS could be used to predict the composition of various animal manure composts. Some of the properties of interest included pH, total organic carbon (TOC), total nitrogen (TN), carbon to nitrogen ratio (C:N) and total phosphorus (TP). A total of 120 samples were used (90 in the calibration set and 30 in the validation set) comprising various types of compost materials (cattle manure, pig manure, chicken manure), different composting methods (windrow and pile) and different grain size (granular and powder). NIRS analysis was carried out on fresh and dry (air dried at 65°C until constant weight, and milled to pass a 1 mm screen) fractions of each sample. The moisture content of the manures ranged from 3.2% to 80.9% with a mean of 22.5%. Most interestingly, in the majority of cases the predictive ability of the fresh-samples model was superior to that of the dried-samples model.

9.4 Summary

It can be seen that NIRS has been applied quite extensively in the analysis of lignocellulosic materials with calibrations for a wide variety of biomass types and constituents. In many cases the calibrations that result have been accurate with good RMSEPs, RERs, and RPDs. However on reading this chapter and looking at the Tables in Appendix B, it should be readily apparent that most of this work has been focused on dry samples of a homogeneous particle size (DG). Having to prepare samples in this form takes away much of the attraction from applications of NIRS as a true **rapid** analytical tool. Not having to prepare homogenous samples is a good step in improving the rapidity of the analysis, and several papers have been discussed where this strategy was applied. These mostly relate to wood samples and have had variable successes. Generally the failure of some attempts can be attributed to the limited number of samples that were included in the model, it is reasonable to consider that an increase in complexity of the material scanned by the NIR (e.g. wood strips versus wood meal) would require more samples in order to develop acceptable calibrations. Indeed, in one paper (Downes et al., 2010a) where a significant number of samples were available the predictive abilities of the model were much better.

The ultimate NIR calibration for rapid analysis would not require the samples to be dried (a process which takes at least 12 hours). This is the basis of the WU category discussed in Section 9.3. It is by far the shortest of the three sections, reflecting the paucity of papers relevant to the topic. However, the limited resources that are available do indicate that reasonable predictions for wet feedstocks can be possible. In particular, the development of calibrations using only a limited spectral region appears to be a useful strategy in developing quantitative models that are not overly influenced by the contribution of water to spectral absorbance. As with the other categories (DG, DU) spectral pre-treatments may also play a role here but it appears the most effective pretreatments to use may change when shifting from dry to wet feedstocks. The online-NIR method discussed in the paper by Axrup *et al.* (2000), while limited in its ability to resolve between monosaccharides with the same number of carbon atoms, is interesting given that reasonable calibrations were developed for some constituents using only a very limited spectral range, while scanning wet unground samples that were travelling at a speed of 1 m/s.

Other important points to take from this literature review include the observations, in several papers, that narrow calibrations were not always the best and including a more diverse range of samples can often improve the predictive ability of the model. These improvements can take place in a wide variety

of statistics, not just an inflation of the RER due to an expansion of the concentration range of some constituents.

In summary, useful information was gathered from this literature review, and the knowledge gained helped inform the practices outlined in subsequent chapters and the development of calibrations that would rank among the best of those in Appendix B.

10 Background on Miscanthus

This Chapter will discuss Miscanthus, a perennial C4 rhizome grass that originated from Asia but was introduced to Europe in the 1930s. At that time its main use was as an ornamental grass. However, the development of Miscanthus *x giganteus*, a sterile hybrid of the *M. sinensis* and *M. sacchariflorus* varieties, led to increased interest in utilizing this variety as an energy crop due to its high yields. The crop is also attractive because it has minimal requirements for fertilizer and pesticides. Miscanthus is closely related to sugar cane (Section 12), and the two often hybridise in the wild (Moller et al., 2007).

Miscanthus is classified as a grass; hence, in order to develop a thorough understanding of this feedstock, some background on the nature and properties of grasses will be provided.

10.1 Background on Grasses

10.1.1 Different Components of Grasses

Figure 10-1 provides an illustration of the major anatomical components of most grasses. The below ground component of plants is usually termed the root and is of little relevance for most biomass schemes. The above ground “shoot” component consists of a stem and leaves. The stem can bear one, two or more cotyledons, depending on the plant (monocotyledons and dicotyledons). All grasses are monocotyledons.

Graminaceae is the term given for the grasses. It chiefly covers herbaceous plants, but also some more woody species such as cereals, bamboo, and reeds. The mainly-holocellulose components of biomass (e.g. cereal straw and sugar cane bagasse) tend to have less established markets than the more valuable components of cereals and sugar canes (e.g. barley/wheat and sucrose) whose high prices preclude their use in lignocellulosic fractionating technologies. It is therefore the lignocellulosic parts of the Graminaceae that will be examined in this study.

The stems are mostly hollow, cylindrical and interrupted at intervals by swollen joints or nodes from which the leaves originate. The parts of the stem between the nodes are termed the internodes. At the tip of a growing shoot sits the terminal bud of the shoot that is surrounded by the leaf (or flower) primordias.

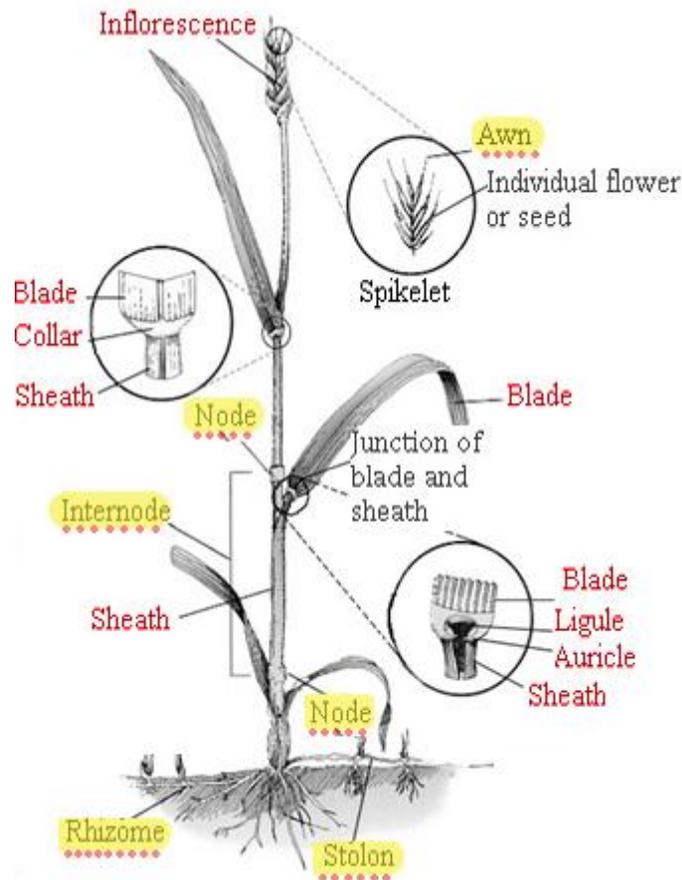


Figure 10-1: An illustration of the main fractions of most grasses. Taken from (Penn State University, 2011)

Rhizomes are a special type of shoot that occur in some grassy species (e.g. *Miscanthus*). These grow underground in a horizontal (plagiotrop) manner and their internodes are usually short. The only leaves present on rhizomes are small and scale-like. Rhizomes serve as survival organs after periods of senescence and also are often used for storage of important nutrients.

The lower portion of the leaf forms a sheath, which encloses and protects the young shoots. The second half of the leaf then opens out into the leaf blade. The midrib is the major vein structure of the leaf. Although it has only a small proportion of the cross sectional area (e.g. 6-13% in grasses), it can comprise 18-28% of the leaf weight and contain 14-24% of the lignified tissues in the leaf (Wilson, 1990).

The proportion of biomass components vary with the species. For example, with barley and wheat straw, internodes dominate (59% and 73% of the dry mass, respectively) while leaves (including sheath

and blade) constitute an intermediate (35% and 21%), and nodes a minor (6.6% and 6.8%) portion of the aboveground dry matter (Aman and Nordkvist, 1983, Theander and Aman, 1984).

10.1.2 C₃ and C₄ Grasses

Grasses can be classified, according to their photosynthetic pathways, as either C₃ or C₄ plants. C₃ grasses that could be considered as energy crops for high-capacity, low-cost, production include reed canarygrass and giant reed while the corresponding C₄ crops include *Miscanthus* and switchgrass. C₄ plants are those of tropical origins that have an extra, preliminary, CO₂ fixing pathway in addition to the Calvin Cycle of C₃ plants. Their principal advantage over C₃ plants is their ability to photosynthesise efficiently under high light intensity and low CO₂ levels. The group also has the highest efficiencies of nitrogen and water usage (Witwer, 1974).

The importance of which pathway is utilised can be understood by observing the Equation (10.1) (Long and Beale, 2001), which summarises the broad physiological processes that determine yield based on the principles developed by (Monteith, 1977):

$$W_h = S_t \varepsilon_i \varepsilon_c \eta / k \quad (10.1)$$

Where: W_h = dry matter at harvest (g/m²); S_t = integral of incident solar radiation (MJ/m²); ε_i = efficiency with which radiation is intercepted by the crop; ε_c = efficiency with which the intercepted radiation is converted into biomass energy; η = amount partitioned into the harvested components; and k = energy content of the biomass (MJ/g)

Of these factors, S_t depends on the site while k varies little between grassy species. ε_i will be determined by the ability of a crop to maintain a closed canopy while, for healthy crops, ε_c varies little within each photosynthetic group (Monteith, 1978). Therefore, theoretical maximal dry matter production, given no other limiting factors, will be determined by photosynthetic type.

ε_c for C₄ plants is approximately 40% higher than that of C₃ species (Monteith, 1978). It should be noted, however, that most C₄ species are tropical in origin and many are not suitable for temperate climates, and hence may not fulfil their growth potential in such climates. These are particularly susceptible to damage at low temperatures during spring and early summer in north-west Europe. For example, even

the maize cultivars bred for western Europe are capable of little photosynthesis at 12 °C, show impaired development of the photosynthetic apparatus in leaves at temperatures below 17 °C, and are subject to light damage (photoinhibition) during low temperature periods (Long, 1999). There are exceptional C₄ species, however, and *Miscanthus* is one of these. It can photosynthesise down to a temperature of <5°C (Long, 1983, Nie et al., 1992).

C₃ and C₄ grasses also have different growing periods, a significant factor when planning biomass-harvesting regimes. C₃ grasses tend to begin growth about 4 to 6 weeks earlier in the spring and to accumulate most of their dry matter before the summer. Then production falls before rising again during the autumn until first frost. Some C₃ grasses, such as reed canary grass, may not go completely dormant in mild winters (Christian and Riche, 2000). Conversely, production in C₄ plants is highest in the summer and stops earlier in the autumn.

10.1.3 Variations in Lignocellulosic Components Between Different Plant Fractions

Compositional data exist for the different plant fractions for various species. These are presented, for cereal straws, in Table 10-1. Similar data for differences in fibre composition were found for the internodes, nodes, and leaves of reed canary grass, a C₃ perennial (Theander, 1991). It can be seen from Table 10-1 that the node fractions generally had greater uronic acids, ash, and arabinose contents than the internodes. This indicates the presence of pectic substances, which would be understandable given the active growth in these regions. Glucose concentrations are significantly higher in the internode regions, which also have the lowest ash and protein contents. This component therefore represents the most attractive part of the plant for acid-hydrolysis technologies. Species with a low leaf/stem ratio and low proportion of nodes to internodes would therefore be preferable.

Liu *et al.* (2010b) separated corn stover and switchgrass plants into their various anatomical fractions and analysed these for a range of lignocellulosic components (glucose, xylose, galactose, arabinose, mannose, KL, acid soluble lignin (ASL), and ash) however the extractives were not removed prior to this analysis. The average results for each section are presented in Table 10-2. If all the constituents for the fractions are summed it can be seen that the total is far from mass closure, attributable to the fact that the extractives were not analysed. It was concluded that the corn husks represented the most attractive fraction of the corn stover given that its total sugar content was high and its lignin and ash contents

relatively low. The observation that the glucan content was greater in the internode sections is consistent with the results for other straws, Table 10-1 and (Duguid et al., 2007). Other noticeable trends in both Table 10-1 and Table 10-2 include larger ash and lignin contents in the leaves and maximum glucose contents in the internodes.

Table 10-1: Mass balance data (% DM) for components of barley and wheat straws. Taken from (Aman and Nordkvist, 1983)

Component	Internode	Node	Leaf	Internode	Node	Leaf
	Barley			Wheat		
Crude protein	1.7	4.0	3.7	3.0	4.5	4.8
Ash	1.6	3.3	4.4	3.8	5.1	9.6
Silica	0.3	0.4	1.1	1.4	1.5	3.9
Dietary fibre	85.1	83.1	79.0	87.2	83.9	83.8
Arabinose residues	1.8	2.7	3.2	1.7	4.1	2.4
Xylose residues	19.9	20.8	21.1	20.7	18.7	20.8
Glucose residues	43.3	33.2	36.4	41.1	32.7	32.3
Uronic acid residues	2.2	4.6	3.2	2.2	4.1	2.6
Klason lignin (KL)	17.6	16.7	14.3	21.6	21.7	26.0

Table 10-2: The average chemical composition of various anatomical fractions (% whole dry mass) of corn-stover and switchgrass samples. Taken from (Liu et al., 2010b)

Constituent	% of Dry Matter										
	Corn Stover							Switchgrass (Alamo variety)			
	Whole	Husk	Sheath	Leaf	Node	Rind	Pith	Whole	Leaf	Node	Internode
Glucan	33.2	37.6	39.6	30.8	29.2	37.8	39.0	36.4	34.7	35.2	41.0
Xylan	18.9	22.2	19.4	16.1	15.9	16.6	17.1	20.2	17.2	22.8	21.4
Galactan	2.2	2.5	2.1	2.2	2.2	1.5	1.7	3.1	2.2	2.3	1.9
Arabinan	3.1	4.8	4.0	3.1	3.9	2.3	2.9	3.7	3.7	4.8	2.8
Mannan	1.1	1.6	1.3	1.2	1.0	1.1	0.8	0.6	0.8	0.9	0.6
Total Lignin	22.1	16.1	16.3	24.0	23.6	22.7	19.9	22.9	24.4	21.4	21.4
Ash	3.4	2.4	5.4	7.4	3.7	3.8	4.9	3.9	4.9	2.6	2.2

10.1.4 Extractives in Grasses

Extractives tend to be more prevalent in grasses than woody biomass. In particular, concentrations of water-soluble carbohydrates tend to be relatively high. Glucose and fructose are the two most important free-monosaccharides that occur in grasses and can have dry matter concentrations of approximately 1-3% (McDonald, 1981). More prevalent is the disaccharide sucrose, which can take up 2-

8% of the dry matter of some forage grasses (Smith, 1973a). In certain grasses, such as perennial ryegrass (*Lolium perenne*) and cock's-foot (*Dactylis glomerata*), oligosaccharides such as raffinose and stachyose have been detected, but only in relatively small concentrations (Laidlaw and Reid, 1952). The polysaccharide fructans tend to be the most abundant of the water soluble carbohydrates, constituting around 5-9% of the dry matter, although values as high as 12% have been reported for perennial ryegrass (Laidlaw and Reid, 1952). Fructans in grasses consist of (2→6)-linked β-fructofuranose units terminating in sucrose residues (McDonald, 1981). These tend to have a low degree of polymerisation and are concentrated most in the stem – Mackenzie and Wylam (1957) found that, in perennial ryegrass, the fructans contents of the leaf were never greater than 4% while the content in the stem frequently exceeded 15%. Extractives content is highly dependent on the maturity of the plant and the environmental influences.

10.1.5 Development of Grasses

There are two distinct physical developments that take place with increasing maturity – (i) a change in the proportion of biomass components; (ii) a change in the tissue types, and their properties, of each component.

The leaf to stem ratio of grasses decreases with increasing maturity. Table 10-3 illustrates the effect this has on polysaccharide-sugars contents. It shows the chemical development of alfalfa (*Medicago sativa* L.) and timothy (*Phleum pratense*) harvested at different stages of maturity. For alfalfa during maturation the stem fraction increased from 19.0% to 51.0% of the aboveground dry matter with a corresponding decrease in the leaf fraction. This shift results in the most important fibre residues shifting from uronic acids, glucose, arabinose and galactose residues in the young plants to glucose, uronic acid and xylose residues, along with KL, in the mature plants.

With regard to changes in tissue structure (which is also responsible for some of the variation in Table 10-3), this only occurs for some components. While the proportions of the different tissue types in a blade do not appear to change within a leaf as it ages (Cherney and Marten, 1982), stem tissue characteristics change greatly with age (Cherney and Marten, 1982). Principally, there tends to be a

secondary thickening of the cell wall and extra lignification. The result is that the lower stem is generally more highly lignified than the upper stem during the developmental process in grasses.

Table 10-3: Chemical composition (% DM) of two grasses - whole crop alfalfa (Medicago sativa) (Nordkvist and Aman, 1986) and timothy (Phleum pratense) (Aman and Lindgren, 1983) at different stages of maturity.

Constituent	Lucerne Crop (% DM)			Timothy (% DM)		
	11 May	8 June	13 July	17 June	7 July	20 July
Ash				12.4	8.5	6.2
Crude protein	36.8	18.1	12.9	22.4	11.4	8.1
Dietary fibre	32.9	52.2	64.8	51.2	70.9	78.8
Rhamnose	0.4	0.5	0.5	0.5	0.3	0.2
Fucose	0.1	0.2	0.1	Trace	Trace	Trace
Arabinose	2.7	2.5	2.2	2.9	3.8	3.9
Xylose	1.3	5.8	7.6	8.9	16.0	18.7
Mannose	0.8	1.4	1.5	Trace	Trace	Trace
Galactose	2.0	1.9	1.8	0.8	0.8	0.9
Glucose	9.5	19.2	24.7	23.1	32.2	36.4
Uronic acid	11.5	10.8	11.6	2.5	2.5	2.9
Klason lignin (KL)	4.3	9.8	14.7	12.5	15.3	15.9

Farmers that grow forage crops for consumption by ruminants sometimes harvest early in order to minimise cell lignification, maximise protein content, and improve digestibility. However this is at the expense of reduced yield. Lignocellulosic fractionating technologies, which look for high polysaccharides-sugar contents and are unable to utilise protein and ash, would not favour such a harvesting regime, unless they are susceptible to lignin-induced process inhibition.

10.1.6 Environmental Effects on Grasses

Given that the grassy biomass species under consideration for lignocellulosic fractionating technologies are harvested annually, seasonal conditions are likely to be highly important in dry matter and lignocellulose yields. These would be more important than the annual conditions in the year of harvest of woody species, since those crops tend to have longer cutting cycles. Indeed, the plant environment can exert a great influence on the proportion of components and the chemical composition of many grass species.

For example, external stresses that slow the growth and development of plant tissue, tend to increase the leaf/stem ratio (Van Soest, 1982). Given the higher proportion of carbohydrates in species with lower leaf/stem ratios, the reduction in carbohydrate yield would be unattractive for lignocellulosic fractionating technologies as, obviously, would be the reduction in yield. Due to this response, the total sugar yield would be lower than that predicted from a simple stress-induced yield reduction.

10.2 Establishment and Development Cycle of *Miscanthus*

There are numerous varieties of *Miscanthus* that have been grown experimentally (Clifton-Brown et al., 2001a) but fewer varieties tend to be grown commercially. In Ireland *Miscanthus x giganteus* is the only commercial crop that has been established so far, although there have been experimental plots of other varieties, principally *M. x sinensis*. Elsewhere in Europe, particularly in the regions that experience colder winters, *M. x sinensis* has been grown commercially.

Establishment:

M. x giganteus is sterile, and so must be propagated vegetatively. This can either involve the planting of rhizome cuttings or plantlets (significantly more expensive). The plantlets are much more susceptible to frost-mortality and so tend to be planted later (April to May) than rhizomes (March to April). Rhizomes can be collected from nursery fields where *Miscanthus* has already been established – these are broken up, collected and planted using existing agricultural equipment such as potato harvesters and planters.

Miscanthus can grow on a wide variety of soils (e.g. sandy soils as well as high organic matter content soils) and, while the ideal pH range should be between 6.6 and 7.5, it can tolerate more acidic/alkaline conditions (Moller et al., 2007). The main problems associated with the establishment of *Miscanthus* crops are those relating to over-wintering. *Miscanthus* plantations in the northern regions of Europe (*M. x giganteus* varieties in particular) have suffered from poor survival over the winter in the first year after planting (Christian and Haase, 2001). Trials at a site in Cashel, Co. Tipperary, showed that micro-propagated plants are much more susceptible for winter mortality than rhizome propagated plants – there was a 95% survival rate for the rhizomes but the survival rate for the micro-propagated plants was

only 17% (Christian and Haase, 2001). The level of winter-failure was attributed to killing by late spring frosts of the first shoots produced. If this happens, the plants will not re-sprout. The failure of the micro-propagated plants to develop a sufficient rhizome system to retain enough reserves for early shoot growth may be the reason for their high mortality. Overwintering is only a problem in the establishment year of the plantation, after that period the rhizomes are developed enough to escape frost mortality.

Other *Miscanthus* genotypes, such as *M. x sinensis*, are more resistant to cold winters due to their higher frost resistance. However, in periods where winter rhizome destruction is not a problem, *M. x giganteus* yields tend to be higher than those of *sinensis*. The reasonably mild winters experienced in Ireland, particularly the west coast, may reduce the risk of overwintering for rhizome-propagated plants. However, the establishment costs for *Miscanthus* are currently significant, between €1,060 and €2,555 per hectare in Ireland (Styles et al., 2008).

Growth:

Spring growth will start once daytime temperatures exceed 10°C. In May, June, and July, growth is very rapid and results in cane-like stems that may reach a height of 3 m or more. Once the canopy closes, the lower layers of leaves begin to senesce. Shoot growth continues through August and September with full senescence occurring following the first frosts of the autumn. During the end of the growing season, nutrients are translocated from the stems and leaves to the rhizomes for storage and utilisation the following season. The efficient use of nutrients by *Miscanthus* varieties means that fertilisation levels need not be high. The low fertiliser and pesticide requirements of *Miscanthus* mean it is a relatively environmentally friendly crop - Spink & Britt (1998) identified *Miscanthus* as being one of the most environmentally benign alternatives to permanent set-aside land.

Fertilisation should be carried out after the harvest but before shoots grow sufficiently high that they become damaged by the equipment. While the plots at Cashel are still productive in their later years, ten years after planting potassium deficiency became relevant and tenth year yields from unfertilised plots were significantly lower than for fertilised plots (Clifton-Brown et al., 2001b).

If *Miscanthus* is harvested in the autumn, extra leaf mass is taken from the field (Lewandowski et al., 2003a). This means that less will decay on the ground and, hence, less nutrients will be transferred to the soil. More (perhaps up to 100%) fertilisation will therefore be required (Kristensen, 1999).

10.2.1 Harvesting Window

Harvesting of *Miscanthus* is carried out annually and can take place using conventional harvesting equipment. It can occur after crop senescence until just before re-growth in the following spring. It is important that the crop has senesced so that translocation of assimilates to the rhizomes has occurred. Irish growth will cease at the first frosts of autumn or winter. After this period crop senescence accelerates, nutrients are sequestered to the rhizomes, and moisture falls. Harvesting is necessary before spring growth to avoid sprout damage. Current practices involving *Miscanthus* seem to be geared towards harvesting the biomass in the spring before the start of the growing season (Bullard and Nixon, 1999). The reasoning behind this is that the later that the harvest can be carried out, the lower will be both the moisture content and the inorganic mineral content; these are important qualities in biomass combustion.

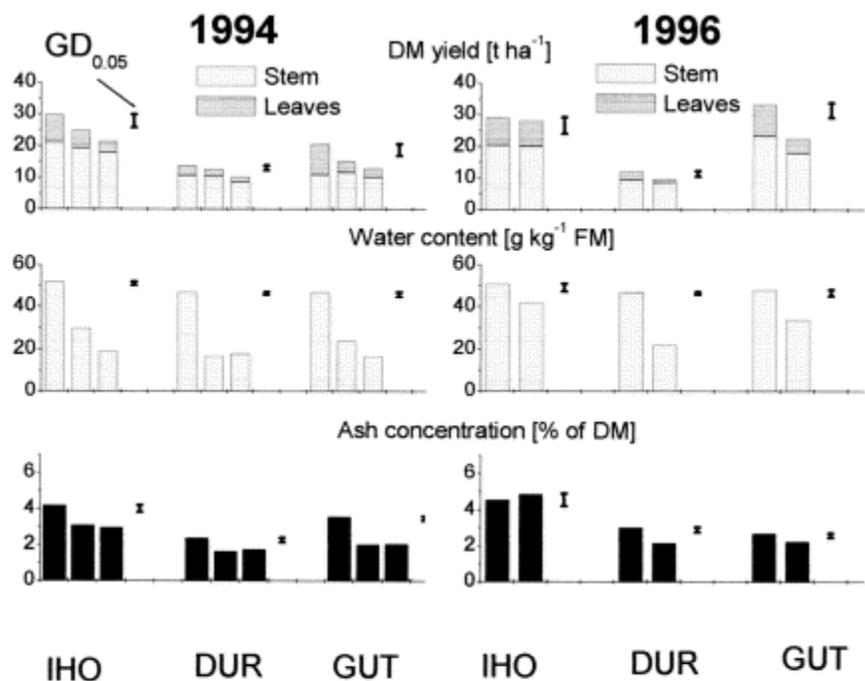


Figure 10-2: *Miscanthus* dry matter yield, moisture content, and ash concentration at three sites for harvests at December and February (for the crop grown in 1994 and 1996) and at March (for the 1994 crop). These dates are plotted from left to right for each site. The three different locations in Southern Germany are Inger Hof (Iho), Durmersheim (Dur) and Gutenzell (Gut) - Taken from Lewandowski and Heinz (2003).

The research of Lewandowski and Heinz (2003) illustrates well the dynamics involved over the harvest window. Figure 10-2 shows the effects, at three locations, of delaying the harvest. *Miscanthus* that was grown in 1994 was harvested in December, February, and March. *Miscanthus* that was grown in 1996 was harvested in December and February. It can be seen that, except for one site in 1996, delaying harvest to February always resulted in a loss of yield, while the moisture content also fell dramatically. On average, dry matter yields were reduced by 18% between December and February and by an additional 16% in 1995 between February and March. Figure 10-2 shows that the principal reason for this reduction in dry matter was due to the loss of leaves; the stem component does not decrease to a great degree in most cases. As mentioned before, the concentration of carbohydrate and lignin per tonne of harvested biomass is therefore likely to change significantly given the loss of this component. With regards to the ash content, a harvest delay from December to February reduced this significantly but a delay from February to March did not.

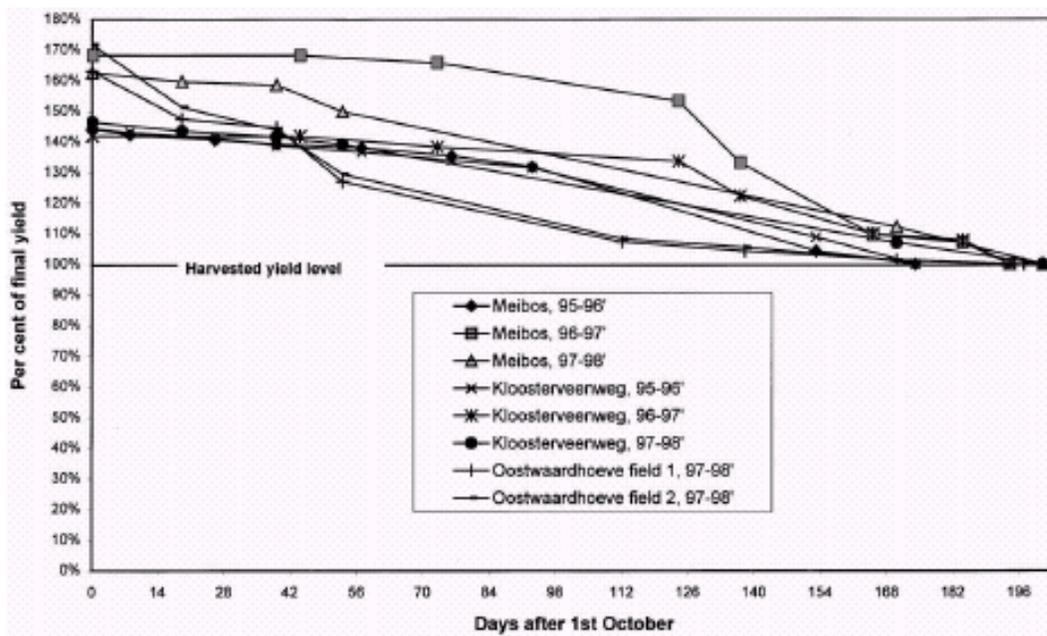


Figure 10-3: Decrease in the harvestable amount of *Miscanthus* over the course of six months after the 1st of October, expressed in terms of the final yield harvested in April, for various years and locations in The Netherlands. Taken from (Long and Beale, 2001)

Experiments at the Cashel site also back-up these data – the ceiling figures of between 14 and 16 t dry matter ha⁻¹ y⁻¹ that were attained after the first four years of growth, were figures for the March harvest.

They found that there were dry matter losses of almost 40% which occurred between the first air frost in autumn and final harvest in March (Clifton-Brown et al., 2001b). Figure 10-3 shows, for several years and stands in The Netherlands, that the standing harvestable-biomass after senescence can be as much as 70% higher than that available in the spring (Long and Beale, 2001).

10.3 Yields

Research throughout Europe has found that the yields associated with different *Miscanthus* varieties can vary dramatically, e.g. between 2 and 44 t DM ha⁻¹ (Lewandowski et al., 2003b). Highest yields tend to occur in warmer, sunnier, southern European climates although, in such locations, water availability becomes a limiting factor. In more northern areas, where global radiation and average temperatures are lower, and hence become the limiting factors, yields without irrigation are more typically 8-25 t DM ha⁻¹ (Lewandowski et al., 2003b).

It generally takes several years for annual productivity to reach ceiling yields, representative of the time required for full establishment of the *Miscanthus* stand. For example, it was found in that, in southern Germany, autumn yields of *Miscanthus x giganteus* increased from about 2-3 t DM ha⁻¹ in the first year after planting to 22-30 t DM ha⁻¹ in the third year (Clifton-Brown et al., 2001b). It has also been reported that the period required to obtain ceiling yields is longer in temperate climates (up to 6 years) than in warmer climates (within 2 years). For example, a site at Trinity College Dublin took 5 years to reach its ceiling yield of 14 t DM ha⁻¹ (Clifton-Brown, personal communication 2003).

10.4 Costs Involved in *Miscanthus* Production

There have been extensive studies on the costs and profitabilities of various *Miscanthus* scenarios (Bullard, 2001, Rutherford and Bell, 1992, Christian and Riche, 2000, Venturi et al., 1999, Bullard and Nixon, 1999). Bullard and Nixon (1999) estimated the cost for *Miscanthus* plantations in the UK. They assumed a ceiling yield of 18 odt/ha that was reached after 3 years. Bullard and Nixon (1999) assumed a unit price for rhizomes of 7.5 cents, resulting in total rhizome costs per hectare (with a plantation

density of 20 000) of €1,500. Given land preparation and planting operations, the total establishment costs were estimated at approximately €2,000/ha. Annual husbandry costs (fertiliser, herbicides) were low at approximately €100/ha. They found that the cost of harvesting one hectare of *Miscanthus* using a baling system and subsequently storing for 6 months was €305/ha. The associated cost for chopping the *Miscanthus* with a forage harvester was €316/ha. It was assumed that 20% of the stand biomass was lost through storage and harvesting losses. The result of their studies was that the break-even cost for the production of baled *Miscanthus*, with no subsidies, was €69/odt. This fell to €33/odt when set-aside support was included.

The assumption of a yield of 18 odt⁻¹ ha⁻¹ in that study was somewhat optimistic. A subsequent paper by Bullard (2001) considered Irish conditions, among those of other countries, and varying yields. Break-even costs, under no subsidy, ranged from €72.9 for 12 odt/ha to €37.96 for 24 odt/ha. Venturi *et al.* (1999) found that the break even cost, excluding land costs, was €40 per tonne for baled *Miscanthus* and €28 per tonne for chopped *Miscanthus*. They found that the corresponding costs for chipped and whole stem willow were €50 and €35, respectively.

10.5 Lignocellulosic Properties of *Miscanthus*

Most analysis on *Miscanthus* has focussed on its moisture content, energy value, and ash composition – all important qualities for combustion (Lewandowski and Kicherer, 1997). Less data exist for the lignocellulosic components of the feedstock. General characteristics of grasses (Section 10.1) should apply – there should be a large concentration of cellulose and hemicellulose with the hemicellulosic fraction being closely related to xylans, with a low amount of glucomannans present.

Kaack *et al.* (2001) examined various properties of *M. x giganteus* stems, selected at 5 different points in the harvest window (between mid-November and mid-March) at a stand in Denmark. These properties included the stem length, internode length, and the number of internodes per stem. It was found that the average stem length increased from 197 cm at the first harvest to 206 cm at the second harvest, but then decreased linearly over time to 169-170 by the last two harvests. Furthermore, the maximum number of internodes (12) was found in the second harvest period, but this value fell to 9 by the last

harvest. These reductions were attributed to lodging occurring because of abscission of internodes during the period from December to March.

Faix *et al.* (1989) attempted to find the absorption coefficients for the acid soluble lignin (ASL) content of *M. x sinensis* by obtaining UV spectra of acid hydrolysates of the milled wood lignin (MWL) of this feedstock. The resulting spectra and the absorption coefficients obtained are provided in Figure 10-4. For their research, which concerned both *M. sinensis* and *Arundo donax*, the authors chose to use the absorbance at 280 nm since this provided similar absorption coefficients for both feedstocks. This and other results from their analysis are included in Table 10-4.

One of the most extensive characterisations of *Miscanthus* was carried out by Visser and Pignatelli (2001). The results of their analyses are provided in Table 10-4. The samples used were composed of a 52% basal section, 33% centre section and 15% top of the stem of a second-year crop of *Miscanthus x giganteus*. They also found a degree of polymerisation of approximately 1 300 for cellulose. Table 10-4 also provides data from other sources, including those obtained by Papatheofanous *et al.* (1996) for *M. x sinensis* (using detergent analysis methods).

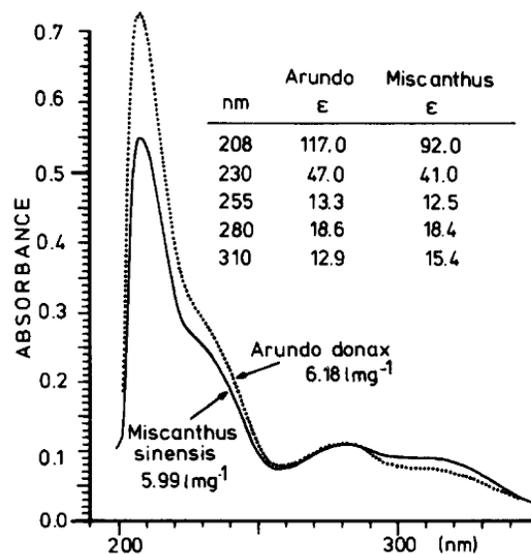


Figure 10-4: The UV spectra of milled wood lignins from *Arundo donax* and *M. sinensis*. The absorption coefficients that were calculated for several wavelengths are also included. Taken from (Faix *et al.*, 1989).

A paper by Le Ngoc Huyen *et al.* (2010) is especially relevant. It discusses the sampling at two points in the harvest window of a two year stand of *M. x giganteus*, an early harvest (November) and a late harvest (February the next year). These dates corresponded to a dry matter yield of 21 t DM ha⁻¹ in the early harvest and 15 t DM ha⁻¹ in the later harvest. These samples were then either analysed whole or separated into the following fractions which were then analysed separately: leaves, sheaths, lower internode section, upper internode section. These internode sections were selected by numbering each internode from the base of the stem and selecting the basal section (number 2, represented by IN2 in Table 10-5) and the apical regions (number 11, represented by IN11 in Table 10-5). All other internodes and all nodes were discarded and not analysed. The analytical protocol involved obtaining the neutral detergent fibre (NDF) fraction which was then hydrolysed using the conventional two-stage acid hydrolysis procedure (Section 3.1.2) and the liberated sugars quantified by HPAEC-PAD. The analyses of acetyl content and numerous lignin components were also carried out. The values for sugars, acetyl, lignin etc. in the paper are presented on a % NDF basis; however, Table 10-5 has corrected these for the NDF content of the sample so that they are on a whole mass basis. Interesting points to take from Table 10-5 are that the ratio of xylose to arabinose is approximately three times higher in the internodes than it is in the leaves and sheaths, it was suggested that this indicates a higher proportion of primary cell walls in the leaves/sheaths. Lignin was also highest in the lower internode section, as would be expected. Galactose is higher in the leaves and sheaths than in the nodes and the values for the acetyl content seem quite high for all fractions, particularly when compared to other lignocellulosic feedstocks, e.g. 1.7% in wheat straw (Kabel *et al.*, 2007). Interestingly Le Ngoc Huyen *et al.* (2010) found barely any uronic acids in their analysis (0.01% of dry matter NDF).

Table 10-5 shows that there is a clear increase in the NDF content of the whole plant associated with the later harvest as opposed to the early harvest. That would be attributable to the loss in leaves that occurs over this period. However, regarding the composition of the NDF fraction itself, the authors concluded that the major influence due to harvesting date was on the cell wall phenolic fraction of the whole biomass (lignin composition, phenolic acids) whereas no significant differences were found for the glucan or total lignin content. The authors also undertook experiments involving the enzymatic saccharification of these samples and found that the lower internode sections were the most recalcitrant, whereas the leaves and sheaths displayed similar susceptibilities to the enzymatic conversion of cellulose and arbinoxylans (being 2.5 to 3 times greater than for the internodes). It was theorised that the lignins in the leaves/sheaths were less recalcitrant than those in the internodes.

Table 10-4: Lignocellulosic contents of *Miscanthus*, taken from several sources.

Source	% Dry Matter				
	(de Vrije et al.)	(Visser and Pignatelli, 2001)	(El Hage et al., 2010)	(Papatheofanous et al., 1996)	(Faix et al., 1989)
Variety	<i>Giganteus</i>	<i>Giganteus</i>	<i>Giganteus</i>	Sinensis	Sinensis
Section	Stems	Stems	Spring harvest	Spring harvest	Stems
Cellulose	38.2			43.1	
Hemicellulose	24.3			26.7	
Glucan	39.5	38.8	45.6		48.3
Xylan	19.0	24.3	22.5		19.0
Arabinan	1.8	2.3	2.3		1.4
Galactan	0.4	0.6	0.4		0.5
Mannan			0.2		
Uronic acids	1.8	-			
Total Lignin	25.0	25.2	26.5	22.1	23.5
Klason Lignin	24.1	23.6	26.0		
ASL	0.9	1.6	0.5		1.6
Solvent Extract.	4.2*a	2.2*b	1*c		
Water Extract.	1.4				
Ash	2.0	3.0	2.8	3.9	5.7

*a = ethanol:toluene (2:1); *b = ethanol; *c = dichloromethane

Table 10-5: The relative amounts (% dry matter) of lignocellulosic constituents in the various anatomical fractions of *Miscanthus x giganteus* harvested in November and February. S/G = syringyl/guaiacyl lignin ratio. Data adapted from (Le Ngoc Huyen et al., 2010)

Constituent	Amount (% dry matter) For Corresponding Fraction								
	Early Harvest (November)					Late Harvest (February)			
	Whole	IN 2	IN11	Green Leaves	Green Sheath	Whole	IN 2	IN11	Sheath
NDF	76.37	89.16	91.05	76.86	78.27	86.75	91.19	91.13	83.98
Arabinose	2.49	1.19	1.35	3.50	3.38	2.41	1.40	1.65	3.27
Galactose	0.47	0.28	0.24	1.03	0.89	0.30	0.30	0.28	0.94
Glucose	37.82	46.74	46.00	30.07	35.30	43.78	46.04	45.23	36.66
Xylose	16.90	15.85	18.10	15.59	16.15	18.81	16.53	17.62	16.54
Xyl/Ara Ratio	6.78	13.37	13.73	4.44	4.78	7.81	11.90	10.73	5.07
Total Sugars	57.68	64.05	65.65	50.20	55.71	65.31	64.27	64.78	57.41
Acetyl	3.22	3.04	3.06	2.54	2.89	2.98	3.16	2.99	2.94
KL	14.40	19.87	15.67	13.41	12.78	16.68	17.36	17.35	15.25
S/G Ratio	0.56	0.64	0.64	0.43	0.34	7.81	0.83	0.83	0.48

Hodgson *et al.* (2011), as part of the European *Miscanthus* Improvement project that included the growth of 15 *Miscanthus* genotypes at a plot in Rothamsted Research in the UK, analysed five of these genotypes for their acid detergent lignin (ADL), cellulose (acid detergent fibre (ADF) – ADL), hemicellulose (NDF-ADF) and ash contents and compared these properties between November (early) and February (delayed) harvests. The results are provided in Table 10-6. It can be seen that there are

clear differences in compositions between the genotypes; for example, *M. x giganteus* and *sacchariflorus* were generally higher in lignin and cellulose and lower in hemicellulose than the *M. x sinensis* genotypes. There were less significant differences between the harvest dates, although the ash contents did fall.

Table 10-6: Cell wall composition of *Miscanthus* species and genotypes in November and February harvests. Ho:Lg = Holocellulose content divided by lignin content. Lig = lignin, Cell = cellulose, Hc = hemicellulose. Taken from (Hodgson et al., 2011).

Variety (M. x)	Genotype	Amount of Contituent (% DM) According to Each Harvest Date									
		November Harvest					February Harvest				
		Lig.	Cell.	Hc.	Ho:Lg	Ash	Lig.	Cell.	Hc.	Ho:Lg	Ash
<i>Giganteus</i>	EMI01	12.0	50.3	24.8	6.3	2.7	12.6	52.1	25.8	6.2	2.7
<i>Sacchariflorus</i>	EMI05	12.1	49.1	27.4	6.3	2.3	12.1	50.1	28.1	6.5	2.2
<i>Sinensis</i> (hybrid)	EMI08	9.3	43.1	33.1	8.2	3.5	9.7	45.3	33.0	8.1	2.7
<i>Sinensis</i>	EMI11	9.7	47.6	34.0	8.0	3.2	10.3	45.5	30.6	7.7	3.0
<i>Sinensis</i>	EMI15	9.2	46.7	33.0	8.8	2.4	9.3	52.2	30.3	8.9	2.2

10.6 Previous NIRS Research with *Miscanthus*

NIRS and *Miscanthus* are linked in the literature typically in situations where the feedstock occurs as one, or as a small number of, sample(s) in a diverse mix of interspecies samples that have been used to develop a global calibration (e.g. (Sanderson et al., 1996)). The Author found few papers in the literature that focused on the development of specific quantitative calibration equations for *Miscanthus*; however, the two most relevant papers in this area are discussed below and the relevant statistics from these are included in the Tables in Appendix B.

Fagan *et al.* (2011) undertook experiments to predict, using a global dataset containing both feedstocks, the moisture, calorific value, ash, and carbon content of *Miscanthus* and short rotation coppice willow (SRCW) samples. Two varieties of SRCW were used, Tora and Karin. All samples were subjected to different fertilisation treatments (using differing amounts of waste water from a dairy facility). The *Miscanthus* was harvested in January and the willows in October. Only the lower 1.5 m of the stems were sampled, and the branches and leaves were removed from these sections, prior to analysis. The variations in moisture contents were artificially generated by drying subsamples for differing periods of time in a convection oven at 105°C. Following drying, all samples were ground to pass a 3 mm mesh. NIR data collection took place over the range 400-2500 nm on a FOSS 6500 system using a circular reference

cell. Two spectra were taken per sample and the average was used. Partial Least Squares regression (PLS) and full cross-validation were employed and various combinations of spectral pre-treatments (multiplicative scatter correction (MSC), standard normal variate (SNV), first derivative (1D), 2D) and wavelength ranges (400-2500, 400-750, 400-1100, 750-1100, 1100-2500) were tested in the development of calibrations. Heating values were determined using an adiabatic bomb calorimeter and carbon values with a carbon analyser.

The quality of the resulting calibrations varied. The best moisture content calibration involved using MSC and 1D, the 1100-2500 nm spectral region, and 7 factors, and resulted in an r_{CV}^2 of 0.99, root mean standard error of cross validation (RMSECV) of 0.90%, RER of 39.3, and an RPD of 13.5. The calorific value calibration was also good, with the best model using a 1D pretreatment, the 1100-2500 nm spectral range, and 5 factors, providing an r_{CV}^2 of 0.99, RMSECV of 0.13 MJ/kg, RER of 39.9, and an RPD of 70.8.

The results for carbon content, on both wet and dry basis, were reasonable. The best carbon model (wet basis) involved using the second derivative spectra over the range 1100-2500 using 4 factors, and provided an r_{CV}^2 of 0.88, an RMSECV of 0.57%, an RER of 10.4, and an RPD of 4.6.

However, the results for the ash content, whether on a wet or dry basis, were poor. The best ash model involved the MSD and 1D pretreatment on the 750-1100 nm spectral region and 5 factors, providing (for wet basis values) an r_{CV}^2 of 0.58, RMSECV of 0.42%, RER of 7.73, and an RPD of 3.6.

The only other paper concerning the specific NIRS analysis of Miscanthus that could be found in this literature review was by Hodgson *et al.* (2010). They analysed various samples obtained from the EU Funded European Miscanthus Improvement (EMI) project mentioned in Section 10.1.4. However, this time the study covered all 15 varieties and five different sites across Europe. The authors used samples from two periods (autumn 1999 and winter 2000). This resulted in a total of 366 samples. The material was oven dried (105°C) and milled (whole plant basis) prior to chemical and NIRS analysis. Wet chemical analysis was carried out for acid detergent lignin (ADL), ADF, and NDF. For NIRS analysis a 76 sample set that combined spectra from all five countries was used, and these samples were scanned in the 1100-2500 nm region. SNV-DT transformations were applied to the spectra along with either a D-1,4,4,1 or D-2,6,4,1 derivative treatment, depending on the constituent of interest, and PLS was used for the generation of calibration equations. The resulting calibration equations were then used to predict the ADL, ADF, and NDF contents of the remainder of the 366 EMI samples. Cellulose was estimated as ADF

minus ADL and hemicellulose as NDF minus ADF. As shown in the relevant Tables in Appendix B, SECV (r_{CV}^2) of 1.74% (0.90), 1.36% (0.93), and 0.69% (0.75) were obtained for ADF, NDF, and ADL, respectively, for the 76 calibration samples. When these calibrations were applied to the larger data-set it was found that the site had a significant effect on cellulose and hemicellulose content in both harvests (Autumn 1999 and Winter 2000).

10.7 *Miscanthus* in Ireland

All commercial *Miscanthus* plantations in Ireland are of the *M. x giganteus* variety. There have been a number of varieties grown at the Oak Park Teagasc experimental site in Carlow, as discussed in Section 16.2, where *Miscanthus* has been in production for over 15 years. *Miscanthus* has also been grown in Cashel, Co. Tipperary, since 1997. The primary use for commercially-grown *Miscanthus* is as a fuel, supplied as co-feed with peat in the Bord na Móna power station in Edenderry, Co. Offaly. Up to 30% biomass as a co-feed is mandated for 2015 and the power plant operator expects that *Miscanthus* will contribute 10% to the total fuel mix. As a result of this end use for the crop it is typically harvested in the Spring (March/April) so that the ash, moisture contents, and leaf:stem ratios are at a minimum. There are currently no uses for early harvest, i.e. wet, *Miscanthus* in Ireland.

The total area of land under *Miscanthus* cultivation in Ireland has expanded in recent years, primarily as a result of the introduction of the “Bioenergy Grant Scheme for Willow and *Miscanthus*” which provides establishment grants, up to €1 300 per hectare, to cover 50% of the costs. According to the official statistics for applications made to this scheme 599 hectares were planted in 2007, 752 in 2008, 704 in 2009 and 182 hectares in 2010, totalling 2238 hectares over this period. In comparison, funding was sought for establishment grants for a total of 577 hectares of willow short rotation coppice plantations over this same period. Data are available on a per-county basis and these are provided (for the combined 2007-2010 period) in Table 10-7. It can be seen that the counties of Cork, Kilkenny, Limerick, Tipperary, and Wexford contribute 65% of the total *Miscanthus* area. This is in contrast to the coppice plantations where these counties only contribute 18.1% towards the total (with no hectares planted in Limerick and Kilkenny) with Meath (24.5%), Cavan (17.1%), and Monaghan (12.4%) contributing the majority of the plantations. In 2011 the Sustainable Energy Authority of Ireland (SEAI) launched the “Bioenergy Mapping Scheme” website. This can be accessed at <http://maps.seai.ie/bioenergy/> (last

accessed by the Author on 16/4/11) and allows the known locations of numerous crops to be plotted on a map of Ireland. Figure 10-5 shows this map when all known Miscanthus plots were selected.

Table 10-7: Total hectares (ha) applied for Miscanthus establishment grants, by county, over the period 2007-2010

County	Total ha	% of Total
Carlow	39.4	1.76%
Cavan	15.4	0.69%
Clare	11.3	0.51%
Cork	327.6	14.64%
Donegal	6.9	0.31%
Dublin	0.0	0.00%
Galway	99.8	4.46%
Kerry	109.5	4.89%
Kildare	37.4	1.67%
Kilkenny	233.6	10.44%
Laois	49.8	2.22%
Limerick	365.4	16.33%
Longford	17.9	0.80%
Louth	10.6	0.47%
Mayo	37.3	1.67%
Meath	31.7	1.42%
Monaghan	16.4	0.73%
Offaly	32.6	1.46%
Roscommon	25.7	1.15%
Sligo	6.9	0.31%
Tipperary	351.7	15.72%
Waterford	97.6	4.36%
Westmeath	62.4	2.79%
Wexford	203.3	9.09%
Wicklow	47.5	2.12%

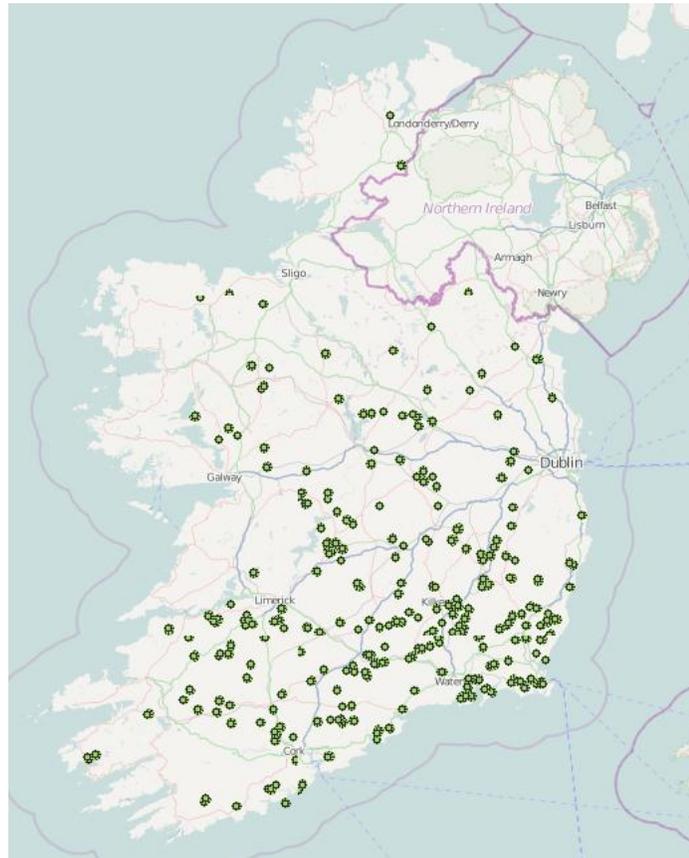


Figure 10-5: A map showing the Miscanthus plantations across Ireland. Taken from <http://maps.seai.ie/bioenergy/> on 16/4/11

Styles *et al.* (2008) undertook, for Irish conditions, a life-cycle-assessment for a 16 year (14 harvests) Miscanthus plantation. This involved taking, from the literature, the costs associated with each activity in the life cycle and inflating these to 2006 prices. The paper considered the use of the feedstock for electricity/heat production and assumed payment prices for Miscanthus at 70, 100 and 130 € t⁻¹ DM for mid, low, and high estimates. It was found that annualised production costs for Miscanthus ranged from €430 to €559 ha⁻¹, equivalent to between €37 and €48 t⁻¹ DM. Under the mid-costs and mid-price scenario the annual gross margins for Miscanthus production were €326-383 ha⁻¹ with these margins

rising to up to €586 ha⁻¹ when using the low cost estimates. It was found that most of the Miscanthus scenarios were highly competitive with all other agricultural land uses with the exception of dairy.

Walsh (1999) carried out a reasonably simplistic evaluation in the Irish context. She assumed a productivity of 15 dry tonnes per hectare and found that annual costs were €803/ha (including a €480 land rental), equating to a cost of €53.53 per tonne with no subsidy or €25.80 with set-aside payments.

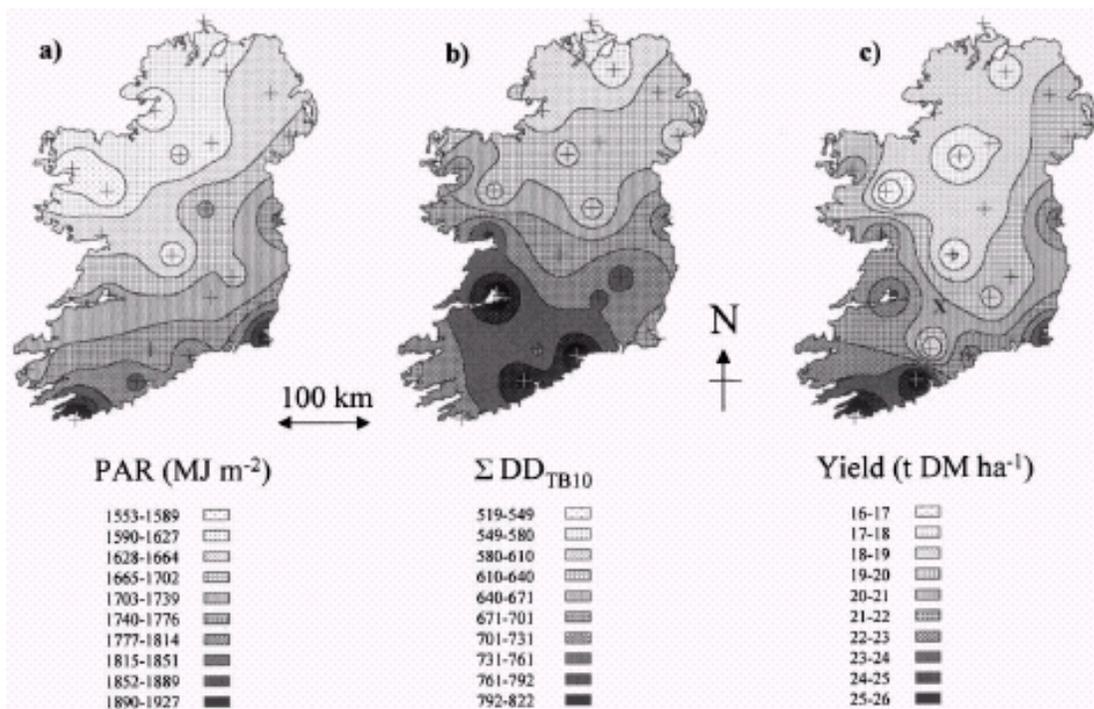


Figure 10-6: Miscanthus modelled productivity maps for Ireland. (a) Total annual mean radiation (MJ/m²); (b) degree days above 10°C; (c) mean simulated yield at the end of the growing-season for *M. x giganteus*. Taken from (Clifton-Brown et al., 2001b).

A very useful yield prediction model was designed using measurements from a four year old field trial of *M. x giganteus* in Cashel (Clifton-Brown et al., 2001b). It assumed that rainfall was not limiting and that fertiliser application was adequate; hence growth was dependent on air temperature and solar radiation. By linking empirical growth and climate parameters (Clifton-Brown, 1997), with data from meteorological stations in Ireland, a geographic map was produced to indicate the potential primary production of above-ground dry matter, as shown in Figure 10-6. The yields predicted by the model were the standing biomass yield in early autumn when the first frost occurred. The predictions for the

trial site over two years were 17.3 and 18.4 t ha⁻¹, 27% and 31% higher than the December-harvested dry matter yields of 13.6 and 14.0 t ha⁻¹. While the overestimation may be partly due to the relative simplicity of the model, the dry matter loss experienced between senescence and harvesting (e.g. Figure 10-3) could also be responsible.

The model predicted substantial interannual variation in yields, mainly as a consequence of the changes in the length of the frost free period. Nixon and Bullard (2001) developed a similar model for the United Kingdom. This model also incorporated soil series data for England and Wales to allow the user to also calculate potential yields where water availability may be a limiting factor.

The MiscanMod model (Clifton-Brown et al., 2002) that was subsequently developed by Clifton-Brown has allowed predictions of the production potential throughout Europe for *M. x giganteus*, based on local climatic conditions (temperature, radiation, rainfall and soil water holding capacity).

11 General Analysis Methodology

This chapter will provide the general methodology employed for the processing and the, NIRS and reference, analysis of biomass samples. There was some variation in the methodologies according to the stage of the project and the sample type analysed, and these differences will be outlined in the relevant chapters.

As outlined in Section 3.9 it was decided that a hydrolysis procedure similar to that available for download at the NREL website would be carried out for biomass that had been previously extracted with 95% ethanol using the Dionex ASE 200. The hydrolysate would allow for the analysis of the liberated monosaccharides (arabinose, galactose, rhamnose, glucose, xylose, and mannose) via HPAEC-PAD (Section 4), and the acid soluble lignin (ASL) content via UV-spectroscopy. For a limited number of samples uronic acid (UA) analysis took place using this hydrolysate. The acid-insoluble residue, once weighed and then ashed, would allow the Klason lignin (KL) content to be determined. Figure 11-1 illustrates this sequence of steps starting from a dry and sieved (DS) sample of biomass.

Of course sample preparation steps would be necessary prior to this reference analysis. As discussed in Section 3.9, it was decided that all biomass would be decreased to a particle size less than 850 microns and that the fraction between 180 and 850 microns would be used for the majority of the reference analysis. This fraction is termed “DS” (dry and sieved). The fraction containing the part of the sample less than 180 microns was termed “DF”, meaning “dry fines”, and retained for possible future reference analysis. The first Section of this chapter details the steps involved in getting to the DS and DF fractions.

11.1 Sample Preparation

As described in Section 5.3, the Author was targeting the development of robust and accurate calibration equations for material that is wet and of a heterogeneous particle size. This is in contrast to the vast majority of existing publications regarding the use of NIRS for the characterisation of lignocellulosic materials since these have focused on dry samples that have been ground down to a homogeneous particle size (< 1 mm), i.e. DS-type samples.

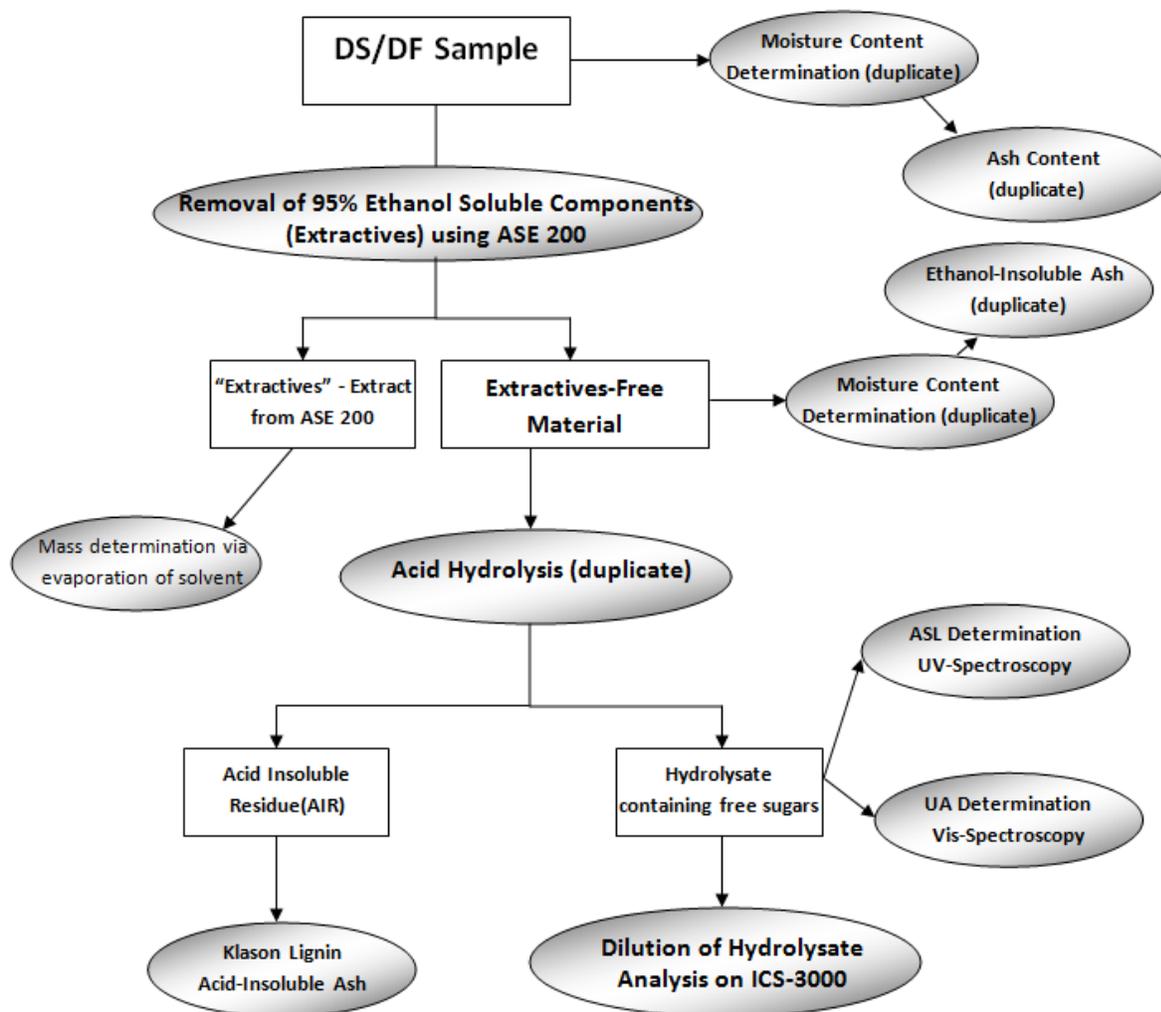


Figure 11-1: An illustration of the sequential method analytical sequence employed for the analysis of structural carbohydrates in prepared biomass samples.

The Author was aware of the risks involved in trying to develop such a calibration due to the interferences of water and the numerous effects of variable particle sizes. Therefore an NIRS analysis methodology was designed to allow the scanning of samples in various other states of preparation, integrating this spectra collection protocol with the sample preparation methodology necessary to obtain samples in the DS form ready for reference analysis. It was considered that this strategy may facilitate the development of more accurate calibrations for the “easier” sample states (e.g. calibrations for a dry feedstock), and it was also considered that multiple calibrations, along the sample-preparation sequence, would allow an understanding to be reached regarding the effects that one change in the

sample state could have. A total of 5 separate NIRS analyses were originally conducted based on the sample in the following states of preparation:

- WU (Wet Unground) – Feedstock not dried and minimally processed.
- DU (Dry Unground) – After WU sample had been air-dried.
- DG (Dry Ground) – The DU sample is comminuted until all is less than 850 μm particle size.
- DS (Dry Sieved) – The DG sample is sieved, and DS is the fraction with a particle size $180 \mu\text{m} < x < 850 \mu\text{m}$. DS is the fraction used for lignocellulosic analysis in the laboratory.
- DF (Dry Fine) - The DG sample is sieved and DF is the fraction with a particle size of less than 180 μm .

It was planned that NIRS calibration would be attempted, for the important lignocellulosic components of biomass, with the WU, DU, DG, and DS spectra. It should be expected that the DS spectra would give the best results since this is a dry material of a homogenous particle size, and is a representation of the biomass fraction that is used directly in the wet chemical analysis.

The DG spectra differ from the DS spectra in that they also include material that has a particle size of less than 180 μm . If an acceptable calibration equation can be developed for the DG spectra, then it means that the DF material is not significantly chemically different from the DS material. A poor calibration for DG will indicate that the chemical composition of DF will need to be taken into account in the DG (and DU and WU) calibrations.

The DU spectra represent the air dried sample with a heterogenous particle size. If an acceptable calibration equation can be developed for the DU spectra then it means that this variation in particle size is not prohibitive for calibration.

The WU spectra represent a wet sample of a heterogeneous particle size. If an acceptable calibration equation can be developed for the WU spectra then it means that this moisture content is not prohibitive for calibration for the lignocellulosic components of interest.

Should only WU and DS scans be taken it would not be possible to determine with confidence the reason for a poor WU NIRS calibration on the basis of a good DS NIRS calibration since there are several factors that vary between these scans (the WU scan includes the presence of DF material that is not present in the DS scan, as well as a heterogeneous particle size, and the presence of moisture).

However, with the range of spectra outlined above a much more detailed understanding can be reached regarding how the quality of the calibration varies with sample preparation.

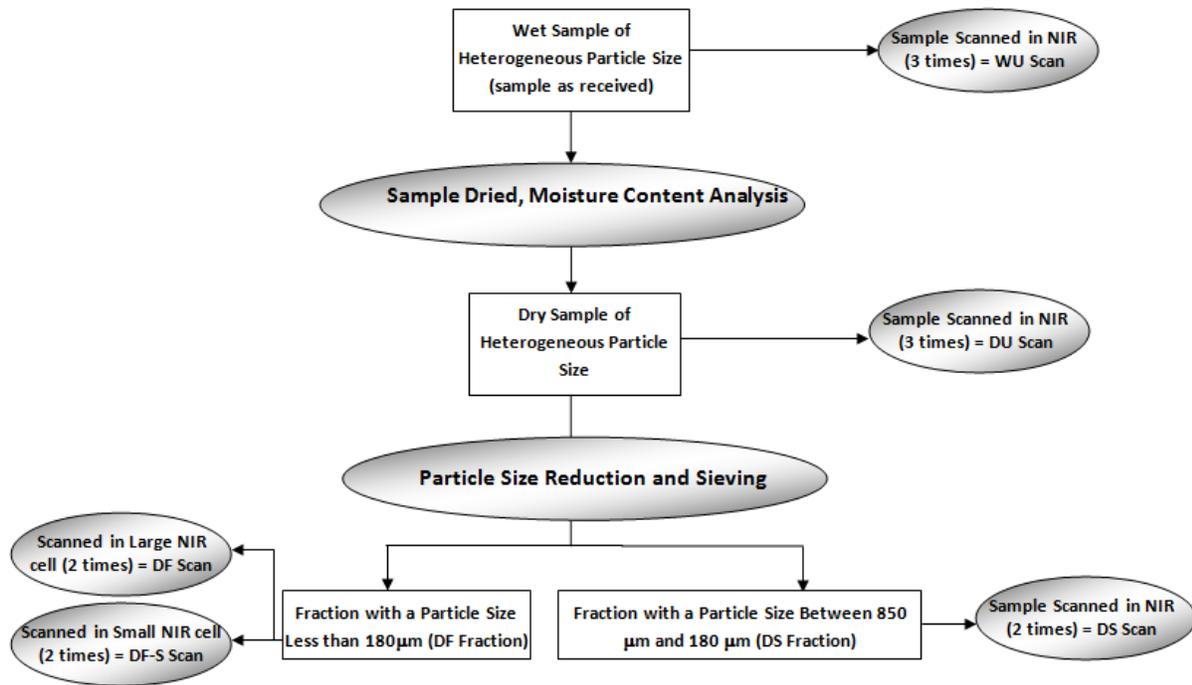


Figure 11-2: The methodology employed in processing biomass, including the collection of NIRS spectra, so that a dry sample of an appropriate particle size can be obtained for subsequent wet chemical analysis.

Figure 11-2 illustrates the steps involved in the preparation of biomass samples. These are also detailed below:

1. A representative sample was taken from storage and placed on a tray in preparation for air drying. If it is in a state that allows direct presentation to the NIRS solid transport cell (for example spent mushroom compost) then proceed to Step 2. Alternatively, the sample is put through a Retsch SM2000 chipper with a 20 mm sieve aperture. The resulting material is a wet sample of a heterogeneous particle size that can easily be presented to the NIRS cell.
2. A subsample was taken from this tray and placed into the solid transport cell of the FOSS XDS NIR and scanned. Then the subsample was returned to the tray and mixed with the rest of the sample. This step was repeated two more times, providing a total of three separate scans of the sample. An average was taken of the three scans to provide a single scan for the sample. This was labelled WU-A.

3. In some cases the moisture content of the sample was determined via the drying, at 105°C, of a subsample taken from the tray (moisture analyses were usually carried out in duplicate).
4. The tray was left to air dry in the laboratory. Each day the sample was moved around on the tray in order to ensure thorough drying. The tray was also weighed each day to monitor the moisture loss. The drying phase was considered complete when the mass loss between 2 consecutive days was less than 0.5%.
5. Step 2 was repeated on this air-dried sample, with three DU scans of each sample taken. The scans were averaged to produce a single scan for the sample. This was labelled DU-A.
6. The later analytical phases involve relatively small subsamples of the feedstock; however, it is important that these are representative of the whole sample. Hence comminution of the biomass, to ensure a more homogeneous particle size distribution (DS), was necessary. Various methods were employed for this (chipping with the Retsch SM2000 chipper using a smaller sieve aperture; using a FOSS Cyclotec mill; using a combination of the two pieces of equipment; hand grinding with a mortar and pestle) depending on the sample in question. The target was to maximise the DS to DF ratio.
7. The combined DS/DF fraction is termed DG and is scanned twice on the FOSS XDS NIR. The two scans were averaged to produce a single scan for the sample; this was labelled DG-A.
8. The DG fraction was then sieved to provide separate DS and DF fractions which were weighed. The DS fraction was scanned on the FOSS XDS NIR (2 scans per sample). The two scans were averaged to produce a single scan for the sample; this was labelled DS-A.
9. For some samples two NIR spectra were collected for the DF fraction, with the average labelled as DF-A.
10. The DS and DF fractions are stored in air-tight containers for future analysis.

It should be pointed out that the samples are all scanned inside the “coarse sample cell” described in Section 5.3.2.3. If there was insufficient DF material to cover the window of this cell then the DF sample was put in a smaller circular cell and scanned twice with the average labelled as DF-S-A.

It was discovered during the development of calibration equations for DS Miscanthus samples that, when calibrations that were based on the average spectra of 2 DS scans (i.e. DS-A) were applied back to the original non-averaged spectra, there was, in some cases, quite a difference between the predicted compositions of the two scans. The corresponding differences between the two DF scans were much

lower, however. The Author came to the conclusion that there was a bias in the way in which the DS samples were being presented to the NIR unit.

Standard practice for all fractions (WU, DU, DG, DS, DF) involved spreading out the sample on a bench so that there was no bias in the collection of the subsample that was presented to the cell. However, this sample was then simply released by hand into the cell and, once sufficient material was inside, the lid was closed and the cell scanned. After the observation of a significant variation between duplicate scans, the Author examined the window of the cell after the sample had been introduced using the standard method outlined. It was found that there was a visually-apparent difference in particle size when looking at the window from below compared to looking at the cell from the top down. Particles at the window were smaller and those at the top were larger, representing the effect that gravity would have in allowing the smaller particles to filter through the available gaps until they met the base. The degree to which this would happen would depend on the way in which the sample was put in the cell; hence the variability in predictions. Typically, the appearance of the sample in the cell from the top was much more representative of the sample as whole than was the appearance of the sample from the bottom.

It is important to consider how the sample is analysed in the NIRS system. The sample cell can contain approximately 430 cm^3 of sample. However, this sample is scanned from below over an area of approximately 82 cm^2 . If the sample is presented in a consistent manner, with no air gaps present, then the light from the system will only penetrate through a few millimetres of the sample at most. Hence, the material at the base of the cell is the part of the sample that will contribute most to the resulting spectra and to any calibration equations that are ultimately developed. There could therefore be a bias to the NIRS analysis of the small particles compared with the reference analysis of the DS sample as a whole. Hence, a new sample presentation method was employed involving the following steps.

1. Place the sample in the cell as normal and close the lid.
2. Turn the cell upside down while holding the lid.
3. Pull back the lid so that there is space in the enclosed area.
4. Shake the contents of the cell up and down. Due to gravity the smaller particles will fall away from the cell window.
5. Push the lid back so that there is no space for particle movement. Observe that the particle size distribution appears more representative, if not repeat steps 3 and 4.
6. Turn the cell back to normal orientation and scan.

This new scanning method could also be applied, in the same manner, to the DG and DU fractions (for which the same small-particle bias was also observed). DS samples scanned in the new method were labelled as “DT”, DG samples as “DH” and DU samples as “DV”. The particle size bias was not an issue in WU samples since the moisture present in the sample acts as a “glue” binding the smaller particles to the larger ones. That meant that the sample as presented to the cell window was more representative of the whole sample. The new method was primarily used for *Miscanthus* samples and its effectiveness will be discussed in Section 15.

It should also be noted, regarding the preparation methodology, that in some cases (e.g. straw samples and dry *Miscanthus* samples that had been mailed to the Author from abroad) the samples were already dry. This meant that, after any appropriate comminution to allow the sample to fit in the NIR cell, the DU scan was the first collected in the sequence.

A single coarse sample cell and several smaller circular sample cells were supplied with the XDS system. It quickly became apparent that one large cell was insufficient given the amount of work being carried out in the laboratory, particularly if the user wanted to scan wet samples and dry samples in quick succession. The Author therefore purchased another coarse sample cell. Each was given a separate label (A or B) with one being used for wet samples and the other for dry samples. The small cells were also given labels (C, E, or F). In January 2011 the window of cell B was broken and another window was purchased and installed in the cell and the cell was renamed as “D”. The reasons for labelling these cells differently, and for making reference to them when saving spectra, was that there could potentially be path-length differences between these cells that could result in slightly different spectra. Keeping reference to the cells used would allow tests that could be made if this effect was suspected to be the reason for calibrations of poor accuracy. A comparison, involving 10 scans and several quantitative calibrations, however, found no significant differences between cell A and cell D.

11.2 NIRS Conditions

Section 5.3.2.3 has detailed the specification of the FOSS XDS NIR instrument used in the Carbolea laboratories as well as the specific conditions (data collection method, DCM) used for the collection of

spectra. The relevant Sections in subsequent chapters will detail the development of calibrations for the lignocellulosic properties of interest for a particular data-set. However, reference here will be made to the two scenarios used, for Miscanthus samples only, for the provision of reference data to the WU, DU, DV, DG, and DH products.

In the first scenario, the data obtained for the DS fraction of samples is used directly for these products. In the second scenario, which occurs only if there are DS and DF analytical data for the sample in question, a weighted sample composition is determined according to the compositions of the DS and DF samples and the relative proportion that these contribute to the total weight of the DG sample. Calibrations using this scenario have the DG subscript associated with the product type, e.g. WU_{DG} .

In the third scenario, which applies when there are no reference analytical data for the DF fraction of the sample, the DF composition is determined from the DF spectra of the sample using the DF calibrations that have been developed. A weighted sample composition is then calculated as in the second scenario. Calibrations using this scenario have the DGP subscript (for dry ground predicted) associated with the product type, e.g. WU_{DGP} .

11.3 Moisture and Ash Analysis

The analysis of moisture involved placing a subsample (between 0.2 and 0.5 g) in a crucible of known weight and drying in an oven overnight at 105 °C. This usually occurred in duplicate. In all cases an NIR spectrum of the sample was collected, using the small circular cell, at the same time as moisture analysis. If there was only sufficient sample for the subsequent steps (e.g. extraction, hydrolysis, elemental analysis) then a spectrum was collected but no moisture content analysis was made. It was planned that the moisture content of the sample at that time could be predicted from the spectrum using a moisture calibration that had been developed with other samples/spectra.

Ashing took place after the dry weight of the sample was known and involved the following heating program in the Nabertherm L-240H1SN muffle furnace:

Ramp from room temperature to 105 °C

Hold at 105°C for 12 minutes

Ramp to 250°C at 10°C/minute
Hold at 250°C for 30 minutes
Ramp to 575°C at 20°C/ minute
Hold at 575°C for 180 minutes
Allow temperature to drop to 105°C
Hold at 105°C until samples are removed

The samples were weighed after ashing and then the ash removed from the crucible and the empty crucible weight recorded again. It was found that using this new empty weight to determine the ash content gave more precise results. Proper desiccator practice was employed in all stages of analysis.

11.4 Extractives Content

The steps involved in the removal of 95% ethanol-soluble extractives from DS/DF samples are listed below:

1. Samples were removed from containers and given time to equilibrate with the humidity of the air.
2. An NIR scan, using the small cell, is taken of the sample prior to moisture content determination.
3. Moisture content determination as outlined in Section 11.3, preferably in duplicate, but one analysis or none may take place if the sample is limited. Ash determination follows.
4. An 11 ml capacity ASE (Accelerated Solvent Extraction) cell is filled with a recorded weight of the sample. There may be a single cell or duplicates or triplicates, depending on the amount of sample available.
5. Once the above steps have been repeated for all the other samples, weighed collection vials are placed in the lower wheel of the ASE 200 and the following program is used for all cells:

Pressure:	1500 PSI
Temperature:	100 °C
Preheat Time:	0 minutes
Heat Time:	5 minutes

Static Time:	7minutes
Flush Volume:	150%
Purge Time:	120 seconds
Static Cycles:	3

6. After completion of the sequence, a glass petri-dish of known weight was taken and the remaining biomass from the extraction cell was transferred to this dish.
7. After 2 days the petri-dish was weighed again and its contents scanned in the NIR with the small circular cell. Then the moisture content of a subsample of the extracted biomass was determined (in duplicate). However, if all of the sample needed to be retained only the NIR scan took place and there was no moisture analysis. The remaining extracted biomass was stored for the hydrolysis stage (see Section 11.5). The weight of extractives was determined as the mass loss in the biomass sample due to extraction in the ASE-200 (corrected for moisture).

Prior to the purchase of the Zymark Turbovap II (Section 3.4.4) and its accessory for handling ASE collection vials, the contents of the collection vials were discarded. However, once this system was available the collection vials were instead put inside and the solvent removed. Then the collection vials were placed in a 40°C oven and weighed each day until the loss in weight between days was not greater than 1 mg. That allowed for the extractives content to be determined directly, as well as indirectly. However, it often took many days for the weights of the collection vials to stabilise, and there were often discrepancies between the indirect and direct extractives-content values.

Initially there were only 6 of the 11 ml ASE cells available and this limited the amount of samples that could be analysed each day. Table 11-1 describes the different variants of the above method that were used for extractives removal. It assumes that there is a plentiful supply of the sample available. Method E-3 involved the use of a moisture calibration to predict the moisture content of the material in the third petri-dish. The FE method was employed as a very late experiment towards the end of this study when the Dionex Solvent Controller (Section 3.4.4) was available. It involved the water extraction (2 cells), ethanol extraction (2 cells) and water then ethanol extraction (2 cells) of a sample. For the water then ethanol extraction the water extract was collected in one vial and the subsequent ethanol extract in another. The same ASE program as described above was used for water extraction. The solvents were not evaporated from the extracts but instead the extracts were weighed and then stored in the freezer for possible future analysis (for example for sugars via HPAEC-PAD). The vials corresponding to the

water extract of the water-ethanol extraction were not kept. Hence, in the FE method there was no “direct” measurement of the extractives.

Table 11-1: A summary of the various methods used for the removal of extractives from samples. The “samples per batch” category describes the modal number of samples per batch although the number did vary on occasion.

Method	Used Between	# Cells per Sample	Samples per Batch	“Direct” Extractives?	Comment
E-1	10/3/09→6/10/09	2	3	No	The extract was discarded.
E-2	6/10/09→24/3/10	2	6	Yes	
E-3	24/3/10→Present	3	6	Yes	The third petridish was weighed and an NIR scan taken but no moisture analysis.
FE	7/4/11→Present	6	3	No	2 cells water, 2 cells 95% ethanol, 2 cells water and then 95% ethanol (see above)

11.5 Hydrolysis Procedure

The steps involved in the acid hydrolysis of extracted samples are outlined below:

1. Samples are removed from containers and given time to equilibrate with the humidity of the air.
2. An NIR scan, using the small cell, is taken of the sample prior to moisture content determination.
3. Moisture content is determined as outlined in Section 11.3, preferably in duplicate, but one analysis or none may take place if the sample is limited. Ash determination follows.
4. Approximately 300 mg, exact weight noted, of sample is added to a pressure tube.
5. 3.00 mL of 72% H₂SO₄ is added by means of an automatic titrator, the weight of the acid added is noted.
6. The sample is mixed thoroughly with the acid using a glass rod, care is taken that no sample stays adherent to the sides of the tube, but instead stays in contact with the acid.
7. The tube is transferred to a water bath that is maintained at 30°C.
8. Steps 4-7 are repeated for the duplicate, and steps 2-7 for subsequent samples.
9. Every 5/10 minutes the glass rod for each pressure tube is stirred so that the acid reaches all parts of the sample and complete hydrolysis occurs. **This is a crucial step.**

10. Exactly one hour after it is placed in the water bath the pressure tube is removed and placed on a scales and 84 mL of water added (with the weight of water added noted). Any acid/sample on the rod is removed from the rod at this point using this water.
11. A lid is screwed on the tube and the tube is inverted several times to ensure thorough mixing of the acid.
12. Sugar recovery solution (SRS, see Section 3.1.2) tubes:
 - a. 348 μl of 72% H_2SO_4 is added to a test tube containing a solution containing a known weight (approximately 10 g) of a sugar standard. This standard should be of a similar sugar composition to that expected of the samples being analysed. The acid and sugar solution are thoroughly mixed. This is repeated for 2/3 tubes (2 in early experiments batches and 3 in later experiments).
 - b. The sugar-acid mixture is transferred to a pressure tube which is then sealed.
13. All SRS and sample pressure tubes are placed in an autoclave which is run at 121°C for 60 minutes.
14. The tubes can be removed from the autoclave when the temperature drops to under 80°C, and are left (closed) in the lab until they reach room temperature.
15. The hydrolysates are filtered (using vacuum suction) through filter crucibles of known weight and the resulting filtrate is stored.
16. Any residual solids are washed out from the tube using deionised water until all the residue resides on the filter crucible. This filter crucible is then dried at 105°C overnight and then weighed to determine the Acid Insoluble Residue (AIR) content. The filter crucible is then ashed under the conditions outlined in Section 11.3 to determine the acid-insoluble ash (AIA) and the Klason lignin is determined as AIR minus AIA.
17. Acid Soluble lignin analysis - The hydrolysate from step 15 is placed in a 1 cm path-length (3 mL volume) quartz cuvette and diluted with water so that the UV-absorbance is within a linear region (considered to be between 0.7 and 1.0 absorbance units at 240 nm). This is done twice for each hydrolysate (i.e. four scans will be taken for each sample, 2 for each duplicate).

The subsequent steps involving the dilution (5 times using a solution of known internal standard concentration) and the chromatographic analysis are outlined in Section 4.6.2.2. All hydrolysates and diluted hydrolysates corresponding to batches that were considered to have “good” hydrolysis results

(i.e. no large differences between the duplicates) are still kept frozen in the laboratory, meaning that they can potentially be used in future experiments.

The whole period from starting the autoclave until the solutions in the pressure tubes have reached room temperature takes approximately 2 ½ hours. This is a critical parameter because the time involved will have an effect on the dynamics of the second-stage-hydrolysis and the associated loss of sugars to degradation products such as furfural, HMF etc. The purpose of the SRS solutions is to correct for any local between-batch variations in autoclave conditions. However, ideally the hydrolysis conditions should be relatively stable over time. Fortunately the autoclave operated quite consistently over its whole operation (May 2008 to May 2011) as illustrated in Section 15.2.1, for example.

Uronic acid analysis, when carried out, followed the method outlined in Section 3.2.4. This involved duplicates of each hydrolysate being put through the UA analysis method and, as with the ASL analysis, each of these was analysed twice with the UV-Vis spectrophotometer. The UA content was then determined according to the following formula:

$$UA = \frac{W_u \times F_u \times F_c}{S} \quad (11.1)$$

Where S= weight (dry matter, mg) of original sample; W_u = weight (mg) of galacturonic acid monohydrate/100 mL hydrolysate, obtained from calibration curve; F_u = factor for recalculation of galacturonic acid to polysaccharide residues (0.830); and F_c = compensation factor to adjust for greater degradation of free galacturonic acid as opposed to that of polygalacturonate, under conditions of uronic acid calibration ($F_c = 0.81$ was used).

11.6 Elemental Analysis

The following method was employed for the CHNS analysis of samples. This procedure usually involved characterising both the DS and DF fractions of the same sample in the same batch.

1. Samples are removed from containers and given time to equilibrate with the humidity of the air.
2. An NIR scan, using the small cell, is taken of the sample prior to moisture content determination. Separate Products/calibrations are used for DS and DF samples.

3. Moisture content is determined as outlined in Section 11.3, preferably in duplicate, but one analysis or none may take place if sample is limited. Ash determination follows.
4. Approximately 20 mg of the sample is placed in a tin boat which is then wrapped up to prevent sample leakage. A 5 decimal place (i.e 0.1 mg) analytical balance was used in this step.
5. Standards were prepared in the same way as for the samples. A separate product was used to store the NIR spectra of the standards.
6. The wrapped up samples were placed in the wheel of the vario EL cube and the sequence started using the following conditions:

Combustion Tube Temperature:	1150°C
Reduction Tube Temperature:	850°C
Oxygen Dosing Time:	260 seconds
Desorption Temperature CO ₂ Column:	260°C
Desorption Temperature H ₂ O Column:	170°C
Desorption Temperature SO ₂ Column:	280°C

The sample sequence outlined in Figure 11-3 was typically used for each batch.

For the early experiments with elemental analysis the 5 decimal place analytical balance had not been purchased and instead a 4 decimal place balance was used, and only 5 mg of sample was prepared. The Author calculated that this would lead to an error of 2%. This meant that, for the analysis of carbon, which has a content of around 50% for some samples, there would be no justification in reporting any of the digits following the decimal point. Clearly, having such a large error at the very start of the analysis simply due to the inadequacy of the weighing scales is far from ideal and it was for that reason that the more sensitive analytical balance was bought by the Author.

Another way to reduce the weighing error would be to use a greater quantity of material for the analysis of each sample. This would also help to reduce any effects that may come from sample inhomogeneity, which is an issue for the DS samples but much less so for the DF samples.

However, this was found to give rise to an issue with the sample chosen for use as a standard for determining the daily factor (see Section 3.6). The only standard that was available initially was sulphanimide, C₆H₈N₂O₂S. This has CHNS contents of 41.8%, 4.6%, 16.3%, and 18.6%, respectively. These nitrogen and sulphur contents are far in excess of what would be expected in most of the biomass samples of interest in this study. As with the use of sugar standards in chromatography/hydrolysis, it

would be better to use a standard that is more similar to the samples being analysed. However, a more significant problem was presented by the fact that, at sample weights over 10 mg, the N and S contents of sulphanimide overloaded the adsorption columns and prevented the accurate calculation of daily factors. It was not an option to use differing weights for the standard as compared with the samples, so instead an alternative standard was sought, with CHNS proportions more similar to those of the samples.

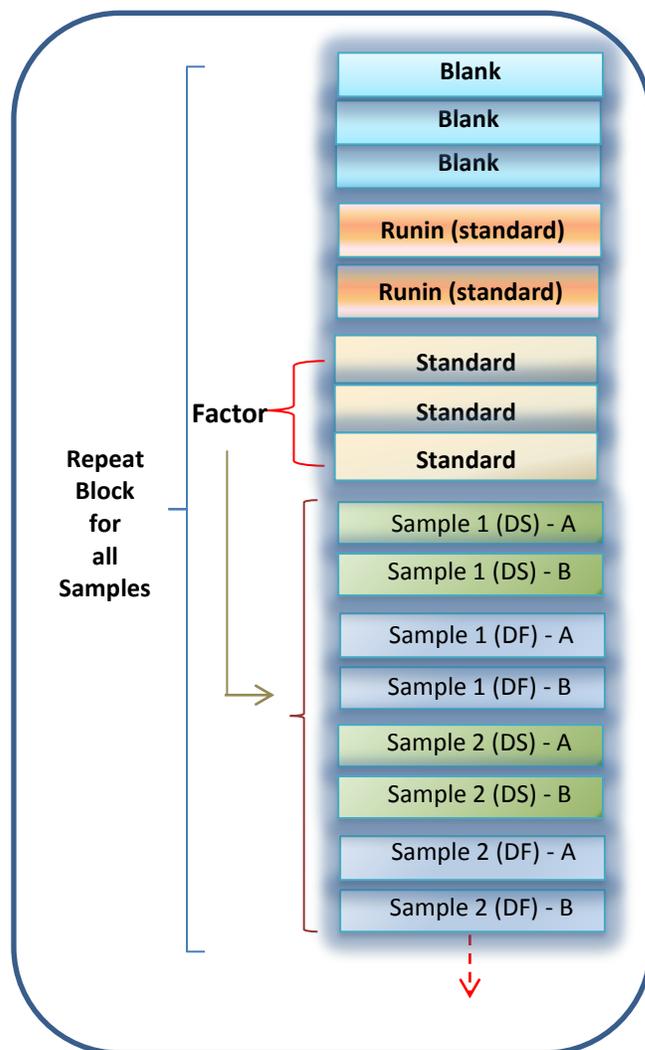


Figure 11-3: The sample sequence used in elemental analysis batches.

It was considered that a commercial standard was not necessary and that instead a cheap consistent material could be used with its elemental contents determined from multiple replicates at lower

weights when using sulphanilamide as a standard. Several food products, including commercial flours and seeds, were bought and tested for their suitability. However, it was difficult to find a suitable material that contained some sulphur. Finally, Cadbury's Cocoa powder was chosen as a suitable material. Its CHNS compositions was determined as 53.5%, 7.3%, 4.0%, and 0.2%, respectively. This standard was kept in an air-tight container and its moisture content measured each day prior to its use as a batch standard.

11.7 Precision Criteria for Reference Analytical Methods

Criteria were set for the wet-chemical analysis of Miscanthus samples so that the precision of the reference methods would be high. These were based on the standard deviation of the duplicate (SDD), i.e. the standard deviation of the constituent values for each of the two replicates. If the SDD of a sample was above the assigned limit for that sample then the reference analysis should be repeated. In some cases analytical data were retained even if the SDD breached the limit. This usually occurred if there was an insufficient amount of material remaining to allow reanalysis of the sample and the inclusion of this sample in NIRS models was considered important. However, no samples with excessive SDD values were kept in models. The SDD limits are presented in Table 11-2, according to constituent.

Table 11-2: Standard deviation of duplicate (SDD) limits for a range of constituents.

Constituent	SDD Limit (%)	Notes
Moisture content (at hydrolysis and extraction stages)	0.20	If this limit was breached then an NIRS calibration was instead used to predict the moisture of the sample.
Ash	0.20	
Klason Lignin (KL)	0.25	
Acid Insoluble Residue (AIR)	0.25	
Acid Soluble Lignin (ASL)	0.20	
Extractives	0.25	
Glucose	0.30	The SDD for glucose was used to represent the precision of the hydrolysis/chromatography analysis for all sugars.
Carbon	0.30	The SDD for carbon was used to represent the precision of the elemental analysis for all the elements.

12 Analysis of Sugarcane Bagasse and Development of Quantitative NIRS Calibrations

This Section will describe the reference analytical and NIRS calibration development work undertaken by the Author on 47 sugarcane bagasse (SB) samples that originated from North Queensland, Australia. These samples were analysed via NIRS, and processed to an appropriate particle size for reference analytical methods, in the laboratories of BSES Limited in Brisbane, Australia. A subset of 30 of these samples was sent to UL at a later date for wet-chemical analysis. Following this analysis, NIRS calibrations for the quantitative prediction of a range of lignocellulosic properties of these samples were then developed for several NIR data-sets consisting of the samples at various stages in the processing procedure. Some of the results obtained are presented in this Chapter with the rest presented in Appendix C.

12.1 Background on Sugarcane Bagasse

12.1.1 Sugarcane

Sugarcane is a close relative of *Miscanthus* (Section 10). It is classified under the Poaceae (grasses) family and so much of the material discussed in Section 10 is relevant to this feedstock. An extremely large number of sugarcane varieties have been developed in order to improve yields, sugar contents, pest resistance, and to optimise the crop towards the region in which it is grown.

Sugarcane is a semi-perennial C_4 plant with the plantation cycle depending on the environmental properties of the region of production. For example, in Brazil the cycle usually lasts six years enabling five cuts to be taken (BNDES, 2008). The crop is typically grown for the production of sugar, which can be extracted from the plant and refined for sale on the open market or used for the production of bioethanol. In recent years, however, there has also been research into the development of lower sugar content but higher fibre and higher biomass yielding sugarcane varieties, often referred to as energy cane (Mark et al., 2009).

Around 20 million hectares of sugarcane were planted and 1.3 billion tonnes of cane produced in the 2006/2007 season with Brazil being the greatest contributing country with plantations covering 7 million hectares (BNDES, 2008). According to a report by Centro de Tecnologia Canavieira (CTC), in Brazil the average productivity of the plant is 70 wet tonnes per hectare per annum (CTC, 2005).

Sugarcane ideally requires two different growing seasons in order to attain maximal yields and high sugar production. A warm and wet season allows the growth of the crop and then a cooler and drier season will enable the maturation of the plant and the accumulation of sucrose in the stems (BNDES, 2008). The first cut of the crop typically occurs a year or more after planting. Following cutting the stems will grow back and are harvested each year, although yields tend to decline with the stage of the cycle.

There are a variety of important compositional characteristics of sugarcane. The Pol (short for polarisation) value is the apparent sucrose content, typically expressed as a percentage of the total dry mass of the cane. In varieties of sugarcane grown for sugar production the pol content of the sugarcane juice varies from 8 to 15% (Tewari et al., 2003). Fibre is the term given to the sugarcane bagasse (see Section 12.1.2) which is considered to be the total of the insoluble solids derived from cane after the milling stage. CTC (2005) reported a 14.2% pol content and a 12.7% fibre content for sugarcane produced in a plantation in Sao Paulo state, which achieved a yield of 87.1 wet tonnes of cane per hectare.

Only the stem of the crop contains sufficient sucrose to warrant its processing in sugar mills. The other parts of the crop, principally the leaves and tops, are often referred to as the “trash” of the plant. It has been estimated that for each tonne of harvested cane 140 kg of trash will exist (BNDES, 2008). Traditionally, manual harvesting practices of sugarcane have been carried out, and these still take place in developing countries, although they are being gradually phased out in favour of mechanical harvesting techniques (for example, the Brazilian government has targeted 2020 as a date when all harvesting will be mechanical). Manual harvesting practices require the crop to be burnt prior to the chopping of the stems by workers. This will burn the leaves of the crop but not damage the stems and roots. Mechanical harvesting usually involves a combine harvester which cuts the crop at the base of the stalk, strips off the leaves, and chops the cane into reasonably consistent lengths which are then blown into a transporter traveling alongside. The “trash” is blown back onto the field.

12.1.2 Production of Sugarcane Bagasse

Following harvesting the cane is transported to a sugar mill where it needs to be processed quickly in order to avoid the loss of sugar. This means that sugar mills only operate during the harvesting season, between April and December in the Brazilian state of Sao Paulo, and between June and December in Australia.

Figure 12-1 shows a simplified representation of a sugar mill (Pennington and Baker, 1990). The process firstly involves washing the cane to remove excess soil and impurities and it is then crushed/shredded/chopped before proceeding to a series of mills that contain three to five rollers. Hot water, or a combination of hot water and impure sugarcane juice, are sprayed onto the crushed cane after it leaves each mill. This is done in order to extract the juice from the cane. Lime is then added to the juice and it proceeds to a clarifier where soluble and insoluble impurities coagulate and settle at the bottom of the tank. This residue is filtered to produce a filter cake that can be used as an animal feed supplement and a fertiliser. Subsequent stages involve the clarified juice being evaporated to produce a syrup that is about 65% solids and 35% water (EPA, 1995). This syrup is then clarified and the sugar starts to crystallise in vacuum pans. The mixture of syrup and crystals is known as massecuite and these are separated in a centrifuge device. The liquor component here is referred to as molasses and it returns to a vacuum pan for the production of more crystals which form part of a second masscuite which is again put through a centrifuge. A third crystal production cycle can then occur but the molasses that follow from this are of low sugar content and cannot be processed further. This type of molasses is known as blackstrap molasses (EPA, 1995) and has use as an animal feed and as a feedstock for ethanol production.

All of the above stages are not necessary in sugar mills that produce bioethanol from the feedstock. The process involves the fermentation of the sugarcane juice, or of a mixture of the juice and molasses. Typically the juice is clarified and partially evaporated in order to increase the sugar concentration prior to fermentation and it is at this stage when molasses can be added. Yeasts are used for the fermentation and the resulting product is a “wine” with an ethanol concentration of between 7 and 10% (BNDES, 2008). After recovery, via centrifugation, of the yeasts, ethanol is produced via the distillation of the “wine” and the subsequent dehydration of the hydrated bioethanol.

Hence, if sucrose is efficiently removed the bagasse will represent a predominately lignocellulosic feedstock. Taking the example of Brazilian sugar cane, one wet tonne of sugarcane that is processed at the mill will yield approximately 280 kg (wet) of SB (Cardona et al., 2010). Typically SB has a moisture content of between 45 and 55% on a wet basis (EPA, 1995). It can be seen therefore that SB is an extremely significant output of the sugar mill. Taking again the example of Brazil, it was estimated that 160 million wet tonnes of SB were produced in 2008 (Cunha et al., 2011).

The primary use for this resource is as a heat and steam provider to satisfy the energy needs of the sugar production or fermentation processes. It has been estimated that a minimum of 50% of the bagasse is required for this (Edwards, 1991). In many cases the surplus bagasse represents a problematic waste that could lead to safety issues (e.g. spontaneous combustion) if stored for a long period of time. For that reason some mills deliberately burn the SB at a low efficiency in order that more of it will be consumed for energy production (Lavarack et al., 2002). It has been estimated that the use of SB in boilers could be reduced by up to 36% if more efficient combustion schemes are employed (Lavarack et al., 2002). Regarding SB in Australia, it has been estimated that there are potentially 2 million tonnes per year that could be available for use after the needs for process heat and steam in sugar mills have been considered (Briody, 1997).

There has been a significant amount of research into the utilisation of SB in biorefining technologies that may produce saleable chemicals from the polysaccharides (Lavarack et al., 2002, Jackson de Moraes Rocha et al., 2011, Cardona et al., 2010, Rodríguez-Chong et al., 2004) or bio-oils via the pyrolysis of this residue (Dynamotive Energy Systems Corporation, 2001, Cunha et al., 2011). The DIBANET project (Section 18), which is currently operational and led by the University of Limerick, is focussed on the production of levulinic acid, furfural and formic acid from *Miscanthus* (sourced in Ireland) and SB and sugarcane trash (both sourced from Brazil). The yields of each of these platform chemicals will depend on the process conditions and on the relative amounts of polysaccharide sugars in the feedstocks.

12.1.3 Compositional Data for Sugarcane Bagasse

Table 12-1 summarises the compositional values of SB found by the Author in his literature review. This Table includes the lignocellulosic compositional values, obtained by Lavarack *et al.* (2002), for SB obtained from Victoria Mill, which is located near Ingham on the North Queensland (Australian) coast.

This mill is located only about 180 km away from the Mulgrave Mill from where the samples analysed by the Author (see Section 12.2) were obtained.

A paper involving the analysis of the hemicellulose in SB was published by du Toit *et al.* (1984). The removal of hemicellulose involved a 24 hour treatment with 4% NaOH, the filtrate was then adjusted to pH 5.0 with 50% acetic acid and the mixture was centrifuged and the precipitate, hemicellulose A, collected and freeze dried. The hemicellulose B fraction was precipitated from the supernatant by adding it to absolute ethanol. The authors obtained a total yield of 32.10% by mass hemicellulose, with 17.75% being hemicellulose A and 14.35% being hemicellulose B. These different hemicellulose fractions were then hydrolysed; this required 2% HCl for the “A” fraction and a stronger acid, 5% H₂SO₄, for the “B” fraction. Table 12-2 contains a summary of the compositions of the different monosaccharides determined from these hydrolysates as well as the composition of the residue left after extraction. The hydrolysis conditions employed for this residue were not considered to be strong enough to hydrolyse the crystalline cellulose; hence the sugars were considered to be from residual hemicelluloses that had not been extracted.

Table 12-1: Compositional data (% of dry matter) for sugarcane bagasse samples obtained from various locations.

Reference	Source	Cl.	Hc.	Glu.	Xyl.	Ara.	Gal.	Man.	Extr.	KL	ASL	Ash
(Yu and Stahl, 2008)	Hawaii	42.9			25.2	1.4				19.0	6.6	4.9
(Lavarack et al., 2002)	Australia									22.3		
(Jackson de Moraes Rocha et al., 2011)	Brazil	45.5	27.0						4.3*b	21.1*a		2.2
(Boussarsar et al., 2009)	Reunion	45	26							20		2.1
(Maeda et al., 2011)	Brazil	34.1	29.6							19.4*a		7.9
(Gámez et al., 2006)	Mexico			38.9	20.6	5.6				23.9		
(Laser et al., 2002)	Hawaii			44	26	2				23		
(Neureiter et al., 2002)	Thailand			40.2	22.5	2.0	1.4	0.5		25.2		
(Ewanick and Bura, 2011)	Brazil			41.3	21.8	1.8	0.5	0.3		20.5	2.9	4.1
(Alves et al., 2010)	Argentina *c			43.1	25.0	1.5	0.4	0.3		23.2*a		2.5
(Sun et al., 2004b)	China	43.6	33.5							18.1*a		2.3
(Sindhu et al., 2010)	India		27	34					17			4

Cl. = cellulose; Hc. = hemicellulose; Glu. = glucan; Xyl. = xylan; Ara. = arabinan; Gal. = galactan; Man. = mannan; Extr. = extractives; KL = Klason lignin; ASL = acid soluble lignin; *a – total lignin; *b – 95% ethanol-soluble extractives; *c – the data for this sample are presented on an (ethanol-toluene) extractives-free basis (this sample also had 1.2% uronic acids, and a 3% acetyl content).

What is particularly interesting about the data in Table 12-2 is that there is a significant amount of glucose present. While it is possible that some of this may have come from hydrolysis of the amorphous

regions of the cellulose that may have been extracted in the procedure or from starch present in the SB, it does appear that glucose is also an important constituent sugar in hemicellulose. This indicates that the glucose liberated under stronger hydrolysis conditions (see Section 11.5) will come from both the cellulose and hemicellulose/starch fractions and classifying total glucose as equivalent to total cellulose would be inappropriate. Also notable in Table 12-2 is the higher relative proportion of galactose, glucose and arabinose in hemicellulose A compared with hemicellulose B. That would indicate that the xylan of the former has a higher degree of branching and that this branching involves greater contributions from galactose and glucose.

Table 12-2: The proportions of the different monosaccharides obtained upon hydrolysis of different hemicellulose fractions: hemicellulose A, H(A), hemicellulose B, H(B), and the solid residue after extraction. These proportions are expressed as a percentage of the extract/residue and as a percentage of the original dry mass of the SB (%SB).

Fraction	Arabinose	Galactose	Xylose	Glucose	Total
H(A) (%)	3.98	0.37	54.66	7.79	66.80
H(B) (%)	16.76	5.32	43.69	19.18	84.95
Residue (%)	2.67	0.80	11.51	4.37	19.35
H(A) (% SB)	0.74	0.07	10.08	1.44	12.33
H(B) (% SB)	2.50	0.99	8.13	3.57	15.19
Residue (% SB)	1.62	0.48	6.96	2.64	11.70

A more recent paper concerning the analysis of the hemicelluloses in SB was published by Sun *et al.* (2004a). This involved the sequential extraction of SB that had previously been ultrasonically treated in distilled water at 55°C for 40 minutes. The yields of hemicelluloses and lignin in each of these sequential treatments are shown in Table 12-3 and the contents of sugars and uronic acids in these fractions are shown in Table 12-4. Table 12-5 takes the data in Table 12-4 and corrects the values to a whole-mass basis using the data in Table 12-3. Notable points from Table 12-4 and Table 12-5 again include the presence of glucose in significant quantities, particularly in the water soluble fraction which also contained contributions from galactose and mannose. The sugar proportions in the alkali and alkaline peroxide treatments, while similar to each other, were significantly different from those in the water extracted fraction. It was therefore concluded (Sun *et al.*, 2004a) that these fractions were mainly composed of glucuronoarabinoxylans whereas the arabinoxylans extracted with water were more highly substituted.

Table 12-3: The yield of hemicelluloses and lignin (% of dry matter) solubilised during the successive treatments of dewaxed bagasse with distilled water under ultrasonic treatment, 0.5 M NaOH, various concentrations of hydrogen peroxide, and 2 M NaOH at 55°C for 2 h. Taken from (Sun et al., 2004a)

Component	WS ₁ ^a	AS ₁ ^b	H ₂ O ₂ Concentration (%) ^c					AS ₂ ^d	Total
			0.5	1.0	1.5	2.0	3.0		
Hemicelluloses	4.5	12.8	4.0	1.5	3.0	0.8	0.4	3.8	30.8
Lignin	0.6	10.8	2.8	0.8	1.0	0.3	0.1	0.1	16.5
Residue	94.2	70.0	63.1	60.9	56.6	55.5	55.0	44.7	

a - Water-soluble hemicellulose and lignin fraction obtained by ultrasonic treatment of the dewaxed bagasse in distilled water at 55°C for 40min, then stirring for 80min at 55°C; b - Alkali-soluble hemicelluloses and lignin obtained by extraction with 0.5M NaOH at 55°C for 2h from the ultrasonic irradiated and water treated bagasse; c - the fractions obtained by successive extractions of the 0.5 M NaOH treated bagasse with different concentrations of H₂O₂ at 55°C for 2h at pH 11.5.; d - the fraction obtained by extraction with 2 M NaOH at 55°C for 2h from the 3.0% H₂O₂ treated bagasse.

Table 12-4: The content of sugars and uronic acids (% dry weight) in the isolated hemicellulose fractions. Taken from (Sun et al., 2004a)

Component	WS ₁ ^a	AS ₁ ^b	H ₂ O ₂ Concentration (%) ^c					AS ₂ ^d
			0.5	1.0	1.5	2.0	3.0	
Rhamnose	1.25	0.42	0.44	0.74	0.26	0.25	Trace	-
Arabinose	12.83	12.57	10.40	10.48	10.38	10.59	8.23	7.09
Xylose	37.38	80.60	78.98	78.44	77.62	77.38	71.22	86.56
Mannose	8.07	0.40	0.32	0.22	0.18	0.18	Trace	-
Galactose	11.61	1.83	1.95	2.35	2.49	3.02	2.45	0.44
Glucose	28.86	4.18	7.93	7.77	9.07	8.57	18.12	6.22
Uronic Acids	5.38	3.50	4.00	4.65	5.04	4.65	4.34	1.87

a, b, c, d – see the explanations in Table 12-3

Table 12-5: The contents of sugars and uronic acids in the isolated hemicellulose fractions on a whole mass basis (% of original bagasse dry matter). These data have been constructed by the Author based on the data in Table 12-3 and Table 12-4

Component	WS ₁ ^a	AS ₁ ^b	H ₂ O ₂ Concentration (%) ^c					AS ₂ ^d	Total
			0.5	1.0	1.5	2.0	3.0		
Rhamnose	0.056	0.054	0.018	0.011	0.008	0.002			0.149
Arabinose	0.577	1.609	0.416	0.157	0.311	0.085	0.033	0.269	3.458
Xylose	1.682	10.317	3.159	1.177	2.329	0.619	0.285	3.289	22.857
Mannose	0.363	0.051	0.013	0.003	0.005	0.001			0.437
Galactose	0.522	0.234	0.078	0.035	0.075	0.024	0.010	0.017	0.995
Glucose	1.299	0.535	0.317	0.117	0.272	0.069	0.072	0.236	2.917
Uronic Acids	0.242	0.448	0.160	0.070	0.151	0.037	0.017	0.071	1.197

a, b, c, d – see the explanations in Table 12-3

Other authors have noted that there can be a wide variation, for bagasse samples hydrolysed without polysaccharide isolation, in the ratio of arabinose to xylose (0.019 to 0.247) and xylose to glucuronic acid

(7.4-100) (Reicher et al., 1994, du Toit et al., 1984, Saska and Ozer, 1995, Lavarack et al., 2002). In the study by Lavarack *et al.* (2002) on the composition of SB from Queensland, an arabinose to xylose ratio of 0.11 (± 0.01) was determined.

The cellulose in SB has also been isolated and characterised. A study by Sun *et al.* (2004b) involved testing three different isolation techniques (which focus on the removal of all other structural components leaving the cellulose as a solid residue) on extractives-free samples of Chinese SB. It was found that the degree of polymerisation of the “cellulose” fraction varied between 1185 and 1406, according to the method of isolation. Sindhu *et al.* (2010) determined a degree of crystallinity of 63.6% for Indian SB cellulose.

12.1.4 NIRS and Sugarcane Bagasse

There have been a large number of publications relating to the use of NIRS for the laboratory analysis of sugarcane and sugarcane products. Typically these are for components relevant to the production of sugar from the original cane feedstock. For example Sverzut *et al.* (1987) developed calibrations for the pol, fibre, and moisture content of whole shredded sugar cane, and Tewari *et al.* (2003) developed a calibration for the pol content of sugarcane juice.

Regarding sugarcane bagasse, most of the publications that were found in a literature review by the Author also focussed on sugar-related properties, such as pol content. For example, Schaffler *et al.* (1993) obtained calibrations for pol, brix and moisture in bagasse. There appears to have been much less research into the development of quantitative calibrations for the lignocellulosic components (e.g. glucan, xylan, KL etc.) relevant to this study. The research that has involved SB in such calibrations tend to have only a few samples of SB in much larger and more diverse global datasets from which the calibrations are developed (e.g. Sanderson *et al.* (1996) and Kelley *et al.* (2004b)).

BSES is the standout research centre in the field of NIRS analysis of sugarcane and sugarcane products. This centre has been active in the development of lab-based NIR calibrations for many years (O'Shea et al., 2011). For example, a calibration developed for the sucrose content of 119 samples of molasses obtained from the Mulgrave Central Mill (North Queensland, Australia) provided an R_{calib}^2 of 0.995 and

a SECV of 1.14% using a FOSS InfraXact spectrophotometer scanning over the wavelength region 800-2778 nm (O'Shea et al., 2011).

BSES has also been a pioneer in research regarding the use of NIRS for the online analysis, at sugar mills, of sugarcane and the outputs of the milling process (Simpson et al., 2011, O'Shea et al., 2011). This research led to the joint development, with FOSS analytical, of three online NIR devices targeted for sugar factories. The first of these, the cane analysis system (CAS), is based on the NIR Systems 5000 device and features a scanning head that is pressed against the mill chute. It has been used for the quantitative analysis of a range of components such as pol, dry matter, ash, potassium, calcium, and fibre, and these have led to the integration of the system for cane payment systems (Pollock et al., 2007), cane quality schemes (Westmoreland et al., 2005, Pope et al., 2004), and for the control of the rate of fibre feed to the process (Jones et al., 2002).

The second device is the sugar analysis system (SAS) (Bevin et al., 2002). It is positioned above the conveyor belt that transports the raw sugar from the dryer. It is also used in process control schemes and to monitor the production of a low glycemic index sugar product known as LoGiCane™ (O'Shea et al., 2010).

The final sugar factory online NIR device developed by BSES and FOSS Analytical, and the one most relevant to this study, is the bagasse analysis system (BAS). It is configured in the same way as the CAS except it is used to analyse the bagasse. Its primary use is in the determination of the pol content of this bagasse. In addition to the pol content, calibrations have been developed for moisture content, ash content, and fibre content, among other components. These range of calibrations, coupled with those developed for sugarcane using the CAS, have enabled a calibration, involving both systems, to be developed for the pol extraction efficiency statistic described in Equation (12.1) (O'Shea et al., 2010). The BAS has also been used to develop calibrations for the calorific value of the bagasse (Staunton and Wardrop, 2006). This calorific value was calculated based on the relative proportions of fibre, ash, and pol in the bagasse and the heating values of these components. It was assumed that all of the fibre was cellulose and that no other components contributed to the heating value. The following statistics, (R_{val}^2) [SEP (MJ/kg)], were obtained: (0.83) [0.41 MJ/kg] and (0.83) [0.47 MJ/kg] for gross calorific value and net calorific value, respectively (Staunton and Wardrop, 2006).

Most relevant for this study, and particularly to the DIBANET project described in Section 18, the BAS has been used to develop quantitative calibrations for the lignin content of the bagasse (O'Shea et al.,

2010). This process involved the sampling of eleven bagasse samples from Mulgrave Central Mill. The lignin analyses of these were then matched with their BAS spectra. Subsequently, 21 more samples were collected and their lignin contents determined via a laboratory NIR device which had already been calibrated for lignin content. The SEC was only 0.09% and the R_{calib}^2 was very high at 0.99. This was based on a set of 33 samples that had a lignin content standard deviation of 2.56%. However, this calibration was not validated, either with an independent set or through cross validation, meaning that the possibility for overfitting exists. Also, the number of PLS factors used in the calibration were not reported.

The CAS and SAS are commercially available, with 23 CAS systems installed worldwide (O'Shea et al., 2010), while the BAS is undergoing development prior to its release. Two prototype BAS systems have been operational, one at Mulgrave Central Mill (Cairns, Northern Queensland, Australia), the site from where the samples analysed in this study were obtained, and the other at Costa Pinto Mill (Sao Paulo State, Brazil). The BAS system in Australia has been operation since 2005 (O'Shea et al., 2010).

12.2 Methodology Employed

In November 2006 the Author visited the research laboratories of BSES Limited in Brisbane, Queensland, Australia. At these laboratories there were 47 samples, stored in a freezer, of sugarcane bagasse collected, in 2006, from Mulgrave mill. This mill is situated in the town of Gordonvale, south of the city of Cairns in North Queensland.

The methods employed in sample preparation and analysis differed, in some instances, from the standard methods that were employed, in the following years, in the University of Limerick laboratories. The general sequence involved at BSES for a batch of samples is listed below.

1. A number of samples are removed from the freezer in the morning in order to defrost. The container is not opened, preventing moisture loss in the sample.
2. In the afternoon when the samples have equilibrated to room temperature, three NIR scans of each sample are taken. The NIR device used was the same model as described in Section 5.3.2.3 and the solid transport device and coarse sample cell were also used. These scans are therefore

equivalent to the “WU” scans described in Section 11.1. An average of the three scans for each sample was also recorded (“WU-A”).

3. Duplicate subsamples (each approximately 3 g) of each sample are placed in crucibles of known weight and then put in an oven at 105°C to dry overnight in order to determine the moisture content. Following this the ash content is also determined for these subsamples.
4. The remainder of each sample is placed on a tray of known weight, the tray is weighed again and then placed in an oven at a temperature of 45°. These samples are dried until the moisture content is consistent.
5. The dry and unground (“DU”) sample is scanned three times as described in Section 11.1.
6. An additional step, not outlined in Section 11, was employed at this point. It became quite apparent that there was a bias with these DU scans towards the smaller particles accumulating at the window of the cell, as described in Section 11.1. The DV method described in Section 11.1 was not employed here, however. Instead the whole DU sample was (hand) sieved using three sieves (2 mm, 850 µm, 150 µm) resulting in four particle-size fractions: $X > 2 \text{ mm}$; $2 \text{ mm} > X > 850 \text{ µm}$; $850 \text{ µm} > X > 150 \text{ µm}$; and $X < 150 \text{ µm}$. Each of these fractions was weighed. For each of the three largest fractions 2 NIR scans were collected and the average taken. The smallest particle-size fraction contributed only a minor component of the total mass (typically < 1%) and was not analysed with the NIR device. Using these scans and the known weights of each fraction a weighted average spectrum was calculated for each sample; this scan was labelled DW.
7. All fractions that had a particle size greater than 850 µm were processed in a rock mill and the comminuted sample was then hand sieved and the 850 µm > X > 150 µm fraction added to the proportion of the DU sample that was already that particle size. In this chapter this fraction is labelled DB. Two DB scans were collected for each sample and the average spectrum calculated.

The original plan for the work to be conducted at the Brisbane laboratories was to undertake a complete analytical method for the lignocellulosic components of the prepared DB samples. Unfortunately, problems with the HPAEC-PAD device, which was in the process of being installed during the Author’s time at BSES, prevented the hydrolysis stage from being carried out. That meant that analysis was only possible for ethanol-soluble extractives and ash.

The ethanol-extractives method was different from that employed in the Carbolea laboratories (Section 11.4). It involved the removal of 80% ethanol extractives in four repeat extractions as discussed in Section 3.4. The steps involved are summarised below; samples were analysed in duplicate.

1. The moisture content of the DB sample is determined, in duplicate, prior to extraction. The ash contents of these samples are also determined (see Section 11.3).
2. Approximately 3 g of the DB sample added to a 150 ml centrifuge tube, exact weight noted.
3. Then 75 ml of 80% aqueous ethanol is added to the centrifuge tube.
4. The tubes are then placed in a ultrasonic water bath for 15 minutes at room temperature.
5. The tubes are then centrifuged at 700 x *g* for 10 minutes and the supernatant is decanted into a separate flask for each sample/duplicate.
6. Steps 2-4 are repeated a further 3 times.
7. A container, including a lid and filter paper, is removed from a 40°C oven and weighed.
8. The sample and the total extract collected over all the extractions are then filtered through this filter paper and the tins are then placed in the 40°C oven where these are left to dry over a period of a few days.
9. The centrifuge tubes are placed in a 105°C oven and weighed the next day.
10. The containers are then weighed and the moisture content of the extracted sample determined from a single subsample of it. The rest of the extracted sample is retained for the subsequent hydrolysis step.
11. The extractives content is then determined as the moisture-corrected mass loss of the sample plus any difference in the weight of the centrifuge tube that occurs (due to residual particles).

In December 2006 the Author returned to Limerick. At this point all 47 bagasse samples had been processed to a DB state and had their spectra collected at the various stages of sample preparation. Duplicate data for the moisture contents and ash contents of all 47 samples, in the WU state, existed as did the ash contents for the DB fractions of all 47 samples. A total of 40 of these samples had 80% ethanol-soluble extractives data.

The DB samples were kept in sealed containers in cold storage and the intention was that these would be delivered to Limerick once the analytical equipment (HPAEC-PAD, NIR, accelerated solvent extractor (ASE) etc.) were installed and operational there. Unfortunately, in this interim period some of the

samples were thrown out of the cold storage unit at BSES meaning that only a total of 30 samples were delivered to UL.

At the point at which these samples were received a Retsch AS200 sieve shaker was installed in the Carbolea laboratories and the Author found that, when the DB samples were sorted in this device there was a proportion that was of a particle size less than 180 microns (the lower limit established for the DS samples in Limerick). The Author wanted to avoid any analytical inaccuracies that may result from variations in the liberation and degradation of sugars in the acid hydrolysis step due to differing particle size distributions between samples (see Section 3.1.2.3). Hence, all the samples received were sieved in the AS200 unit, the true DS fraction ($180 \mu\text{m} < X < 850 \mu\text{m}$) of each was collected and scanned in the Carbolea-laboratory FOSS XDS NIR unit using the coarse sample cell and solids transport device. For two of the 30 samples there was insufficient DS sample to fill the coarse sample cell and no spectra were collected for these. However, the analytical data for these two samples could be used in calibrations involving the spectral data sets collected at BSES.

The reference analytical methods at UL were carried out according to the standard methods outlined in Section 11. However, the small amounts of each sample that were available were a hindrance in this study. In some cases it was necessary to forego a direct determination of moisture in order to preserve the maximum amount of sample for the hydrolysis stage, or to enable the analysis of duplicates. As described in Sections 11.4 and 11.5, an NIR spectrum of each sample was collected prior to its analysis in the ethanol extraction or acid hydrolysis methods. For spectra that had corresponding reference analytical data for the moisture content, a quantitative NIR calibration was developed and used to predict the moisture contents of the samples for which only spectra were collected. A total of three different moisture content calibrations were developed (in addition to the calibration for the moisture content of the WU sample). The first was for the DS sample prior to ethanol extraction, the second was for the moisture content of the extracted sample and the third was for the moisture content at the time of the acid hydrolysis procedure. These calibrations are discussed in Section 12.3.3.2.

As described in Section 11.7 it was general practice to repeat samples for which the data for the duplicates differed over a given threshold. However, the lack of sufficient material for re-analysis of many of the sugarcane bagasse samples prevented this. This is reflected in some relatively high values for the standard error of the laboratory statistic in Table C-3 and Table C-4.

In addition to the data acquired at the UL laboratories, the Author was also supplied with data from BSES concerning the carbon and nitrogen contents (determined via elemental analysis) of the 47 samples. Summary statistics for the PLSR of these are provided in Section 12.3.3.4.

12.3 Results and Discussion

12.3.1 Reference Analytical Data

Table C-1 provides the results (on an extractives-free basis) for each of the 30 samples analysed at the University of Limerick. Table C-2 provides these, and further data, on a whole-mass basis. In the instance of duplicates being analysed the standard deviation of duplicates (SDD) statistic is also included. Reliable extractives content data (based on the change in weight of the sample, i.e. the EXTR_PD value) could not be obtained for sample 50017. That means that only extractives-free data are available for analytes determined after the ethanol-extraction step for this sample. Also, reliable extractives content data (based on the weight of the extract in the collection vial, i.e. EXTR_CV) were not obtained for sample 50047. It was also not possible to determine the ethanol insoluble ash content (and hence associated statistics such as the ethanol soluble ash and acid soluble ash) of sample 50046.

Table C-3 and Table C-4 present some summary statistics and histograms for the most important components in Table C-1 and Table C-2. These statistics are based on a data-set excluding sample 50019 (see Sections 12.3.2 and 12.3.3.3). The histograms represent the distribution of component values, on a whole mass basis, across this data set. A model normal distribution is included, for comparison, as a curve on these histograms. For all components the distribution statistics are provided in whole mass (WM) basis, and for relevant components these are presented on an extractives-free (EF) and/or ash free (AF) basis. The SD rows in Table C-3 and Table C-4 describe the distribution of the data-set of SDDs, (i.e. the SD columns in Table C-1 and Table C-2) for each component. This SDD is calculated on an extractives-free basis for analyses that took place after the ethanol extraction step (SD_{EF} in Table C-3 and Table C-4) and on a whole mass basis for other components (SD_{WM} in Table C-3 and Table C-4). The SEL statistic in this "SD" row is calculated as described in Section 6.7. The SEL statistics in other rows represent the division of the SEL by the average (AV.) value in each row. Therefore, for example, the SEL

of the glucose-EF dataset is 0.20%, and this is equivalent to 0.48% of the mean glucose-EF value of 42.19%.

It can be seen from the data in Table C-3 and Table C-4 that the precision of the data, measured according to the variation of the SDD over the data set, and by the SEL statistic, is generally good. This is despite the presence of a few samples where the SDD for selected components were greater than the ideal limits outlined in Section 11. For example, the SEL for total sugars (extractives free) was only 0.30%, equivalent to 0.43% of the average total-sugars (extractives-free) value. This relative error is much greater, 9.52%, for the EXTR_PD content, but the SEL is still reasonable at 0.40%. It should be noted, however, that there was only sufficient material for duplicate EXTR_PD analysis of 6 samples.

A total of 9 hydrolysis batches were undertaken. Table C-6 presents the data regarding the sugar recovery values for each of these batches and Figure C-1 plots these data. It can be seen that the SRS data are generally very consistent, demonstrating the reliability of the autoclave and hydrolysis procedure over this period. The mannose and rhamnose recoveries varied the most. However, this was a result of their low concentrations in the SRS samples. Since these are minor constituents in bagasse this variation is not that important. It was therefore suitable to use the individual batch SRS-corrected data for the component sugars in all calibrations (i.e. GLU_SRS, XYL_SRS etc.).

As expected, the major components of the bagasse samples analysed are glucose, xylose and Klason lignin. The concentrations of these components, provided in Table C-3 and Table C-4 are comparable to those in Table 12-1. The glucose sugars liberated in the hydrolysis step would primarily have come from cellulose whereas xylose would have come from hemicellulosic polysaccharides (xylans). The other hemicellulosic sugars that were analysed for in this experiment were arabinose, which had an average value of 2.50% on a whole dry mass basis, and galactose, which had an average value of 0.99%. The ratio of arabinose to xylose contents ranged from 0.11 to 0.09, within the ranges discussed in Section 12.1.3. The ratio of galactose to xylose contents ranged from 0.0434 to 0.0297. There were no correlations between this ratio and xylose or between the arabinose:xylose ratio and xylose. The other hemicellulosic carbohydrates analysed for, mannose and rhamnose, are only present in minor quantities, as expected (see Section 12.1.3).

It should be noted that, throughout this Thesis, the constituent sugars in the polysaccharides are expressed using the monosaccharide names (e.g. glucose, xylose, etc) rather than the polymer versions of these names (e.g. glucan, xylan etc.). This is to avoid confusion that a homopolysaccharide (e.g. xylan)

may be being referred to when in fact it is the component sugar of a polysaccharide that is being referenced. However, while the monosaccharide terms are used, the concentration values provided for these constituents reflect the mass of the sugar in the polysaccharide (i.e. with one less hydroxide molecule than the monosaccharide), as described in Section 4.6.2.4.

Table C-5 provides the correlations between many of the chemical components (the WM data were used in all instances), absolute r values greater than 0.5 are highlighted in bold. Some of the correlations are to be expected; for example the positive correlations between many of the ash components. However, other relationships are more interesting. For example, there is a negative correlation between the galactose content and the extractives content. There is also a negative relationship between most non-ash components and total ash, and between these non-ash components and other ash-containing components (e.g. EIA, AIR, etc.). This relationship is expected given the relatively wide variation in the ash contents and the fact that a large increase in one component will reduce the total amounts of all other components.

However, the ash relationship is by far the strongest ($r = -0.795$) for the glucose content. As a reflection of this there are also negative relationships between glucose and: AIR ($r = -0.791$), AIA ($r = -0.685$), and EIA ($r = -0.797$). Regarding the correlations of glucose with other components, there is a positive correlation ($r = 0.785$) with the ASL content. This could be a reflection of the influence of glucose acid degradation products (e.g. hydroxymethyl furfural) in UV-absorption spectra (see Section 3.2). However, there are no strong correlations between ASL and the other polysaccharide sugars. There are also no strong correlations between xylose and arabinose or xylose and galactose. However, there are reasonable correlations between galactose and arabinose ($r = 0.675$), mannose and arabinose ($r = 0.719$) and mannose and galactose ($r = 0.648$).

12.3.1.1 *Extractives Content*

Table C-5 also provides the value for the correlation ($r = 0.904$) between EXTR_PD and EXTR_CV. Figure C-2 (a) plots this relationship. It can be seen that there is a bias towards a higher EXTR_CV content, compared with the EXTR_PD content, and that this principally occurs in samples towards the lower end of the extractives content range. This is reflected in the equation of the regression line in Figure C-2 (a); the slope is less than 1 and the intercept is over 1%. This relationship is also demonstrated in Figure C-2

(b) which plots the extractives content residual, determined as EXTR_CV minus EXTR_PD, against EXTR_PD. The extracts in the collection vials often took many days in the 40°C oven to dry to a steady weight. It is possible that, even at this stage, all of the solvent was not removed or that moisture from the air was retained on the samples. These are possible explanations for the higher extractives contents for EXTR_CV compared with EXTR_PD.

As discussed in Section 12.2, the extractives content was also determined, for a total of 40 samples, in the experiments at the BSES laboratories using the 80% ethanol extraction method. The summary statistics of this analysis are provided in Table C-7, where these are compared against those for the EXTR_PD and EXTR_CV data determined at UL. Figure C-3 provides a histogram of the BSES extractives-content data and Figure C-4 (a) and (b) are scatter plots showing the relationship between the BSES data and the EXTR_PD and EXTR_CV data, respectively. It can be seen from the statistics in Table C-7 that the SEL was much greater for the BSES data. Furthermore, the average, max, and min values are all larger than those for the EXTR_PD or EXTR_CV data. Also, the plots in Figure C-4 show that the BSES data are not following the trends of the EXTR_PD/EXTR_CV data sets. It appears that the extractives data obtained at BSES are poor and of no use for NIRS calibration. The probable reason for the failure of the BSES method relates to the poor accuracies that resulted from the moisture content determination of the extracted samples.

12.3.1.2 *Ash Content and Moisture Content*

As mentioned in Section 12.2, for many samples there was an insufficient amount of material to allow for accurate ash analysis of the DS fraction at the Carbolea laboratories. In total only 20 of the samples in Table C-2 had their ash content analysed at UL. Ash data were obtained at BSES for all 47 bagasse samples and these data could potentially be used to represent those samples missing UL ash data. However, as also described in Section 12.2, these DB samples were of a different particle size distribution compared with the DS samples analysed at UL.

Before including the BSES ash data in Table C-2, and using these in the subsequent development of quantitative NIR calibrations, the Author first undertook an investigation to determine the correlation between the two sets of results. Figure C-5 (a) shows a scatter plot involving those samples that had ash data for both analyses (i.e. BSES and UL). It can be seen that the y:x relationship is not close to one;

however, this is primarily a result of the “outlying” high ash value sample that had a residual (UL ash value minus BSES ash value) of -2.72%. If this point is excluded, then the relationship is more equal, as shown in the scatter plot in Figure C-5 (b). It was therefore considered reasonable to use the ash values obtained at BSES in instances where, due to limited sample amounts, selected samples were not ashed at UL.

The BSES ash data will be used in the development of ash calibrations for the 47-sample NIR data sets that include all the spectra collected at BSES, as outlined in Section 12.2. Table C-8 provides, under the “DB Fraction Ash Content” heading, a summary of these ash data, and Figure C-6 (a) provides a histogram representing these. It can be seen that the histogram and the SEL are similar to those for the UL data-set, Table C-3.

As described in Section 12.2, the subsamples of the WU fraction that were used to determine its moisture content were subsequently ashed. Due to laboratory error the ash fractions of one of the WU analytical batches were lost, meaning that data exist for only 42 of the 47 samples. Table C-8 includes the summary statistics for these analyses under the heading “WU Fraction Ash Content”. It also includes summary statistics for the moisture contents determined for these WU subsamples and Figure C-6 (b) presents a histogram of these data. It can be seen that the SEL values for analyses involving the WU fraction are quite high, 1.81% for the moisture content and 1.10% for the ash content, with the ash content SEL being more than double the value of the same statistic for the DB analysis. This is an expected result given the relatively high degree of heterogeneity with regard to the particle size distribution of the WU fraction compared with that of the DB fraction.

Figure C-7 provides a scatter plot showing the relationship between the ash content determined from the WU fraction compared with the ash content determined from the DB fraction. The relationship is reasonable, although the regression equation indicates that the WU ash values tend to be larger than those of the DB fraction. A possible explanation for this is that there may have been a bias towards the sampling of smaller particle sizes when obtaining WU subsamples for moisture and ash analysis. This bias occurred because it was difficult to fit the larger fractions of the bagasse into the crucibles used for analysis. As outlined in Section 16.3, higher ash contents can be associated with the smaller particle size fractions of biomass samples.

12.3.1.3 *PCA of Chemical Data*

A principal component analysis was conducted for all samples using the following 10 variables: EXTR_PD, ASH, KL_EF, ASL_EF, ARA_EF_SRS, GAL_EF_SRS, RHA_EF_SRS, GLU_EF_SRS, XYL_EF_SRS, MAN_EF_SRS. The extractives-free data were chosen for the lignocellulosic components since there was no reliable value for the EXTR_PD content for sample 50017.

Figure C-8 (a) shows the explained variance plot for up to 10 PCA factors with the blue line representing the explained variance in the calibration set and the red line the explained variance estimated via full cross validation. Five PCs explained a total of 99.50% of the variance in the data set with cross-validation with the first PC explaining 85.5%. Figure C-8 (c) shows the loadings plot for PC1 and PC2. It can quite clearly be seen that the ash and glucan contents are primarily responsible for these first two PCs, with the KL content also making a contribution to PC2. The PC3 vs. PC4 loadings plot shown in Figure C-8 (e) illustrates that the extractives content is the main contributor to PC3 and the KL content (and to a lesser extent the xylose content) to PC4. The PC5 vs. PC6 loading plot (figure not included) showed the xylan content as the major contributor to PC5 and the arabinan content to PC6, although both of these components only explained an additional 0.16% of the X-variance compared against a 4 PC model.

All of these loadings make sense when looking at the data in Table C-3 and Table C-4 – ash, glucose, extractives, and KL contents demonstrate the greatest absolute variation within the data-set. Figure C-8 (d) shows a scores plot for PC1 versus PC2. Here the high positive score along PC1 for sample 50017, coupled with the high positive loading for ash content regarding this PC, reflect the very high ash content observed for this sample (Table C-2).

Figure C-8 (b) shows an influence plot for a 5 PC model. It can be seen that samples 50017 and 50019 have a very high leverage; indeed their calculated Hotelling T^2 scores were above the critical limit. Sample 50038 has a high residual X-variance in this influence plot. This outlying residual variance is removed after PC6 (the arabinan-focussed PC) is considered. As can be seen in Table C-1, this sample has the highest arabinan content of the dataset.

The influence of sample 50017 is clearly due to its ash content which is over 7% greater than any other sample. This high ash content will reduce the whole mass contributions that the other components can make. The extractives-free data were corrected to an extractives-free and ash-free basis, and in the resulting scores plot it was observed that, while sample 50017 still displayed some residual X variance,

the only sample now outside the Hotelling T^2 ellipse in the PC scores plots was sample 50019, and sample 50017 was much less influential to the model until PC6 (which had significant loadings for galactose and arabinose) was considered.

The issue regarding sample 50019 is unclear. Observing its compositional values in Table C-1 and Table C-2 it does not have outlying values for the most important components. However, when looking at the different ash compositions discrepancies appear. For instance, the EIA value is higher than the total-ash value, and this results in a negative ESA content, which is clearly impossible. As shall be seen in Section 12.3.3, this sample was a significant outlier in all of the calibrations developed for lignocellulosic components, although it was not an outlier for total ash, EXTR_PD, or EXTR_CV contents. It appears that this sample may have been mislabelled after the ethanol extraction step, and for this reason it has been excluded from the calculations for the statistics and histograms in Table C-3 and Table C-4, and it is also excluded from the calibrations for lignocellulosic components.

Principal Components Regression (PCR) Using Chemical Data

Using the original variables outlined at the start of this Section, and excluding sample 50019, PCR was used to predict each variable in regressions involving all of the other variables. Summary statistics of this are provided in Table C-9. It can be seen that the predictive ability was reasonably good for glucan. Figure C-9 (a) includes an explained variance plot for this regression. PCs 1, 3, and 5 are the major contributors to the explained variance of this component. Figure C-9 (b) includes a loading plot of PC1 vs. PC3 and Figure C-9 (c) a loading plot of PC3 vs. PC5. It can be seen that, as expected, ash is a major contributor in explaining much of the variance in the glucan content. KL is the most important contributor to PC3 and arabinan, galactan, and ASL are all relevant to PC5. Figure C-9 (d) features a plot of the regression coefficients for each variable in a 5 PC model.

The calibration for ash was also reasonable, and only required 3 PCs as shown by the Explained Variance plot in Figure C-10 (a). A loadings plot of PC1 vs. PC2 is included in Figure C-10 (b) where it can be seen that glucan is the major contributor to PC1 and EXTR_PD to PC2. The PC2 vs. PC3 loadings plot, Figure C-10 (c), shows that KL and xylan are important here. Figure C-10 (d) shows the important contributing variables to the regression coefficients of a 3 PC model. For the ASL regression, sample 50042 was a clear outlier in the predicted vs. reference plot, and when it was excluded from the regression the R_{CV}^2

for the 2 PC model rose to 0.715 and the RMSECV fell to 0.081%. As shown in Figure C-10 (e), in this regression, for PC1, ash had a large positive loading and glucan a large negative loading and, for PC2, glucose had a large positive loading and KL a large negative loading. In this regression glucan had the greatest regression coefficient, followed by KL and then EXTR_PD.

12.3.2 NIR Spectra and PCA of Spectral Data

This section will discuss the NIR spectra of the various data sets analysed at BSES (DB, DU, DW, and WU) and the PCA of these data sets. The discussion and PCA will focus on the 30 samples that were analysed via wet chemical methods at UL, but comments and comparisons will be made concerning the larger sample set of 47 samples. Figure C-11 and Figure C-12 illustrate the explained total X-variance (via full cross validation) with an increasing number of PCs, for the full-spectral-region (400-2500 nm) PCAs (Figure C-11), and for the 1100-2500 nm region PCAs (Figure C-12, referred to as “NIR”) for each of the spectral data sets (DB, DU, DW, WU).

12.3.2.1 DB Spectral Data-Set

Figure 12-2 (a) displays the raw spectra of the 30 samples analysed at UL, and Figure 12-2 (b) displays the raw spectra of all 47 samples.

The second derivative (SG2,2,10,10) was used as a pretreatment prior to PCA. Separate PCAs were carried out for the spectral regions 400-2500 nm (FULL) and 1100-2500 nm (NIR), and on the UL data set (UL) and the total data-set (ALL). The explained variance of the (full) cross validation with increasing PC numbers for these are included in Table C-10. It can be seen that the full spectral region requires more PCs to attain the same level of explained variance as the 1100-2500 nm spectral region.

Figure C-13 (a) to (d) shows the X loadings plots for a PCA carried out over the full spectral region for the UL data set, and Figure C-13 (e) to (h) shows the loadings plots for the PCA carried out over 1100-2500 nm spectral region for this same data set. The loadings plots for PCAs carried out on the 47 sample dataset were very similar and are not included. It can be seen from these plots that, when the visible region is included in the PCA, this region contributes greatly to the PCs that are generated. This

phenomenon continued through the higher PCs. The significant variance within the raw spectra over the 400-1100 nm region can clearly be seen in Figure 12-2, and this variation is not purely a baseline shift effect since it is evident in derivative and scatter corrected (e.g. MSC, SNV etc.) plots and PCAs.

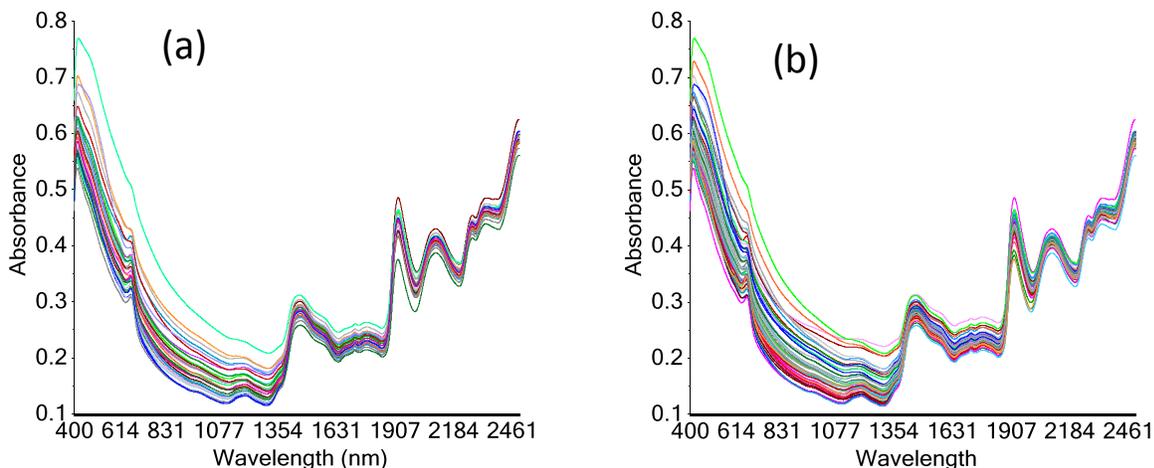


Figure 12-2: The visible-NIR spectra for the bagasse DB scans of: (a) the 30 samples that were analysed at UL; (b) all 47 samples processed.

Given the spectral complexity of this wavelength region and the number of PCs required to model it, it should only be included in regression models if it is relevant to the analyte in question. When only the 1100-2500 nm region is used for the PCA it can be seen that the important wavelengths noticed for this region in the full-spectrum PCA are emphasised with greater relative loadings.

The Unscambler X recommended 11 PCs to model the full spectral region PCA. In a Hotelling T^2/Q -Residual influence plot of this model sample 50034 breached the Q-residual limit ($\alpha = 0.05$) and no sample breached the Hotelling T^2 limit. Regarding leverage, sample 50047 had the greatest, followed closely by sample 50017. Both of these samples also had by far the greatest residual sample variance, under cross validation, in this 11 PC model. Furthermore, in individual sample plots of the sample residuals over the 400-2500 region, the greatest X-residuals were present in the visible region for both of these samples. Indeed, this was the case for sample residuals for all samples, including outlying sample 50034.

Looking at the chemical data of the samples, Table C-1 and Table C-2, and the PCA score plots of these, particularly Figure C-8 (b) and (d), it is clear that sample 50017 is most likely a spectral outlier due to the high ash concentration of the sample and the effect this has on other components of that sample. Ash is

not detected directly in the NIR but it can have indirect effects, particularly in the shorter wavelengths as discussed by (Sanderson et al., 1996).

The raw NIR spectra of samples 50017, 50034, and 50047, along with the average spectrum of the 27 other UL samples, are plotted in Figure 12-3 (b). Explaining the high leverage of sample 50047 and the high residual of sample 50034 purely based on the chemical data is difficult since these are not outliers in the chemical-PCA plots, Figure C-8, nor do they have abnormal values for any particular chemical component, Table C-1 and Table C-2. However, Figure 12-3 (b) shows a clearly different profile to the spectrum for sample 50047 over the visible region. The absorbance starts out at a much higher value than for most other samples (with the exception of 50017) but then drops much more rapidly; hence the effect is much more than just a baseline variation that may exist due to scattering effects, and it will carry through to plots involving derivative spectra.

Examining the PCA involving the 1100-2500 nm spectral region, the Unscrambler software recommended a 6 PC model to explain the UL dataset of 30 samples, with this model explaining 92.69% of the X-variance. In a Hotelling T^2/Q -residual plot of this model, sample 50034 now has average leverage/residuals, while sample 50047 is no longer such a highly influential sample but is close to the Q-residual limit (as is 50007). Sample 50017 is still highly influential.

12.3.2.2 DU and DW Spectral Data-Sets

Figure C-14 (a) shows the DU spectra for the 30 samples analysed at UL, and Figure C-14 (b) shows the DW spectra for these samples. Figure C-14 (c) shows the DU spectra for all 47 samples, and Figure C-14 (d) shows the DW spectra for these samples. There does not appear to be a great deal of difference, visually, between these Figures. Figure 12-3 (a), however, plots the spectra for the different datasets (WU, DU, DW, DB) for just one sample (50041), and here it can be seen that the DU spectrum has, over all wavelengths, a lower absorbance value than the DW spectrum, and that this difference increases with the wavelength of the radiation. This is clearly a multiplicative scatter-related effect with the smaller average particle size of the DU fraction reflecting more radiation. Indeed, the absorbance values of the DU scan are quite close to those of the DB scan at longer wavelengths.

Other differences between Figure C-14 (a) and Figure C-14 (b) can be determined by plots of the standard deviation spectra (using all 47 samples) for DU and DW, Figure C-15. This shows that the variation is less for the DW set. This occurs both for the raw spectra and, to a lesser extent, for the second derivative (SG2,2,10,10) spectra.

It was expected that the DU set would demonstrate less variation as a reflection of the tendency for smaller particle sizes to be presented to the window of the NIR cell. However, the fact that the variance was greater in this set shows that this particle size bias may not have been as consistent as thought, and would, of course, be subject to the variation in the relative amount of small particles present in the sample in question. The DW set, in giving proportionate weight to each particle size fraction, therefore appears to give a better representation of the complete set of samples.

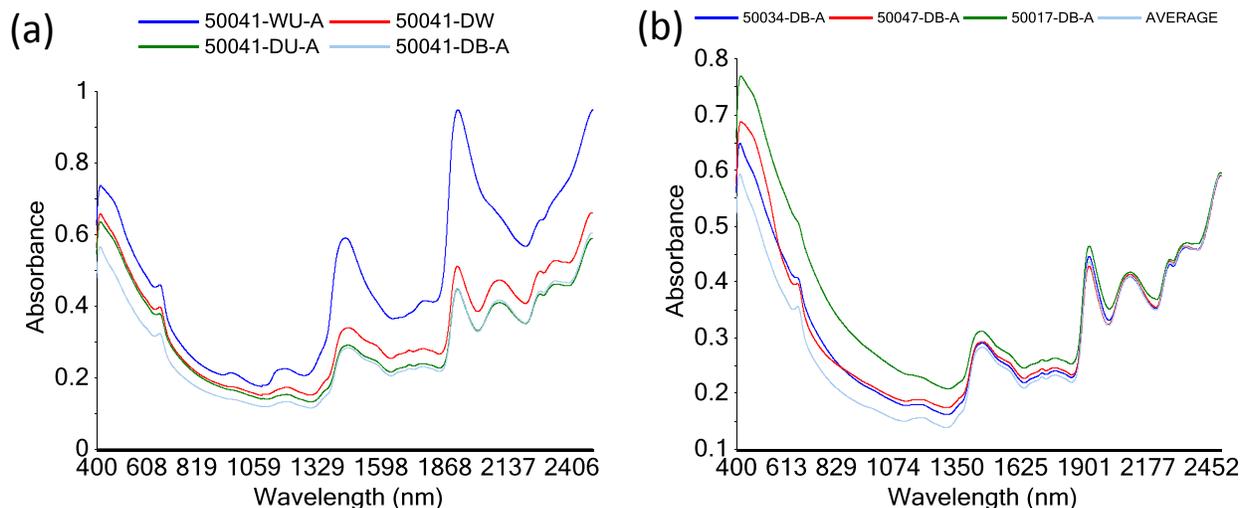


Figure 12-3: Selected bagasse spectra. (a) The spectra of sample 50041 at various stages of preparation (WU, DU, DW, DB) ; (b) A plot of the spectra for samples 50034 (blue), 50047 (red), 50017 (green) and the average of the 27 other samples in the UL data-set (light blue).

Table C-11 shows the total explained X-variance, in cross-validation, for the different sample sets and spectral regions. These are also illustrated in Figure C-11 and Figure C-12. It can be seen that, for both the DU and DW datasets, the early PCs explain less of the X-variance than the same PCs in models on the DB scans. There is less difference in total explained variance between the DU and DW scans. However, for most PCs and datasets, the figure is slightly higher for the DW scans. For both DU and DW scans, as with the DB scans, PCs built on the 1100-2500 nm spectral region tend to explain more of the X variance than the equivalent PC for a full-spectral region model.

Figure C-16 (a) to (d) present the loadings plots for PCs 1 to 4 for a PCA constructed over the 400-2500 nm region using the DU data set and Figure C-16 (e) to Figure C-16 (h) present the corresponding loading plots for the DW data set. Similarly, Figure C-17 (a) to (d) present the loadings plots for PCs 1 to 4 for a PCA constructed over the 1100-2500 nm region using the DU data set and Figure C-17 (e) to (h) present the corresponding loading plots for the DW data set.

The influence plot for the full spectral region model of the DU set was somewhat similar to the DB influence plot; in both cases the Unscrambler selected 11 PCs for the model and samples 50017 and 50047 had high leverage. Regarding the DW (400-2500 nm) 11-PC-model influence plot, sample 50047 is still highly influential but sample 50017 much less so.

For the 1100-2500 nm models, there were similar trends as with the DB model. For example, in both the DU and DW models sample 50047 was no longer such a highly influential sample but sample 50038 was.

12.3.2.3 WU Spectral Data-Set

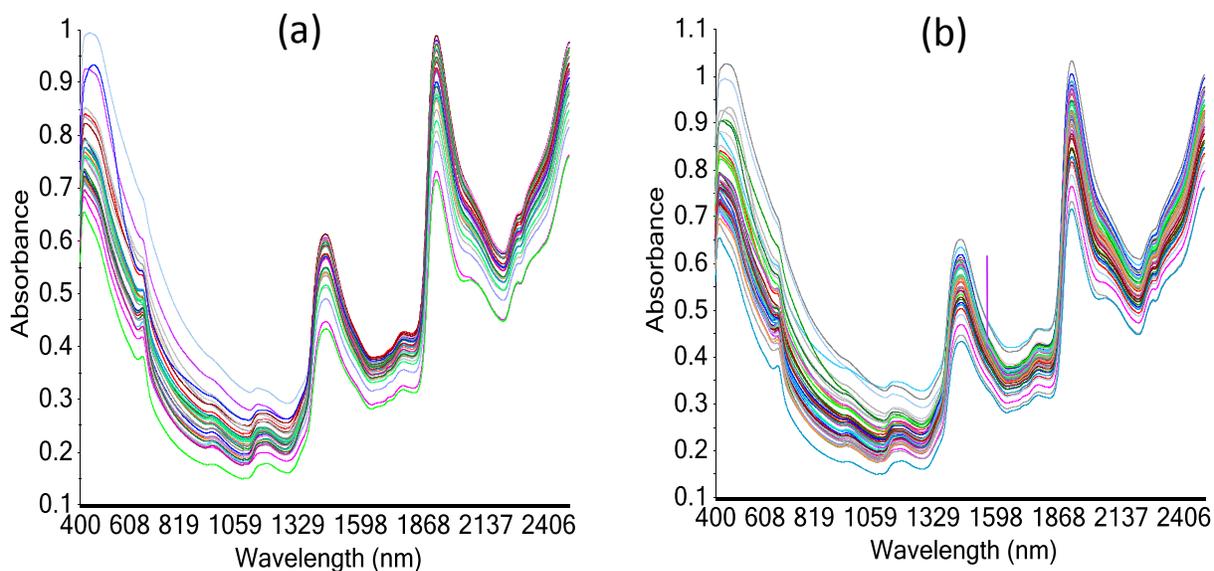


Figure 12-4: WU bagasse spectra. (a) WU spectra of the 30 bagasse samples analysed at UL; (b) WU spectra of all 47 samples.

Figure 12-4 (a) shows the WU spectra of the 30 samples analysed for their chemical constituents at UL and Figure 12-4 (b) shows the WU spectra of all 47 samples. The clear difference between these spectra and those of the DB and the DU and DW sets, Figure 12-3 (a), is apparent in the strong absorbances by water around 1390 nm (associated with the first overtone of an O-H symmetrical vibration) and 1895 nm (associated with the combination band involving an O-H bend and asymmetrical stretch).

There is also significant variation within the WU dataset, Figure 12-4, and this presumably is a reflection of the differing water contents of these samples. The importance of the contribution of water to the WU spectra is clearly demonstrated in the PC1 loading plots for the WU dataset, Figure C-18 (a) and (e). For the first time, the first PC in the 400-2500 nm PCA model has its largest loading values outside of the visible region, Figure C-18 (a). After this first PC, the visible region again dominates the loadings for the subsequent PCs in the full spectral region model. It can also be seen, Figure C-11, that the first PC explains significantly more of the total X-variance than any of the first principal components in other models. Table C-12 summarises the total explained X-variance, under cross validation, for the different WU PCA models. As can be seen in this Table, there is a significant difference, for PCs 1 and 2, between the UL data set and the full dataset values for total explained variance in the NIR region model. This was represented in the PCA loadings plots for the second principal component in these different data sets. In all other datasets the loadings plots for the 30-sample-model and the 47-sample-model were very similar; however, in this case there appeared to be an additional PC after PC1 for the 47-sample-model. This second PC is illustrated in Figure C-20. The loading plots for the previous and subsequent PCs in the 47-sample model [47] looked similar to the corresponding loadings plots in the 30-sample [30] data set (i.e. PC1[47] looked like PC1[30], PC3[47] looked like PC2[30], PC4[47] looked like PC3[30] etc.).

It can be seen in the loadings plot for this PC, Figure C-20, that the important variables again cover the region where water absorbs NIR radiation. The presence of this PC only for the 47-sample dataset might be explained by the fact that this set had a larger range in, and standard deviation of, moisture content values. That would mean that, even after the contribution of PC1, the most significant variation in the spectra, for the NIR region, was still a result of the moisture present in the samples, and this would need to be modelled by PC2.

12.3.3 NIR Calibrations

12.3.3.1 Suitability of Data for Development of Quantitative Calibrations

It is clear from Table C-3 and Table C-4 that, for most components, the ranges in concentrations over the 29 samples are small and the standard deviations of these concentrations are even smaller. For example, the range for KL values, on a WM basis, is only 2.71% and the standard deviation just 0.67%. That would mean that NIR calibrations would need to have extremely accurate predictions in order to describe most of the variance in the data set. Using the quality thresholds outlined in Section 6.11.2, 0.14% would be the upper limit for the RMSEP in order to achieve an RER of 15 for KL and a calibration that could be considered suitable for quantitative predictions.

As can be seen in Table C-3 this value is less than the SEL value of 0.32%. If the RPD statistic were to be used as a quality threshold then it would be even harder to find a suitable calibration given the small SD of 0.67% for KL values. Indeed, for some components the SD is so small that it is only fractionally greater than (ASL, RHA) or equal to (MAN) the SEL.

Table C-3 and Table C-4 show that the ranges and standard deviations of many constituents are significantly less when the data are examined on an ash-free basis. Ash content is the most variable constituent, and clearly this variation is responsible for much of the variation, on a whole mass basis, of other components.

12.3.3.2 Moisture Content for Extraction and Hydrolysis Experiments

12.3.3.2.1 Moisture Content at the Extraction Stage (DS-E Batch)

A total of 21 moisture content values were available for this calibration. Summary statistics regarding the moisture content of these samples are included in Table C-13. It can be seen that there was little variation in moisture content, with the standard deviation being particularly low.

PLS calibrations were undertaken in the FOSS Vision software since this allows moisture predictions to be made immediately upon scanning unknown samples. Vision uses Wold's criterion (i.e. the local minimum in PRESS, Section 6.8) in order to decide how many factors to use. This criterion was retained for these moisture content calibrations.

The calibrations were tested through full cross-validation. Table C-15 summarises the various pretreatments that were employed and the regression statistics that resulted. It can be seen that there was not a great variation in RMSECV values between the different calibrations. Ultimately, the MSC transform over the 1100-2500 nm spectral region was chosen since it demonstrated the lowest SECV with the lowest number of factors (2). Other models, mostly those involving derivatives, often provided lower SECs (but not SECVs), but required more factors. This suggests that these calibrations were being overfitted.

12.3.3.2.2 Moisture Content of the Extracted Samples (DS-E Dishes)

The indirect determination of extractives content (Section 11.4) requires determining the weight loss of the extracted sample; hence its moisture content needs to be known. During the period of time when these extractions were being carried out it was typical to perform duplicate, or triplicate, extractions of the same sample. After the extraction, the extracted material of each replicate was allowed to air dry on a separate petri dish prior to the collection of its spectrum and oven-determination of its moisture content (this oven-determination was only ever carried out for a maximum of 2 of the replicates).

Hence, there could be up to three NIR scans of the extracted residues of a sample from any particular analytical batch in the NIR dataset, representing the replicates that were analysed. Including all of these replicate scans in a calibration that uses cross validation could lead to potential overfitting since the sample excluded from the model in the cross validation stage could most likely be well predicted by its replicate in the calibration set. For that reason, only the scans and data for the first replicate of the sample were included in the calibration.

Table C-13 provides summary statistics for the moisture content at this stage (DS-E Dishes). It can be seen that the range and SD for this stage are much greater than for the DS-E samples. As with the DS-E calibration, a range of pretreatments were tested prior to PLS calibration. However, it was found that the same pretreatment and wavelength range for PLS calibration was ideal. Regression statistics for this

calibration are provided in Table C-14. It can be seen that, while the SECV is higher than for the DS-E calibration, the wider concentration range results in higher values for R_{calib}^2 , RPD, and RER.

12.3.3.2.3 Moisture Content at the Hydrolysis Stage (E-H Batch)

After the determination of the moisture content of the samples in the petri-dishes, the replicates were combined into a single test tube which was sealed and stored for future hydrolysis. The hydrolysis procedure (E-H) requires a determination of moisture content at the start.

Table C-13 includes summary statistics for the moisture contents of the samples used in the calibration set for the E-H regression. As with the other calibrations, a range of pretreatments were tested, but MSC offered the best combination of good predictive ability with few factors. Regression statistics for this calibration are included in Table C-14. The predictive ability of this equation is somewhat poorer than the other two, but still reasonable.

12.3.3.2.4 Cross-Predictions Using Moisture Calibrations

The predictive ability of each calibration was tested by using it to predict the moisture contents of the samples in the other sets (e.g. the DS-E calibration predicted the moisture contents of the samples used in the E-H calibration). The results are summarised in Table C-16 and presented in plots in Figure C-22.

It can be seen that the predictive accuracies were reasonable although there is a bias towards predicting higher-than-reference values when using the DS-E model for prediction, and a bias towards predicting lower-than-reference values for the DS-E set when using the other two models. A probable cause for this is that the DS-E oven-determinations of moisture may involve the volatilisation of some extractives components, as discussed in Section 3.4. In contrast, these components will have been mostly removed, in the extraction process, prior to the determination of DS-E Dishes and E-H moisture contents. These later calibrations will therefore tend to under-predict the “moisture content” of the DS-E samples since these will not consider the loss of extractives in the reference moisture analysis at that stage.

It could be considered that the E-H and DS-E Dishes calibrations therefore reflect more accurate “true-moisture” calibrations since the oven determinations of moisture at these stages are more likely to only remove residual moisture in the sample and not the other components.

12.3.3.3 Limerick Spectral Set (DS)

Initially calibrations were developed for the 28 samples which had their spectra collected at the University of Limerick. Firstly, calibrations were attempted for the composition of the various lignocellulosic components (monosaccharides, lignins etc.) on an extractives-free basis. This was done so that any errors that may be present in the extractives content data would not affect the calibration providing that the ASE did the extraction accurately. This is usually the case; extractives errors tend to be a result of errors in the determination of moisture content rather than in the efficiency of the ASE.

12.3.3.3.1 Descriptions of Calibration Techniques

The Unscrambler X was used for the pretreatment of the NIR spectra, transferred from the FOSS Vision software as NSAS files, and the subsequent development of PLS calibrations.

The data for the total explained variance and for the predicted y-values (in calibration and full cross-validation) for each PLS factor were copied from the Unscrambler software to Microsoft Excel. Within this spreadsheet, and using these data, the statistics outlined in Table C-17 and subsequent Tables were calculated according to the calculations outlined in Section 6.7.

The rows with “F-“ indicate the number of factors chosen according to various factor selection criteria – Wold’s Criterion, adjusted Wold’s Criterion with 0.95 threshold, adjusted Wold’s Criterion with 0.90 threshold, the F-test for factor selection, Haaland’s test, the number of factors associated with the minimum PRESS value (all of these are described in Section 6.8) and the number of factors suggested by the Unscrambler software.

It was decided that models would be based on the number of factors chosen by Haaland’s criterion although Table C-17 and subsequent Tables do also present the RMSECV values for the model using the

number of factors associated with minimum PRESS ($RMSECV_{MP}$). It can be seen that, in many cases, Haaland's criterion involves fewer PLS factors than are required to obtain the minimum PRESS; hence it is considered that the criterion is a reasonably conservative method for factor selection. The various Wolds' criteria were not considered suitable given that these were subject to local minima in the PRESS value in situations where the global PRESS may be significantly lower (as discussed in Section 6.8).

As can be seen in Table C-17 and subsequent Tables, sometimes no, one, or two pretreatments (Pre.) were employed. The "Specific" row provides details about the pretreatment method. These details included the SG conditions and, in the case of scatter correction methods such as SNV, MSC etc., the wavelength region used to calculate the transformation coefficients and over which the transformations were made. This region is expressed in units of 10^3 nm (i.e. μm). The specific section also includes, when relevant, indications of the type of MSC, EMSC, or SNVDT model used. For example, the SNVDT transformation with a "Specific" of 1.1-2.5,3 means that the transformation took place over the wavelength region 1100-2500 nm and that a 3rd order polynomial was used for the detrend (see Section 8.3.1). The MSC "Specific" notation also includes the wavelength region used as well as a letter, with this letter representing the type of MSC model used. "F" represents full MSC, meaning that the a and b coefficients were calculated and used in the transform (see Section 8.1.1).

For EMSC the letter "F" also represents full, meaning that the channel number, squared channel number, and squared spectrum were all modelled and subtracted in the transform (see Section 8.1.1.1). The letter "A" represents EMSC scenario "F" with the exception that the squared spectrum was not included. The letter "B" is similar to EMSC scenario "A" with the exception that the squared channel number was also not included. The Savitzky Golay notations follow the principles outlined in Section 8.2.1.

Table C-17 and subsequent Tables present the RPD and RER statistics. In instances where only cross-validation was employed to test the calibrations the $RMSECV$, instead of the $RMSEP$, was used to calculate these statistics. This is indicated by a subscript of "CV" in their notations. The final row also includes the $RMSECV$ as calculated according to a model using the number of factors that are associated with the minimum PRESS value. The notation $RMSECV_{MP}$ is used for this statistic. The RER and RPD values are clearly dependent on the range and standard deviation of the reference analytical values for the samples and may change if samples are excluded from the calibration (as will other tests, such as the

Haaland's criterion, that depend on n , the number of samples). These changes in n , SD, range, and other properties are accommodated in all calculations used when samples are excluded.

Data are also provided for the slope and offset/intercept of the regression line for cross-validation and for the offset/intercept of the regression line in calibration. No statistic is provided for the slope of the regression line for calibration since this value is equal to the R_{calib}^2 value. Also, no values are provided for the bias in the calibration dataset because these values were always extremely low.

12.3.3.3.2 Glucose Content, Extractives Free (GLU_EF_SRS)

The most clear observation regarding the GLU_EF_SRS calibration, and for all calibrations involving the lignocellulosic components of the 28 samples in the DS dataset, was that sample 50019 was a clear outlier in plots of predicted y vs. reference y . This is illustrated for the case of a PLS regression model for GLU_EF_SRS in Figure 12-5 (b). The incidence of sample 50019 also being an outlier in a PCA of the chemical data is discussed in Section 12.3.1.3. This sample has therefore been removed from all lignocellulosic calibrations.

For GLU_EF_SRS, numerous spectral pretreatments were employed prior to PLSR, and the results of regressions involving these (and experiments with no spectral pretreatments) are summarised in Table C-17. Firstly, a calibration (*1 in Table C-17) was sought using all wavelengths with no spectral pretreatments. It was found that, not only were the predictive abilities (e.g. RMSECV, R_{CV}^2) significantly poorer than when spectral pretreatments were employed, but it also took more factors to get to the optimal results. These optimum results would never be reached if the Wold's Criteria had been used since, as shown in Table C-17, these all choose only 1 factor. This is due to the erratic Explained Variance plot for the CV, shown in Figure C-23 (a), which experiences a drop in the explained variance after Factor 1. For the no-pretreatments test, limiting the PLS spectral region to 1100-2500 nm, however, did improve the situation greatly (see *2 in Table C-17), leading to a lower RMSECV with fewer factors.

Including the 400-1100 nm spectral region in scatter correction pretreatments (e.g. SNV, MSC – see *3 and *5 in Table C-17) also tended to result in more complex and less accurate models compared to the situation when these treatments only considered the 1100-2500 nm region (see *4 and *6 in Table C-17).

Various SG derivatives were also tested (*8 to *15). These covered a variety of derivative orders, smoothing polynomials, and smoothing points. Calibration *8 involved PLS calibration over the whole wavelength region (400-2500 nm) of a first derivative spectral set. It was noticed in a 2-dimensional score plot of this regression (Figure C-23 (b)), and in similar score plots of many regressions involving the visible region as well as the NIR, that sample 50047 was an outlier. However, in scores plots of similar models that instead used only the 1100-2500 nm spectral region (e.g. the F1 vs. F2 scores plot for calibration *9, Figure C-23 (c)), this sample was no longer an outlier. The unusual spectral characteristics of this sample in the visible region are discussed in Section 12.3.2 and are considered to be responsible for it being an outlier in full spectral region regressions. However, since calibrations involving the full spectral region were poorer this phenomenon was not an issue since models involving the 1100-2500 nm region were preferred.

Moving from the 1st to 2nd SG derivative was seen to result in a poorer model (*12 vs. *10 in Table C-17). Furthermore, the derivate treatments typically required more factors than the SNV and MSC calibrations previously undertaken for the same spectral regions (e.g. *3 to *7 vs. *8 to *15).

The next three calibrations involved the use of SNVDT using different order polynomials for the DT transform (2 to 4 orders for *16 to *18, respectively). SNVDT using a second order polynomial (*16) demonstrated the best predictive ability. After these SNVDT calibrations, three EMSC calibrations were undertaken (*19 to *21). Calibration *21 involved EMSC scenario B and gave one of the lowest RMSECVs at min PRESS (0.5181%), but this took 11 factors to reach.

The final two calibrations, *22 and *23, involved combinations of the two best pretreatments (according to RMSECV using the number of factors determined by Haaland's criterion). These were SG-1,3,7,7 and SNVDT using a 2nd order polynomial for the detrend. Calibration *23 used these pretreatments in that order and calibration *22 used the reverse order. However, it was found that neither calibration resulted in a simpler and more accurate model.

It was decided that calibration *16 (highlighted in bold in Table C-17), which involved an SNVDT transform using a second order polynomial for the DT, was the most suitable, based on its predictive ability (RMSECV of 0.547%), and the relatively low number of factors required (4). Figure 12-5 (a) shows the predicted vs. reference y for this calibration (using the cross-validation values).

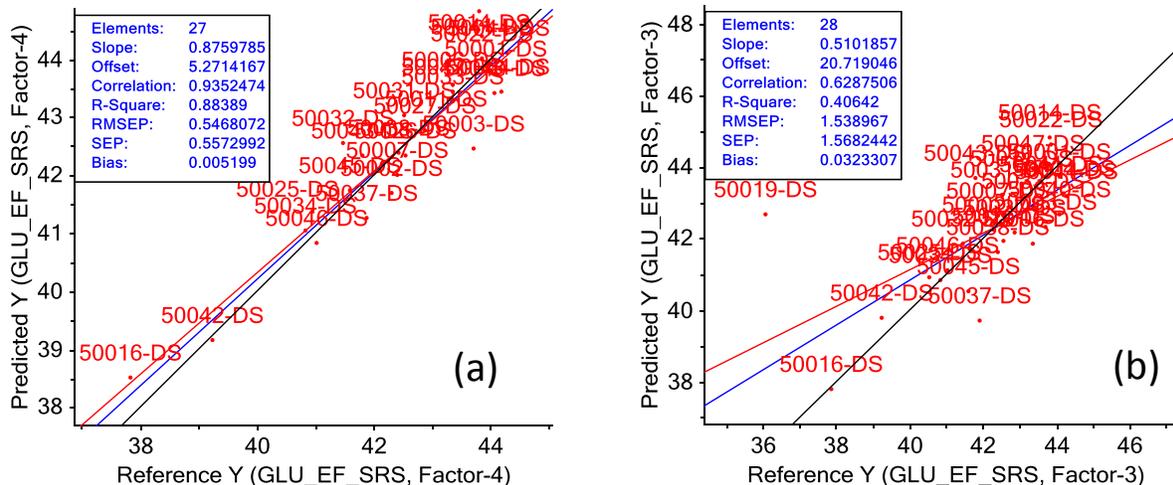


Figure 12-5: Glucose regression plots. (a) The predicted *GLU_EF_SRS* content of the 27 DS samples using model *16 in Table C-17. The red points represent the location of the cross-validation-predicted concentration; (b) the same calibration but including sample 50019 in the regression. The red line represents the regression line in cross validation, the blue line represents the regression line in calibration, and the black line represents a 1:1 (predicted:reference) line.

The RMSECV for calibration *16 is reasonable, however, it can be seen that the RER value of 11.57 is below the value of 15 that is considered (see Section 6.11.2) to represent a calibration that is suitable for quantitative analysis. The RER value for *16 is above 10, however, meaning that the calibration is suitable for screening purposes. As mentioned in Section 12.3.3.1, the low range in the compositional values for the samples analysed make it extremely difficult to attain an RER value over 15. This minimal variation is also the reason why the RPD value for *16 is only 2.83.

Figure C-24 (a) provides the 27 spectra (sample 50019 excluded) that resulted after the SNVDT transform used in calibration *16. It can be seen that the resulting spectra exhibit much less variation and apparent scatter effects than the raw spectra presented in Figure 12-2 (a). Figure C-24 (b) provides the total explained y variance with increasing numbers of PLS factors, for both calibration and full cross-validation, for calibration *16. It can be seen that it is the first three factors that explain most of the y-variance. Comparing X and y, the first factor explained 98% of the X variance compared with 59% of the y variance (in calibration sets). Figure C-24 (c) presents a scores plot of F1 vs F2. It shows that sample 50047 is somewhat of an outlier due to its high score on F2, even though only the 1100-2500 nm region is considered. Figure C-24 (d) presents the F3 vs. F4 scores plot, where 50037 is outlying.

Figure C-25 (a) presents an influence plot for calibration *16, it can be seen that, while no samples have high Q-residual scores, sample 50037 lies outside the Hotelling T^2 limit ($\alpha = 0.05$) and samples 50047 and

50016 lie close to this limit. All other samples are far from the limit. Figure C-25 (b) presents a plot of y -residuals vs. predicted y . It can be seen that there is no apparent relationship between the two, or structure to the residuals.

12.3.3.3.3 Xylose (EF), Arabinose (EF), Mannose (EF), and Rhamnose (EF)

According to the R_{CV}^2 values, no accurate calibration could be developed for xylose. Table C-18 shows the statistics (numbers *1 to *3) for PLSR on the raw spectra and on the best two pretreatments, of a much larger range that were tested. However, none of these achieved an R_{CV}^2 of over 0.5. The limited variation in the xylose content of the samples, as discussed in Section 12.3.3.1, is the probable cause of these poor calibrations.

The situation was somewhat similar regarding the PLSR for ARA_EF_SRS although the predictive abilities of the calibrations (*4 to *11 in Table C-18) were slightly improved. Interestingly, for this constituent the best predictive ability came from using the raw spectral data (*6), although a very large number of factors were required to achieve this calibration (18). Once again, the lack of variation in the reference chemical data (the SD was only 0.11% and the range 0.44%) was a problem in developing accurate calibrations for arabinose. An RMSECV of 0.07% for arabinose (as obtained for calibration *6 in Table C-18) is extremely low, however, and, if reproduced in datasets with wider variations in the arabinose content, could lead to much larger RER, RPD, and R_{CV}^2 values.

Table C-18 also includes the models (*12 to *17) that were developed for the rhamnose (EF) content. Given that the total concentration-of (average of 0.11% on an extractives-free basis) and variation-in this constituent (SD of 0.01%) were so low it is not surprising that the resulting PLS regressions were poor. The same is the case for mannose (*1 to *7 in Table C-19).

12.3.3.3.4 Galactose (EF) Content

The results of several calibrations are included in Table C-19, numbers *8 to *21. Interestingly, it can be seen that, of all the calibrations tested, the PLSR (*8) on the raw spectra and the full spectral region

(400-2500 nm) was the most accurate with a RMSECV of 0.0428% and an R_{CV}^2 of 0.7883, using 7 factors. The best calibrations after this involved the first derivatives over the 1100-2500 nm region (*13 and *14). These first derivative calibrations, which both provided an RMSECV of 0.049%, required only 4 factors compared with the 7 needed for the raw spectra and full spectral region. As can be seen in Table C-4, the range and standard deviation of the galactan-EF data were small. However, the coefficient of variation (CV) of this dataset (i.e. the standard deviation divided by the mean) was 11.2%, significantly larger than the CV for xylose (3.1%), glucose (4.60%) and arabinose (4.6%). Furthermore, expressing the range in terms of a percentage of the average gave galactose a value of 34.8%, much greater than xylose (16.0%), glucose (19.5%), and arabinose (18.26%). It is possibly this greater relative variation in the galactose dataset that enabled higher R_{CV}^2 compared with xylose and arabinose.

12.3.3.3.5 Ash Content

Regression statistics for the ash calibrations are presented in Table C-20. It appears from these data that the wavelengths shorter than 1100 nm make an important contribution. This is demonstrated in the first few calibrations, involving no pretreatments, where the RMSECV for the 1100-2500 nm region (*2) is greater than the equivalent value for other calibrations either involving the full spectral region (400-2500 nm, *1) or the visible-only regions (400-750 nm, *3), although the 1100-2500 nm region calibration did require fewer factors.

This introduced a difficulty into the decision of how to progress with evaluating various spectral pretreatments. It is usually the case that using scatter correction methods, such as MSC and SNV, and incorporating the visible region in the transform will not result in “cleaner” less scattered spectra. This is considered to be due to the excessive noise that can exist in the visible region. Firstly, these scatter correction methods were tested on the 1100-2500 nm spectral regions (e.g. *5). It was found that the resulting PLS1 calibrations were poor predictors of sample 50047. For example calibration *5 gave a residual of -2.99% for this sample, more than 1% greater than the residuals for any other sample, and equivalent to a residual *t*-test value of 2.9. Removing this sample improved the calibration significantly, as shown in *6. Also, removing high residual sample 50007 improved the calibration still further with the RMSECV falling to 0.63% (*7).

However, samples should not be removed unless it is certain that they are X- or y-outliers. For several calibrations involving the whole spectral region the residual for sample 50047 was not high (for example, the residual for this sample in calibration *10 was 0.6%). A test was made whereby the MSC transform was applied to the whole spectral region (*13) and, while there was not the same visual improvement in the spectra as compared with a transform only focussing on the 1100-2500 nm region, the resulting calibration was more accurate (although more factors were required).

The next test (*15) involved applying the MSC transform to the 1100-2500 nm region, but carrying out the PLS regression over the whole region. This resulted in a further improvement of the calibration. The resulting spectra after this transform are shown in Figure C-26 (d). The first derivative of this was taken (*16) which helped to reduce the number of PLS factors used for only a minor loss of accuracy. Removing sample 50007 from this set was tested to see if the calibration would improve. It was observed that this sample demonstrated high residuals in many other calibrations and could be considered to be a predicted y vs. reference y outlier due to analytical error in the determination of the ash content. Taking this sample out (*17) did result in a more accurate model (RMSECV of 0.77% compared with 0.93% before).

The unusual spectral absorbance values, over the visible region, of sample 50047 have been discussed in Section 12.3.2. Deciding which calibration to choose from those presented in Table C-20 will depend on whether the samples for which the ash content is to be predicted bear resemblance to sample 50047 or to the other samples in the dataset. If the latter is the case then calibration *7 appears the most accurate of all those developed. However, should samples more closely resemble 50047 then sample *16 should be chosen. This could be considered to be the “safest” calibration for future prediction because it does not exclude any samples. Figure C-26 (a) presents the predicted y vs. reference y for *16, Figure C-26 (b) presents the same plot for *7 and Figure C-26 (c) presents the same plot for *5 (which is equivalent to *7 but with no samples excluded).

Calibrations were also carried out for the, extractives-free, ethanol insoluble ash content (EIA_EF) Regression statistics for four of the numerous calibrations that were tested are provided in Table C-21. It was noted in these calibrations that there was no significant advantage associated with including the 400-1100 nm spectral regions in calibrations. It appears that any ash components that may have been influencing the spectra in this region must have been removed in the ethanol extraction procedure. This meant that developing the calibration was simpler given that only the 1100-2500 nm region needed to be focussed on. It can be seen, however, from the data in Table C-21 that the predictive abilities of the

models for this component were poorer than many of those models developed for total ash; demonstrated by the higher RMSECVs and lower R_{CV}^2 values.

Another set of calibrations were attempted for the (extractives-free) acid insoluble ash content, i.e. the ash component that remains after acid hydrolysis. This is an important property with regard to many biorefining technologies. The results of these calibrations are presented in Table C-21 (*5 to *13). Numerous calibrations were attempted and it was noted that sample 50047 was consistently outlying. As shown in Figure C-27 (a), the NIR model clearly predicts a much higher AIA value than that which was determined in the laboratory. It should be noted that this reference analysis was carried out in duplicate and the duplicate values were similar to each other. As mentioned with regard to total ash, however, sample 50047 is problematic in calibrations for ash, particularly when excluding the visible region. In this case it was considered feasible to also exclude this sample, resulting in calibration *10. This calibration was selected in preference to *13 due to its simplicity (1 factor vs. 7). The predicted vs. reference y plot for calibration *10 is included in Figure C-27 (b). Calibration *13 also excluded sample 50037 which had an AIA_EF content of 8.07%, which is the highest of any of the samples in the DS dataset (sample 50017 was not included in this set). It was considered that calibration *10 was also superior for keeping this sample, and thereby extending the range of prediction for the calibration, even though it was somewhat outlying (see Figure C-27 (b)).

12.3.3.3.6 Klason Lignin (EF), Acid Insoluble Residue (EF) and Acid Soluble Lignin (EF)

Results for different calibrations for the KL_EF content are provided in Table C-22. Ultimately a SNVDT (2nd order polynomial) pretreatment was chosen (*5) since this provided among the best predictive ability of all the KL calibrations in Table C-22 and was also a good predictor for many other constituents. The R_{CV}^2 of 0.7883 for this constituent is far from ideal; however, the RMSECV of 0.321% is reasonable, particularly given that the SEL for KL_EF was virtually the same (0.32%).

The acid insoluble residue (AIR) content is equivalent to the KL content plus the AIA content. Numerous calibrations were attempted for this on an extractives-free basis and the results of these are presented in Table C-23. Similar trends to the KL and AIA calibrations were found, particularly relating to the presence of samples 50037 and 50047 as outliers. Using the chosen pretreatment technique for KL_EF and AIA_EF (i.e. SNVDT with a 2nd order polynomial) resulted in *11, the reference y vs. predicted y plot

of this is included in Figure C-28 (a). It can be seen that the predicted AIR_EF content for sample 50047 is higher than the reference value (to be expected given the similar trend for AIA, Figure C-27 (a)). However, in this AIR calibration sample 50037 is more outlying than it was in Figure C-27 (a) and it has a significant effect on reducing the slope of the cross-validation regression line to a value much less than 1. It can be seen in Table C-2 that the acid soluble ash content (i.e. the ethanol insoluble ash (EIA) content minus the AIA content (whole mass basis)) for this sample is negative. Clearly this is impossible and it suggests that there may have been a mistake in the laboratory determination of either the EIA or the AIA. It therefore seems reasonable to exclude this sample on the basis of laboratory error. The predicted y vs. reference y of the calibration that resulted after the exclusion of these samples, *12 in Table C-23 is included Figure C-28 (b). It can be seen that the slope of the (cross-validation) regression line is much closer to 1. It should be noted that four factors would have been selected according to the minimum PRESS criterion and the criteria used by the Unscrambler software, which would have resulted in an RMSECV of 0.407% and a R_{CV}^2 of 0.920. Nevertheless, the RMSECV of 0.462% for *12 is reasonable. Also, since the range and SD in AIR values are greater than those for KL, the resulting RPD and RER statistics are much improved. For example, the RER increases from 8.42 for KL_EF model *5 to 12.93 for AIR_EF model *12.

Regarding the extractives-free acid soluble lignin content (ASL_EF), numerous calibrations were tested covering a range of pretreatments (some of these are provided in Table C-23, *10 to *15). However, none of these gave an acceptable R_{CV}^2 . This is most likely due to the extremely limited variation in this component between all samples (see Table C-3).

12.3.3.3.7 Extractives Content

Table C-24 provides the regression statistics for numerous calibrations that were undertaken for the extractives content using either the EXTR_PD or EXTR_CV data.

EXTR_PD calibrations involving the NIR and the visible region (*1, *3, and *7) and calibrations involving only the visible region (*5 and *6) were tested. These were undertaken because it was considered that the colour of the sample may reflect, to some degree, the extractives content. However, it can be seen that, with the exception of the calibrations involving no spectral pretreatments (*1 vs. *2), the

predictive ability when using the full 400-2500 nm spectral region was poorer than when the 1100-2500 nm region was used. Furthermore, there was no predictive ability when using just the visible region.

The use of a full MSC transform to the 1100-2500 nm spectral region (*8) resulted in an improvement in the simplicity of the model; only 2 factors were required to attain R_{CV}^2 and RMSECV values comparable to the best of the previous treatments (*1 to *7). However, it was clear in a Predicted vs Reference plot of this treatment that sample 50025 was an outlier. This was also the case in the calibrations developed previously but, in the case of the MSC transform calibration, the t test for the residual was greater than 2.5 whereas in previous treatments it was slightly below this level. The presence of this sample as an outlier is represented by the encircled blue dot in Figure C-29 (a). Various other pretreatments were tested, with and without sample 50025, and some of these are shown in Table C-24. Finally, the SNVDT pretreatment, using a third order polynomial, and a PLS calibration over the range 1100-2500 nm, and excluding sample 50025, was chosen (*18). This model provided the lowest RMSECV of all the methods (0.41%) and it achieved its optimal results in a low number of factors (2). It was also considered an advantage that the pretreatment method is only based on each spectrum (rather than with MSC and EMSC where the mean spectrum of the data set is used). The RMSECV for this model is reasonable given that the SEL for EXTR_PD was 0.40%.

The best pretreatments were then tested in calibrations for the extractives content as measured via the weight in the collection vials (i.e. EXTR_CV, *19 to *22). Since 50047 was missing CV data it was excluded from this data set, but sample 50025 was included in most cases. It was found that, for *19, sample 50025 had a large residual of 0.5% (corresponding to a t-test for the residual value of 1.9), but that sample 50026 had a larger residual of -0.86% (corresponding to a t-test for the residual value of 2.7). When this sample was removed the predictive ability improved. In the end it was found that the best pretreatment (*22) was the same as for EXTR_PD (i.e. SNVDT using a third order polynomial), but ideally the EXTR_CV data for sample 50026 should be removed from this regression.

Since the EXTR_PD data are those used to correct extractives-free data to whole mass data, calibration *18 in Table C-24 will now be discussed. Figure C-29 (b) plots all 28 DS spectra after this SNVDT treatment. In this calibration Factor 1 explained 99% of the X variance but only 2% of the y variance, while F2 explained 1% of the X variance and 80% of the y variance. Figure C-30 (a) presents the F1 vs. F2 scores plot. It can be seen that there is quite an even spread of the samples in this plot. Clearly F2 is the important factor in this plot with regards to the extractives content and the distribution of samples

along this axis makes sense in this regard; for example sample 50003 has a low score and sample 50006 a high score, reflecting their low and high relative extractive contents, respectively (see Table C-2). Figure C-30 (b) presents an influence plot for this 2 factor model. It can be seen that all samples are within the limits and samples 50016, 50037, 50042 and 50047 are much more influential than the others. Figure C-30 (c) presents a y residuals vs. predicted y plot; the lack of structure to this plot indicates that the residuals are not influenced according to concentration.

12.3.3.3.8 Lignocellulosic Components on a Whole Mass Basis

Table C-25 presents the results of calibration for the whole-mass-basis concentrations of the lignocellulosic components of the DS samples. These calibrations used the best conditions that were identified for the calibration of the corresponding component on an extractives-free basis. The best calibrations for each constituent are highlighted in bold. If the calibrations in this Table are compared with the chosen (in bold) calibrations for each component (on an extractives-free basis) in previous Tables, it can be seen that the values for the R_{CV}^2 , RMSECV and the number of factors are similar in most cases for the constituents that could be modelled with reasonable accuracy. There was an improvement in the calibration for xylan content – RMSECV decreased from 0.429% in the XYL_EF_SRS model to 0.367% in the XYL_SRS model, while the R_{CV}^2 increased from 0.398 to 0.638. While this is an improvement, the accuracy of the whole-mass model is still far from ideal.

Figure C-31 shows the regression coefficients plot over the wavelengths 1100-2500 nm for AIR, KL, TOT_SRS, GLU_SRS and Ash constituents. These represent the major constituents for which the best R_{CV}^2 values were obtained, on a whole mass basis. All of these models were based on the SNVDT pretreatment over the wavelength region 1100-2500 nm and using a second order polynomial for the detrend.

Upon visual inspection there are several apparent key trends. Firstly, the regression coefficients plot for TOT_SRS tends to follow that of GLU_SRS, an understandable phenomenon given that glucan is the greatest contributor to total carbohydrate and the high positive correlation seen between these two values (see Table C-5). Secondly, the regression coefficients plot for AIR much more closely resembles the ash regression line than it does the KL regression line. This makes sense since, despite KL being the

main component in the AIR, it is the ash component that represents the greatest variability and which correlates most with the AIR values (see Table C-5).

Looking at the KL regression coefficients plot, it can be seen that there is major peak at around 2260 nm. There is also a peak at 1667, close to the 1672 nm identified as representing the 1st overtone of a C-H stretch in the aromatic rings of lignin (Shenk et al., 2008). There is also a peak at 1434 nm; this is most likely due to the first overtone of the O-H stretch in the phenolic hydroxyl groups.

For GLU_SRS and TOT_SRS there is a broad peak around the 1500 nm region; this is most likely due to the first overtones of O-H stretches for crystalline cellulose (1480 nm) and for semi-crystalline cellulose (1488 nm), as well as the O-H stretch first overtones of interchain hydrogen bonds (Tsuchikawa and Siesler, 2003b, Tsuchikawa and Siesler, 2003a). Another peak for the GLU_SRS curve occurs at around 2090 nm, considered to be an O-H combination absorbance band in polysaccharides. There is also a peak at around 1930 nm for GLU_SRS and the regression coefficients here are even greater for TOT_SRS. This wavelength region has been associated with O-H stretch and OH bend in polysaccharides (Shenk et al., 2008).

Interestingly, regarding the regression coefficients plot for total ash, it appears to follow an inverse relationship to the GLU_SRS and TOT_SRS plots, particularly in the regions with the largest peaks/troughs. Given the correlation between GLU_SRS and Ash contents (Table C-5) it is possible that the NIR absorbances by molecular bonds in cellulose, and possibly hemicellulose, are being used to indirectly model the total ash content since this constituent does not directly absorb NIR radiation. A test for this hypothesis would involve the model prediction of ash content for a sample that has a relatively large ash content but where this does not affect the glucose content, i.e. a sample that would be an outlier in a glucose content vs. ash content scatter plot. Examining such a plot for the 27 DS samples it appeared that sample 50025 was the furthest from the regression line. In the ash calibration this sample does exhibit a higher than average residual (predicted y vs. reference y) but not the largest residual of the set.

12.3.3.4 Calibrations Involving the DB, DW, DU and WU Spectra Sets

12.3.3.4.1 Discussion

Table C-26 and Table C-27 include some of the calibrations that were tested for a range of chemical constituents for the DB spectral set. Table C-28 and Table C-29 provide the same, but for the DW data set, while Table C-30 provides the results for the DU dataset. Table C-31, Table C-32, and Table C-33 provide regression statistics for calibrations attempted on the WU dataset. All the constituents in these Tables are determined on a whole-mass basis.

Typically, the best calibration found for the DS dataset was applied to the DB dataset. Due to the significant change in particle size structure for the DW dataset a wider range of calibrations were tested here to examine if these could help correct for the added complexity of the sample at this stage. The best calibrations identified for this dataset were then applied to the DU dataset, and new calibration conditions were only attempted when a significant drop occurred in predictive ability. Since the WU dataset represented another significant change in the structure of the raw NIR spectra, much more time was spent experimenting with a range of different calibration conditions in order to find the best for predictive accuracy.

The most accurate calibrations for each dataset (i.e. DS, DB, DW, DU, and WU) are then compared: in Table C-34 for the GLU_SRS, XYL_SRS and TOT_SRS constituents; in Table C-35 for the ARA_SRS, GAL_SRS, and MAN_SRS constituents; in Table C-36 for the KL, ASL and AIR constituents; in Table C-37 for the ASH, EIA, and AIA constituents; and in Table C-38 for the EXTR_PD and Carbon constituents. Comparisons are not provided for the Nitrogen and RHA_SRS constituents given their poor predictions with all models. No calibration that offered any meaningful R_{CV}^2 value was identified for the GAL_SRS constituent and the DU dataset so this field is left blank in Table C-30.

It should be noted that the conditions employed for the best calibration for each constituent for each dataset were cross-checked with all the other datasets to see if an improvement could result over the previously-identified best calibration for that dataset.

Table C-32 provides regression statistics for a whole range of new pretreatments and PLSR wavelength regions, involving the WU dataset, that were not previously attempted for the other datasets. It was

observed that the best pretreatment and calibration conditions found for the WU dataset, when applied to the DU dataset, did improve the predictive ability for the TOT_SRS component compared with all the regressions that have previously been attempted. It was noted that it was the limited wavelength region of this newly-attempted calibration, using the 1100-1800 nm region rather than the wider 1100-2500 nm wavelength region, that improved its predictive ability (with an RMSECV of 1.149% for the new calibration compared to an RMSECV of 1.334% for the calibration which used the same EMSC pretreatment but with the PLSR over the wider wavelength region). This new calibration is included in *14 in Table C-34. The best calibration conditions for TOT_SRS in the WU dataset did not, however, result in more accurate model predictions, compared against the models previously selected, for the DW or DS datasets. It did improve slightly the predictive ability of a model based on the DB dataset; the RMSECV fell to 0.914% compared against 1.037% for the calibration (SNVDT, 2nd order polynomial) previously considered to be the best for this dataset. However, Haaland's criterion selected 14 factors for this new calibration (compared with 3 for the previous calibration). It was considered that the number of factors was too great for this new model; hence the original model was retained in *12 in Table C-34.

Extended Concentration Range:

The first major point to make regarding the difference between the DS calibrations and the calibrations for the other datasets is that these datasets cover all 47 samples scanned with the FOSS XDS device (in the BSES laboratories). Due to the loss of samples however, extractives data are only available for 30 of these samples and lignocellulosic data for 29 (since outlying sample 50019 is excluded). However, total ash, carbon content, nitrogen content, and moisture content data are available for all 47 samples using the values determined at BSES.

Hence, regarding the calibrations for lignocellulosic components and extractives, the calibrations for the DB, DW, DU, and WU datasets involve all the samples in the DS dataset plus samples 50017 and 50015. As discussed in Section 12.3.1, sample 50017 has by far the highest ash content of all the samples and this has an effect on the relative amounts of many other components. Sample 50015 is much less exceptional, however, with no outlying values for any components. Unfortunately, reliable EXTR_PD data were not available for sample 50017. However, an EXTR_CV value of 5.60% was determined. EXTR_PD calibration *17 in Table C-24 was applied to a scan of sample 50017 that was taken in the small circular NIR cell prior to the extractives removal procedure. An EXTR_PD of 5.09% was predicted. Hence, given the RMSECV of the EXTR_PD calibration, the EXTR_CV content of 5.60% seems reasonable, and

was used in subsequent calibrations for the content of lignocellulosic components on a whole mass basis, allowing the inclusion of sample 50017 in these calibrations.

If sample 50017 could be well modelled then it could be expected that the R_{CV}^2 , RPD and RER values would increase due to the expanded concentration ranges brought forward by this sample. For example, in the DS dataset the range in GLU_SRS (whole-mass basis) values was 6.67% and the standard deviation was 1.54%, while in the other datasets these values were 8.51% and 1.98%, respectively. These differences are summarised in Table C-39. Where the range and standard deviation (SD) in component concentration is greater in the Others (i.e DB, DW, DU, WU) datasets this value is highlighted in bold. It can be seen that the extension in concentration range does not occur for all components – there is no change for KL, GAL_SRS, ARA_SRS, RHA_SRS, or MAN_SRS. Hence, for a constituent such as KL, a calibration involving the 29 samples will need to model the “difficult” 50017 sample but has no extension in the concentration range to help improve the R_{CV}^2 and RER statistics. In this instance poorer calibrations may be expected for the “Others” datasets.

Trends Observed Between Datasets

The differences in the summary statistics of the constituents shown in Table C-39 can help to explain many of the trends seen between the different datasets in Table C-34 to Table C-38. Taking the GLU_SRS calibrations as an example (Table C-34), it can be seen that the R_{CV}^2 , RPD and RER values increase for most of the other datasets when compared against the DS dataset calibration. However, these changes reflect the extension in the concentration range provided by sample 50017. This dynamic is reflected in Table C-39 and also in Figure C-32, where Figure C-32 (a) shows the predicted y vs. reference y for the DS calibration and Figure C-32 (b) shows the same plot for the DB calibration.

In fact, the accuracy of all the other calibrations are inferior to that of the DS calibration when the RMSECV value is considered. Regarding the differences between the DB, DW and DU datasets, their ranking in RMSECVs is as expected. DB has the lowest, reflecting the more homogeneous particle size of the DB samples. The DW dataset then has the second lowest of these RMSECVs, while the RMSECV for the DU dataset is the greatest of the three.

It would be expected that the DW calibration would be more precise given the extra work involved in trying to calculate a spectrum that is more representative of each sample; hence the heterogeneity associated with cell repacks should be reduced when compared against the DU samples. Nevertheless,

the RMSECV values for the DB, DW and DU datasets for GLU_SRS are quite similar and the slight increase in precision associated with the DW dataset compared against the DU dataset is probably not worth the extra time involved in collecting the requisite spectra for the determination of the weighted DW spectrum.

The fact that the RMSECV value for the DB dataset is higher than that of the DS dataset can be explained in several ways. Before explaining these, it should be noted that the number of factors chosen by the Unscrambler for the DB GLU_SRS calibration was 4, one more than the number selected according to Haaland's criterion and equal to the number selected by Haaland's criterion for the DS calibration. If 4 factors were to be used in the DB calibration then the R_{CV}^2 would increase to 0.901 and the RMSECV would fall to 0.618%, equivalent to the statistics for the 4 factor PLSR model for the GLU_SRS content in the WU dataset and reducing the RMSECV differential when compared with the DS calibration to approximately 0.05%. However, selecting minimum PRESS for factor determination for the DS dataset would also reduce the RMSECV of this calibration and the difference in RMSECV values (between DS and DB) of approximately 0.1% would be equivalent to the difference when using Haaland's criterion.

As explained in Section 12.2, there were differences in particle size distribution between the DS and DB fractions. The DB samples contained "fines" of a particle size less than 180 microns which were removed from the sample for the collection of DS spectra and reference chemical analysis. Therefore, fractions of the sample that were not analysed via the wet chemical procedures were presented to the NIR cell for collection of the DB spectra. If there were no chemical differences between these fractions then there should be no major influence on the calibrations developed between these datasets. However, if there were differences then the wet-chemical data would be less representative of the DB datasets than they would be of the DS dataset. This relationship could be increased further if the fine particles in the DB fractions had an increased tendency (as discussed in Section 11.1) to accumulate towards the bottom of the NIR cell and therefore be over-represented at the NIR cell window.

The other major difference between the DB and DS datasets, and indeed between all other datasets and the DS dataset, is that these other sets need to model sample 50017 which is chemically (see Section 12.3.1) and spectrally (see Section 12.3.2) different from the other samples. The complexity that this sample brings to the model could explain, in addition to the particle size heterogeneity effects, the larger RMSECVs seen for the DB, DW, DU, and WU datasets as well as the differences in loadings/regression-coefficients plots. The importance of this sample is illustrated in a Factor 1 vs.

Factor 2 scores plot for the DB GLU_SRS calibration (Figure C-32 (c)) and in a Q-Residual/Hotelling T² plot for this same model (Figure C-32 (d)). It can be seen from the influence plot that sample 50017 is the most influential in this 3 factor model and its Hotelling T² value lies outside the limit (as do the values of samples 50047 and 50006).

This high leverage of sample 50017 is repeated for the other datasets and many of the other constituents. Clearly such a situation is far from ideal; the gaps in concentration should be filled in with other samples so that the spread in concentration range is mostly even. However, this was not possible due to the limited number of samples that were available. It was decided that, despite its high influence, sample 50017 would be kept in all calibrations that were developed since its inclusion led to a more robust model that could be applied to samples whose concentrations of constituents might lie outside the range covered in the DS dataset.

WU Dataset

Regarding the calibrations developed for the GLU_SRS content of the WU dataset, an unexpected improvement in the RMSECV was observed when compared against the DB, DW, and DU datasets (although the DB RMSECV value is similar if 4 factors are used in the model). This is an extremely promising result and indicates that a reasonably accurate calibration can be developed for glucose (cellulose) content of wet bagasse samples, so reducing the need for lengthy and costly sample preparation methods.

GLU_SRS was considered to be the most important of the lignocellulosic parameters of the bagasse samples since it constitutes the largest (by mass) polysaccharide-sugar in the feedstock. For this reason a lot of time was spent examining different WU calibrations for this parameter, some of the calibrations attempted are included in Table C-31. It can be seen in this Table that some calibrations only used the 1100-1800 nm spectral region. This was chosen so that the massive influence that water has on the spectra at longer wavelengths (see the PCA loadings in Section 12.3.2.3) would not need to be modelled by the calibration. The statistics in Table C-31 suggests that this limited wavelength region can give predictions of similar precision to models using the 1100-2500 nm region, and often with a lower number of factors.

The largest RER value in Table C-31, for the GLU_SRS content, comes from calibration *9. This value of 13.56 is close to the threshold of 15 suggested for the suitability of a calibration for future quantitative predictions. Regarding the largest RER value for the TOT_SRS concentration, a value of 15.31 for *17 in

Table C-31, it can be seen that the threshold value of 15 is passed. This is an excellent result for bagasse samples that have not been processed in any way prior to the collection of their spectra.

The calibration experiments attempted for the XYL_SRS constituent using the WU spectra, Table C-32, are very interesting. It was found that superior predictions (in terms of the RMSECV value) for this dataset, compared to the DU and DW and DB datasets, could be obtained when high order (3rd and 4th) derivatives were used in the SG transform, along with relatively large regions for the application of the smoothing polynomials. Table C-32 goes through the variety of conditions that were tested. It can be seen that the R_{CV}^2 for these vary greatly. Firstly (*3 to *8), low order derivatives (e.g. 1st and 2nd) did not result in good PLSR models. A combination of higher order derivatives (*9 to *15) did improve the R_{CV}^2 , however. Furthermore, an expansion of the wavelength region used for smoothing improved the calibrations (e.g. *10 vs. *9 for the 3rd order derivative pretreatment) and this was particularly true for the 4th derivative (*12 and *15 vs. *11). However, there did appear to be a limit over which an increase in the region for smoothing did not improve the RMSECV and, instead, started to see it increase (*15 vs. *12). Also, application of higher than 4th order smoothing polynomials over these regions did not improve the models (*13 and *14 vs. *12). Importantly, these higher order SG derivatives only led to reasonable calibrations when the PLS regression only considered the 1100-1800 nm region, with the precision being less when the 1100-2500 nm region was used (data not shown in Table C-32). The most precise calibration found, of the more than 30 that were tested, is presented as *17 in Table C-32. This is quite a complex model that first required scatter correction (over the 1100-2500nm region), followed by SG4,4,30,30 and then PLSR over the 1100-1800 nm region. The resulting RMSECV of this calibration is actually less than that of the best DS model (*6 in Table C-34) but it required a total of 13 factors, an excessive number for a model with only 29 samples. It was decided that calibration *12 in Table C-32 would be selected as the most appropriate for the WU dataset since it required 4 fewer factors than this model. This number of factors is still large, however. Indeed, the XYL_SRS constituent required more factors for all datasets, when compared with the GLU_SRS and TOT_SRS calibrations (Table C-34). Hence, there are questions regarding the suitability of these calibrations and more samples, spreading the concentration range more evenly, will be needed to have improved confidence in them.

The best WU XYL_SRS calibration conditions were tested on the other datasets but these did not improve the precision of any of them. For the DU set the best XYL_SRS calibration found was poor, when compared against all the other datasets. It is unclear why this is the case. However, the increased

sample heterogeneity associated with these spectra may have been a critical hindrance in the development of XYL_SRS calibrations for this dataset.

Regarding the constituents in Table C-34 to Table C-38 it can be seen that the WU models provided a lower RMSECV than the DB, DW, and DU models for: GLU_SRS, XYL_SRS, TOT_SRS, total ash, and carbon. These are interesting results. The most likely explanation for this relative success when compared against the DU dataset is that the wet nature of the WU samples meant that the fine particles of the sample were “stuck” to the larger particles meaning that the view of the sample at the NIR cell window would be more representative of the sample as a whole than the DU sample. In the DU scan more fine particles may accumulate at the base, and this could be a highly variable phenomenon with each sample repack. That could lead to spectra of differing relevancies for each sample.

The DW spectral set was designed to reduce this effect and, in most but not all cases, the precision is improved compared to the DU set. However, the collection of DW spectra takes a significant amount of work, compared with the collection of DU spectra. Hence, the relative importance to the analyst of the constituents being examined, and any advantage that the DW calibration may bring, may need to be considered before deciding whether or not to use the DW method. The relative advantage of the WU models over the DW models for some components indicates that there still remains some variability in the DW scans – only three different particle sizes were scanned to produce the weighted DW scan, and there may have been significant particle size effects within these separate scans.

The calibrations developed for the Klason lignin content demonstrated reasonable RMSECVs for most datasets, with the exception of the DW set. The AIR calibrations, in terms of R_{CV}^2 , are more consistent between the datasets. As for the DS calibrations, an improvement in the R_{CV}^2 over the DB model is expected due to the much wider variability between the samples in AIR content, compared with the KL content.

Poorly Predicted Constituents

The “best” ARA_SRS, MAN_SRS, RHA_SRS models for the DB, DW, DU and WU datasets all demonstrated less precision than the DS models. Given the poor accuracy of the DS models for these components (RPD less than 1.5 in all cases) it is reasonable to say that no suitable model can be developed for these components at the levels of variation seen in these bagasse samples.

There are also questions regarding the suitability of the DS and DW calibrations for the GAL_SRS content given the substantial drop in precision seen in the similar DB and DU datasets. Certainly, more samples are needed to test the predictive ability of the DS and DW models for this component.

There was an interesting phenomenon whereby the best DB, DW, DU, and WU models all provided lower RMSECVs than the DS model for the ASL content. Given that the RPDs of all models are poor, it is possible that it occurred by chance, however. Once again, more samples and a wider range in component concentrations would be needed to see if this phenomenon is consistent.

Perhaps the most interesting results are those concerning the EXTR_PD component, Table C-38. For all of the DB, DW, DU and WU datasets the models generated were very poor, particularly when compared against the DS model. A possible explanation is that the particle sizes retained for the DS samples were not entirely representative of the entire original bagasse feedstock. This is important because there can be a variability in the extractives contents according to particle size (Section 16.3). It was noted that the extractives contents of the DB samples, as analysed via the 80% ethanol method at BSES, tended to be greater than those of the DS samples which may have been a reflection of the higher amount of extractives in the smaller particles of this fraction. However, NIR calibrations for the DB and DW datasets using the BSES extractives data for 40 samples did not result in improved models. The hypothesis of differential extractives contents with varying particle sizes may still be valid, however, given the poor precision of the 80% ethanol method (see Section 12.3.1.1).

It is important to note, however, that good calibrations were possible for some lignocellulosic constituents on a whole mass basis in the DB, DW, DU, and WU datasets, a result that would be difficult if the results of extractives analysis were massively incorrect. Another theory for the poor links between the UL extractives data and the NIR spectra collected at BSES is that the extractives contents of the samples may have changed during their 3 ½ years in storage.

12.3.3.4.2 Comparisons Between Data Sets Regarding Loadings and Regression Coefficients Plots

For all constituents for which reasonably precise NIR calibrations could be developed, there were significant differences between the WU dataset and all other datasets with regard to the loadings and regression coefficients plots. Since GLU_SRS is the most important constituent, the differences between datasets for it will be discussed in this Section. For this comparison the same pretreatment (SNVDT) and wavelength region for calibration (1100-2500 nm) were used for all datasets.

Figure C-33 (a) presents the X-loadings plot for factor 1 and Figure C-33 (b) the X-loadings plot for Factor 2. The loadings for all datasets follow a similar course over wavelength with regard to Factor 1 and mostly resemble the mean spectrum of the SNVDT pretreated spectra. However, in a Factor 2 plot clear differences can be seen. The WU line shares very few similarities with the DB, DW and DU lines, which are all very similar. Interestingly, the F2 X-loading for the DS set is also somewhat different. It was theorised that highly influential sample 50017 (not included in the DS set) could be responsible for this difference. However, a GLU_SRS calibration involving the DB dataset but excluding this sample provided a very similar F2 loading plot to that seen in Figure C-33 (b). The use of a different instrument for collecting the DS spectra is another explanation. However, this should not matter since both the UL and BSES devices were reference standardised.

Figure C-34 (a) presents the regression coefficients plot for the GLU_SRS content for all 5 datasets. Here the difference between the WU and all other sets is even more apparent; the regression coefficients are much greater and complex. Figure C-34 (b) uses a secondary y-axis for the DB, DW, and DU plots so that the trends over the wavelength region for these datasets can be seen and still compared with the WU dataset. Here the DS line, which is very similar to the one presented in Section 12.3.3.3.2 for the GLU_SRS (EF) content, follows a roughly similar trend to the DB, DW, and DU lines but tends to have greater coefficient values in some wavelength regions. The WU line differs greatly in shape from the others, however. Interestingly, it can be seen that high absolute coefficient values exist over the 1800-2000 nm wavelength region indicating that this region, where water absorbs strongly, is still used for prediction of the GLU_SRS content.

12.3.3.4.3 Bagasse WU Moisture Content Predictions

Table C-33 presents the regression statistics for 3 of the numerous calibrations that were investigated for the moisture content of the WU samples. It was found that in all cases samples 50023 and 50040 had high residuals; for example in one model the residual (in calibration) for 50023 was equal to 5.23%, equivalent to a t test for the residual of 3.24. The removal of these samples (*15 in Table C-33) improved the predictive of the model greatly and resulted in an RMSECV of 1.45%, which is particularly good considering that the SEL was 1.81%.

However, 45 samples is a sufficient number to allow for an independent test set. Hence 30 samples were randomly selected for the calibration set and 15 for the prediction set, and the same calibration conditions as for *15 in Table C-33 were used. The model that was developed predicted the unknown samples well, with an R_{Pred}^2 of 0.929 and an RMSEP of 1.21%. The regression coefficients plot of this model was complex with important wavelengths not just in the typical regions for water absorbances, although the greatest absolute value for the regression coefficient was observed at 1913 nm.

12.3.3.5 NIR Predictions of BSES Samples

The best WU models for GLU_SRS, XYL_SRS, and TOT_SRS and the best DB model for KL were used to predict the concentrations of these constituents for sample 50019 plus the 17 samples scanned at BSES but not analysed at UL. For GLU_SRS predictions, in a plot of inlier distance vs. the Hotelling T^2 statistic all samples were within the limits except sample 50005 which breached both. This sample had the highest prediction deviation (see Section 6.6) of 0.854%. The average deviation was 0.450% for GLU_SRS predictions and 0.498% for XYL_SRS predictions. For XYL_SRS seven samples breached both the inlier distance limit and the Hotelling T^2 limit, indicating that the calibration used was not ideal for prediction of these samples.

For KL the average deviation was 0.384%. A total of 5 samples breached the inlier distance and Hotelling T^2 limits, with samples 50005 and 50020 having by far the greatest values. These samples were also identified as highly influential inliers in the TOT_SRS prediction (it had an average deviation in prediction of 0.79%).

It can be seen in the plots in Figure C-35 (a) and (b) that these predicted samples help to improve the spread of all 47 samples across the concentration range for these components. The effect is most marked for XYL_SRS where the SD increased by a relative 31.5% (from 0.80% to 1.05%) and for KL where the SD increased by a relative 24.9% (from 0.687% to 0.858%). The relative increases were less for GLU_SRS (2.80%) and TOT_SRS (8.57%). Having these extra samples in calibration models may have helped to reduce the RMSECVs and increase the RPDs for KL and XYL_SRS, if these predicted concentrations can be trusted.

12.4 Summary

It can be seen that reasonably accurate calibrations can be developed for glucan, ash, total polysaccharide sugars, and acid insoluble residue contents of both dry and wet bagasse samples. The development of these calibrations was possible despite the limited chemical variability of the samples analysed. This variability differed between the various chemical components with the least variable components typically resulting in models with lower R_{CV}^2 values.

In comparing the results of the regressions used in this study against those carried out by other researchers (Appendix B) it can be seen that, while the R^2 values may be less in some instances, the RMSECVs obtained are often superior. For example, the GLU_SRS RMSECV values of 0.567% for the DS model and 0.616% for the WU model are lower than any of the RMSECV or RMSEP values for glucan content in Appendix B. Similarly, the RMSECVs obtained for xylose, arabinose, galactose and KL are lower than any in Appendix B. However, these RMSECVs must be taken in context, particularly with regard to the standard deviation in concentration values which was often extremely low for the minor constituents. When using the RPD or RER statistics as a comparison, the calibrations for the minor constituents, KL and xylose, appear less accurate than those of other researchers, however the glucose, total sugars and ash calibrations compare well against many of the calibrations in Appendix B.

The results of the calibrations are probably as good as can be expected given the SELs of the various components. These SEL values were reasonably low but, in some instances, were inflated due to the lack of available sample material for repeat analyses. However, the critical limiting factor in the quality of the calibrations is the limited variability of the samples analysed. Section 12.3.3.5 indicates that some of the 17 samples that were discarded and not analysed at UL could have helped to increase the RER and RPD values of the calibrations. Furthermore, all of the samples analysed in this study came from the same sugar mill and a single harvest window. A much wider sample collection strategy might allow for the analysis of more varied samples and improved calibrations.

Nevertheless, the results for the WU calibrations of glucan, ash, and total sugars are highly promising and the calibrations developed in this section offer potential in the screening of samples for wet chemical analysis. Also, the methodologies developed and selections made regarding the appropriate spectral pretreatments and PLS regression techniques have relevance to the development of calibrations for the much larger peat and Miscanthus data sets (Sections 13 and 15).

13 Development of NIRS Quantitative Calibrations for the Lignocellulosic Components of Peat Samples

In September 2008 the Author submitted a proposal to Bord na Móna (BNM) for a €42,871 research project. BNM is a semi-state owned company in Ireland whose activities include the mechanical harvesting of peat samples from Irish bogs. A key component of the proposed research involved the analysis of a wide variety of peat samples in order to determine the lignocellulosic compositions of these and hence whether peat may have value as a biorefining feedstock. The proposal was successful and research commenced in January 2009.

According to the description of work of the funded proposal 25 samples would be fully analysed. Following the analysis of these samples and the completion of the project, the Author undertook some preliminary tests involving quantitative NIRS calibrations for various lignocellulosic components and found that NIR prediction appeared possible for both wet and dry samples. However, it was considered that this was an insufficient number of samples for the development of a robust model that could be tested with samples from an independent validation set. Hence, the Author requested additional samples from BNM and the company supplied a further 28 samples.

This chapter will discuss the collection and analysis of these peat samples and focus, in particular, on the development of the NIRS calibrations. Many of the Figures and Tables referred to in this chapter are provided in Appendix D.

13.1 Background on Peat

13.1.1 Peat Formation, Classification, and Composition

Peat is a substance that forms from the incomplete degradation of organic matter. Such incomplete degradation occurs in waterlogged land where the limited availability of oxygen means that the rate of accumulation of plant debris is greater than the rate of its decomposition (Fuchsman, 1980). This accumulation of peat is a slow process, with only approximately 3 mm formed every 100 years (Doyle,

1997). The composition of the peat will vary according to the type of vegetative matter that formed it and the relationship between the water of the peat bog and the ground water system. A “low moor”, or “blanket bog”, peat is one where the peat bed is low lying and the bog water and ground water are continuous between each other. In contrast, a “high moor”, or “raised bog” peat is one that forms where the bog water is above the groundwater system. Reeds, sedges, and woody plants such as willow, birch and conifers are associated with low moor peats while sphagnum mosses, cotton grass, and heath plants are associated with high moor peats (Fuchsman, 1980). The level of plant decomposition is greater in low moor peats compared with high moor peats (Fuchsman, 1980). There are also “fen” peatlands. These are described as minerotrophic (mineral rich), and are often described as “black bogs” whereas raised and blanked bogs are ombrotrophic (mineral poor). Fens are typically located at the edges of raised bogs and along river valleys.

Peats can also be classified based on the degree of humification, i.e. the degree of decomposition of the original lignocellulosic plant materials. A Von Post Decomposition Scale of Peats (Von Post, 1924) rates the degree of humification from H-1 (undecomposed) to H-10 (no plant traces remain) and is determined via squeezing peat and its water through the fingers. Alternatively, this ten point scale can be reduced to three classes: R1 for the lowest levels of humification, R2 for the medium levels, and R3 for the highest (Barron et al., 1987).

The chemical composition of peat will depend on the relative degradation of the different chemical groups of the vegetative species. For instance, plant waxes are relatively resistant to degradation while water soluble sugars and starches are highly susceptible (Mal, 1979). Hemicellulose, cellulose, and lignin can be resistant to degradation and found in peats, but to varying degrees. In particular, while lignin is readily degraded by microbes in aerobic conditions to form the humic layers of soil, it is much more resistant to such degradation in anaerobic conditions such as those seen in peat bogs (Fuchsmann, 1980). The degradation of peat can produce humic acids, which have some polyphenol structures and with high carboxylic acid contents and can contain significant quantities of nitrogen.

As well as the carbohydrates derived from the plant materials, peats can also contain those carbohydrates that are released upon the death of the microbes in the peat. This can result in the presence of deoxy-sugars such as rhamnose and fucose in the peat samples and these are absent (or only present in minute quantities) in the vegetative matter (Doyle, 1997).

As would be expected, the proportions of cellulose and hemicellulose present in the peat will decrease with increased levels of humification, and so with increased depth in a peat bog. Conversely, the relative proportion of the lignin, or at least that which would be quantified as Klason lignin under the analytical method outlined in Section 11.5, would increase. Doyle (1997) analysed an H-5 and an H-8 peat and found cellulose (lignin) contents of 15.4% (36.4%) for the former and 2.9% (46.1%) for the latter. Hemicellulose typically decreases more slowly than cellulose with depth (Fuchsman, 1980) and this rate of degradation differential is more marked in low moor peats than in high moor peats. Table 13-1 presents data from Fuschmann (1980) comparing the lignocellulosic components in both peat types. Table 13-1 also includes the average lignocellulosic compositions of the plants growing on these bogs. Regarding fens, these minerogenic peats typically have lower hemicellulose and cellulose contents but higher lignin and ash contents than ombrogenic peats.

Table 13-1: Compositions of a high moor peat and a low moor peat, as well as that of the vegetative matter that grows on them. Data from (Fuchsman, 1980).

Component	% of Dry Matter			
	High Moor Peat		Low Moor Peat	
	Plant	Peat	Plant	Peat
Ether Extractives	1 - 5	2 - 6	1 - 3	1 - 3
Water Extractives	4 - 8	1 - 2	3 - 12	2 - 3
Hemicelluloses	19 - 31	9 - 21	18 - 21	6 - 10
Cellulose	21 - 25	12 - 19	12 - 31	0
Lignins	7 - 21	25 - 52	21 - 42	38 - 46
Protein	4 - 6	5 - 6	4 - 15	22 - 23
Ash	3 - 4	1 - 2	3 - 5	10 - 13
Total	73 - 86	73 - 87	93 - 98	88 - 91

13.1.2 Peat in Ireland

It has been estimated that peatlands originally covered 17% of the land surface of the Republic of Ireland with 65% of this area being low moor, 26% high moor and 8% fen (Hammond, 1979). In terms of geographical distribution, the low moors typically occur in the mountainous areas in the west of Ireland whilst the high moors are typically found in the Midlands (Cross, 1983). Bord na Móna owns and works 88 000 hectares of peatland, which represent approximately over 7.3% of Ireland's total peat reserve (CEN, 1999). The company harvests the bogs in the Midlands since these receive lower annual rainfall quantities and so offer more potential as combustible fuels. This represents the primary market for peat, although Bord na Móna does sell horticultural products that include peat. There are three main peat-

fired power stations in the Republic. Two of these, with a combined capacity of 225 MW, are located in County Offaly (Shannonbridge and Edenderry) and a 125 MW plant is located in Lanesborough, County Longford. Edenderry power station was purchased by BNM in 2006, but the other two facilities are operated by the national Electricity Supply Board (ESB) which buys the peat. In Edenderry BNM co-feed approximately 7% of biomass (including *Miscanthus*) with the peat and plan to increase the proportion to 30% by 2015.

BNM, in obtaining peat for power generation, harvests it as milled peat. The milling process produces an air-dried peat that is more powdery than the original peat material. In order for milled peat to be harvested the bog needs to be drained over a period of approximately 5 years. This will reduce the moisture content of the surface layers of the bog from approximately 95% to approximately 80%. The harvesting of peat is highly susceptible to the weather conditions which will influence the moisture content of the peat. Bord na Móna has a target moisture content of 45% for peat to be used in power generation. The milling process involves scraping the surface of the bog (to approximately 15 mm depth) and this peat is then allowed to air dry. To facilitate drying the layer of peat is turned over several times by use of a harrow. Once the target moisture content is reached the surface layer is pushed into ridges at the centre of the field by use of a ridger. These ridges are then transferred to larger piles that run parallel across the bog, often beside railways which are used to transport the peat to central facilities. BNM produces approximately 4 m tonnes of milled fuel peat per year with 3 m tonnes used for electricity production and the remainder for briquette production. The company also produces 1.3 m³ of horticultural peat per year; this product can have a higher moisture content than peat for combustion.

13.1.3 NIRS and Peat

There have been few publications relating to the development of quantitative NIRS calibrations for peat and the Author could not find any publications covering all of the lignocellulosic components involved in this study. Malley *et al.* (2007) examined the use of a field portable NIRS device to calibrate for the moisture content, organic matter content, and carbonate content of peat samples from three locations in Canada: a rich, intermediate, and poor fen. These categories of fens were differentiated by the number of indicator species with the rich fen having the largest number. Samples were collected from the peat surface to the mineral substrate (ranging from 0.5 to 3.5m depth) from two sampling locations

along two transects in each fen type. These core sections were stored frozen, and then cut into 10 cm long sections, each of which was used as a sample for NIRS and reference analysis. In total there were 227 samples with two thirds being in the calibration set and the remainder in the validation set. NIRS analysis involved the wavelength region 600-1690 nm. It was found that untreated spectra gave the best calibrations for moisture (which ranged between 60.2 and 92.6 %), with an R_{calib}^2 of 0.83, an SEP of 1.80%, an RPD of 2.38, and an RER of 11.60, using 10 components and with 3 outliers removed. For organic matter content it was found that the best calibrations were obtained from spectra that had been smoothed over 4 wavelength points and transformed to the second derivative with a gap of 4 points. R_{calib}^2 , [SEP], [[RPD]] ((RER)) were 0.89, [0.71%], [[3.07]], and ((14.87)) for organic matter using 6 components and with 1 outlier removed.

Beining *et al.* (2000) attempted NIRS calibrations for gross heating value (17.82-20.76 MJ kg⁻¹), carbon (49.24-57.53%), cellulose (11.46-20.06%), hydrogen (5.01-6.04%), N (0.84-2.93%), and ash (0.92-7.30%) for freeze-dried, pulverised samples of various Irish peats (both milled and non-milled). The best results were found using PLSR calibrations of original absorption data for cellulose; 1st derivatives of absorption data for heating value, C, N and ash; and 2nd derivatives for H. The following were the best R_{calib}^2 obtained: heating value 0.864, Carbon 0.975, cellulose 0.575, H 0.902, N, 0.979, and Ash 0.931. It was not clear from the unpublished paper how many samples were used or if there were separate calibration/validation sets for the PLS models.

13.2 Methodology

13.2.1 Collection of Peat Samples

The collection of peat samples comprised two phases, the first being the original project funded by BNM, under which 25 samples were collected. The second involved the supply by BNM of a further 28 samples for NIR model development.

In the first phase the peat samples were collected over two dates, the 19th of January 2009 and the 9th of February 2009. It was requested that a wide variety of different peats should be sampled (ranging from high to low levels of humification, as well as peats from stranded bogs) in order to enable a good

understanding of how lignocellulosic properties vary under these differing conditions. If certain types of peats appeared to be more attractive for biorefining purposes then these could be focused on in future studies. Such variation between the samples would also facilitate the development of robust NIR models.

On 19/1/09 the power stations at Edenderry (Co. Offaly), Lanesborough (Co. Roscommon) and Shannonbridge (Co. Offaly) were visited. Over these three sites a total of 15 bagged samples of milled peat were collected. On the 9th of February a further 10 samples of peat were collected. Information was given by BNM as to the degree of humification associated with these peats and this is provided in Table D-1. All of the samples were of milled peat with the exception of the peat from Prosperous, Co. Kildare (25016), which was sampled directly at the edge of the horticultural bog and, hence, was not milled meaning that a high moisture content would be expected.

13.2.2 Processing and Analysis of Samples

The general analytical methods outlined in Section 11 were employed for the peat samples with the following key differences:

- Only WU, DU, and DS scans were collected.
- Only 2 replicate scans of each sample were collected for the WU, DU, and DS datasets. The average of these was taken and used for model development.
- Initial attempts to reduce the mean particle size of the DU samples involved the use of the FOSS Cyclotec 1093 mill; however, it was found that only approximately 1/3 of the sample ended up in the DS particle size fraction with the remainder either present as DF samples or lost as dust to the air. The only suitable method found for the maximising the relative proportion of the DS fraction for the first 25 samples was to use a pestle and mortar. This was a highly time consuming process but resulted in approximately 70% of the original DU material ending up in the DS fraction. For the second batch of samples the SM2000 chipper, with a sieve aperture of 1 mm, was used and allowed improved productivity for the generation of DS samples.

13.3 Results

13.3.1 Reference Analytical Data

The relative amounts of various lignocellulosic constituents, expressed on an extractives-free basis, for the 53 samples are presented in Table D-2 and Table D-3. No reliable EXTR_PD (or EXTR_CV) data were obtained for samples 25024, 25025 and 25028 and there was insufficient sample remaining to repeat the extractions. The EXTR_PD PLS calibration outlined in Section 13.3.3.1 was applied to the samples to predict their EXTR_PD values so that their lignocellulosic components, which had been analysed to a satisfactory degree of accuracy, could be expressed on a whole mass basis. These EXTR_PD values, as well as those of all the other samples and the values of other components are presented, in a whole mass basis, in Table D-4 and Table D-5. Table D-6, Table D-7 and Table D-9 present histograms, with associated statistics (including the SEL), for selected components, and Table D-8 presents the correlation coefficients between some constituents for all 53 samples.

Clearly, the composition of these samples is substantially different from that of bagasse and Miscanthus samples. Total carbohydrate is significantly lower and KL is much higher. This is to be expected given the decomposition of the vegetative matter that has led to the formation of peat. There are also differences in the relative proportions of the lignocellulosic sugars. In bagasse samples, an average of 95% of the total sugars analysed for were either glucose or xylose, whereas for the peat samples this proportion was only 80%. Rhamnose and mannose are typically present in higher quantities than in bagasse/Miscanthus. These are likely to arise from microbial communities upon their death.

The absolute correlation coefficients between the different constituents are also greater, in many cases, than for the other feedstocks. Some of the most interesting relationships are listed below:

- Ash negatively correlated (-0.837) with extractives content.
- KL negatively correlated with most sugars.
- KL strongly negatively correlated (-0.923) with ASL. A positive correlation might be expected, however this negative relationship might indicate that ASL lignin is the least recalcitrant lignin component to microbial degradation. Hence, as the relative proportion of KL increases (due to

the loss of other components through degradation) that of ASL falls. This relationship is consistent with the acidic nature of bogs.

- Ash (and ASA) negatively correlated with most sugars except arabinose (0.806).
- UA positively correlated with all sugars.
- Strong positive correlations between rhamnose and galactose (0.950), mannose and rhamnose (0.957) and mannose and galactose (0.975).
- Very strong correlation between glucose and total sugars (0.992), which is significant even after considering that glucose is a major component of total sugars (66.2% on average).

These relationships all need to be considered in the context of varying degrees of humification of the peat and Figure D-1 and Figure D-2 facilitate this by presenting quantile plots for selected components for the following groups: the 13 samples classified as “Low” (L); the 10 samples classified as “Medium” (M); and the 8 samples classified as “High” (H) by BNM; as well as the three samples (labelled “3”) that appear outlying in PC1 vs. PC2 scores plots of a PCA model (see Figure D-9 (d)). These three samples also appear as outliers in the ARA_SRS, Ash, XYL_SRS, MAN_SRS, EXTR_PD and GAL_SRS histograms in Table D-6 and Table D-7 as well as in the raw spectra in Figure 13-1 (b). The other samples (the remaining unclassified samples Table D-1) are not included in these quantile plots.

Figure D-1 and Figure D-2 show the following trends with increasing degrees of humification: all sugars decrease; KL, AIR and ash increase; and there is a relatively small decrease in extractives. The “3” group is somewhat analogous, since, by not having consistent trends between its constituents, it does not appear similar to any of the assigned classes. For example, while its rhamnose, mannose, galactose and extractives contents are even lower than the “High” class samples, suggesting an even more decomposed peat, it has intermediate glucose values, amongst the lowest KL values, and by far the highest xylose and arabinose values. The ash contents of these samples are also far in excess of those of the other groups and result in the AIR values appearing normal (given the low KL content).

Quality of the Analytical Data and Suitability for Quantitative NIRS Calibrations

Table D-2 to Table D-5 present the standard deviation of the duplicate (SD) for each sample and component. Table D-6, Table D-7, and Table D-9 also present the SEL for selected components and express it as a percentage of the mean (see Section 12.3.1 for detailed explanations). For many constituents the SEL values are greater than those obtained for the bagasse and Miscanthus samples.

This is primarily due to the increased heterogeneity of the DS samples used for extraction/hydrolysis. This occurred because the particles were denser than those of other feedstocks, but the weights used (in the hydrolysis phase) were equivalent. Also, in the case of the SELs for AIR and KL, these are also increased in absolute terms due to the greater relative proportion that these constituents provide to the total mass balance.

Improved analytical methods would utilise larger volumes of peat samples. However, additional equipment would be required for this. Despite the increased SELs, the spread of samples across the concentration range (as shown in the histograms) is often superior to that of the bagasse and Miscanthus samples, which should benefit regression methods. This situation is not the case for those constituents (XYL_SRS, ARA_SRS, Ash) where the “3” samples are significantly outlying.

Given that fucose can be present in peats, initially this sugar was not used as an internal chromatography standard and selected peat samples were analysed using external standards. It was found, however, that the mass of fucose present in these peat samples was minimal (between 0.08 and 0.16%). It was determined that the analytical error associated with using external standards was greater than that associated with using fucose as an internal standard. Hence, fucose was kept as an internal standard for peat samples.

Table D-10 presents summary statistics for the sugar recovery (SR) rates of the 13 hydrolysis batches and Figure D-3 presents these as a plot with increasing batch number. There is an increase in SR rates seen from batch 659 onwards; a result of reducing the number of samples hydrolysed per batch to 4, meaning that a total of only 11 pressure tubes (rather than 13) were in the autoclave. This reduced load will lessen the time required to reach the target temperature for the secondary hydrolysis stage (121°C) and also the time required for cooling before the autoclave can be opened (at 80°C). Considering equivalent autoclave loadings, the recovery rates are also slightly higher than those of the Miscanthus and bagasse sugar recoveries. This is most likely a result of the lower quantities of sugars in peat sugar recovery solutions (SRS) compared to the bagasse/Miscanthus SRSs. Sugar loss has been observed to be a function of sugar concentration (Templeton et al., 2010).

13.3.1.1 PCA of Chemical Data

A principal component analysis was conducted for all samples using the following 10 variables: EXTR_PD, ASH, KL, ASL, ARA_SRS, GAL_SRS, RHA_SRS, GLU_SRS, XYL_SRS, and MAN_SRS.

Figure D-4 (a) shows the explained variance plot for up to 5 PCA factors with the blue line representing the explained variance in the calibration set and the red line the explained variance estimated via full cross validation. The first two PCs explained a total of 97.5% of the variance in the data set (in cross-validation).

Figure D-4 (b) shows the loadings plot for PC1 and PC2. The KL and GLU_SRS contents are primarily responsible for PC1 and Ash for PC2. Figure D-4 (c) shows a scores plot for PC1 versus PC2. Here the high ash samples of group “3” can be clearly seen as outliers with high PC2 scores. These samples (25046, 25047, 25048) also have high leverage values in a 3 PC model, as shown in the influence plot in Figure D-4 (d). These values were above the Hotelling T^2 limit in a plot of this statistic.

PCR Using Chemical Data

Using the 10-constituent PCA variables, PCR was used to predict the sample values for each variable in regressions involving all of the other variables. Summary statistics of this are provided in Table 13-2 which shows excellent predictive abilities with R_{CV}^2 values of over 0.9 for all constituents except arabinose and extractives. Regression coefficients bar charts for each of the constituents, except extractives, are provided in Figure D-5 and Figure D-6. Since the coefficients will be inflated for minor components it is also important to discuss the relative proportions that the important PCs play in explaining total Y-variance and which constituents have high absolute loading values for these PCs. As shown in Figure D-4 (b), for the 10 PC model PC1 had a large positive loading value for KL and a large negative loading for glucose, whilst PC2 had a large positive loading for ash. The PCRs that were not looking to predict these four constituents will be discussed first. For extractives, 64% of the Y-variance was explained by PC2, with this figure being 68% for arabinose. For ASL, 80% of the Y-variance was explained by PC1. The first two PCs combined explained 85% of the Y-variance of galactose, 75% of the variance of xylose, and 83% of the variance of mannose.

For glucose, PC1, which had a high positive loading for KL, explained 77% of the Y-variance and PC2 (a high positive loading for ash) explained an additional 15%. The ash content was mostly explained by PC2

(79% of Y-variance) for which extractives, glucose, and KL all had high positive loading values. For KL, 91% of the Y-variance was explained by the first two PCs and for both of these PCs ash and glucose had high positive loading values.

Table 13-2: Statistics for the PCR for a range of peat constituents where the predictive variables are the other constituents in this table.

Component	# PCs	R^2_{Cal}	RMSEC (%)	Offset	R^2_{CV}	RMSECV (%)	Slope	Offset
EXTR_PD	5	0.779	0.595	1.433	0.734	0.655	0.735	1.660
ASH	5	0.952	0.995	0.270	0.940	1.110	0.941	0.331
KL	5	0.980	1.058	1.116	0.975	1.206	0.980	1.162
ASL	4	0.926	0.154	0.154	0.909	0.173	0.914	0.179
ARA_SRS	7	0.864	0.096	0.056	0.777	0.128	0.787	0.081
GAL_SRS	7	0.974	0.094	0.060	0.967	0.109	0.952	0.081
RHA_SRS	7	0.960	0.071	0.022	0.951	0.083	0.958	0.032
GLU_SRS	7	0.977	0.601	0.345	0.966	0.729	0.968	0.387
XYL_SRS	7	0.944	0.187	0.129	0.909	0.239	0.933	0.148
MAN_SRS	7	0.975	0.090	0.027	0.9657	0.112	0.974	0.037

13.3.2 NIR Spectra and Their PCA

Figure 13-1 (a) presents the WU-A spectra of the 53 peat samples, Figure 13-1 (b) presents the DS spectra for these samples and Figure 13-1 (c) shows the WU, DU and DS spectra for one peat sample. As with the spectra of sugarcane bagasse (Section 12.3.2) and Miscanthus (Section 14.2) samples, the peat WU scans exhibit large absorbances in the regions of 1450 nm and 1930 nm, a result of the significant absorbances of NIR radiation in these regions by the molecular bonds of water. However, there are clear differences between these peat spectra and those of bagasse and Miscanthus. For instance, the absorbances in the visible region are significantly greater, unsurprising given the dark colour of most peat samples. Additionally, there appears to be less detail in these untreated spectra, when compared against those of Miscanthus and bagasse samples. For the DU and DS samples the spectra between 500 nm and 1300 nm appear as a smooth curve with little apparent deviation along this region. Even at longer wavelengths, there are fewer peaks and troughs than for the spectra of the bagasse and Miscanthus samples.

Figure 13-1 (b) shows three outlying spectra with greater absorbance values than the other samples. These three spectra come from the samples of group “3” (samples 25046, 25047, and 25048) discussed in Section 13.3.1. However, apart from baseline shifts, the differences between the spectra of the

various peat samples appear less substantial than the differences between the different bagasse/Miscanthus samples. Upon closer examination, though, there are some differences. These include different slopes in the 500-1300 nm curve and also in the undulations between 2000 and 2250 nm. These differences are more clearly illustrated in the MSC-corrected spectra, Figure 13-1 (d), where the deviation of some spectra can be observed.

Regarding Figure 13-1 (c), it can be seen that there is little baseline variation between the DU and DS scans, particularly when compared to the differences seen between these scans in Miscanthus and bagasse samples. This is to be expected given that the mean particle size of the sample in the DU (and WU) state was significantly less than in those other feedstocks, and much closer to that of the final DS fraction.

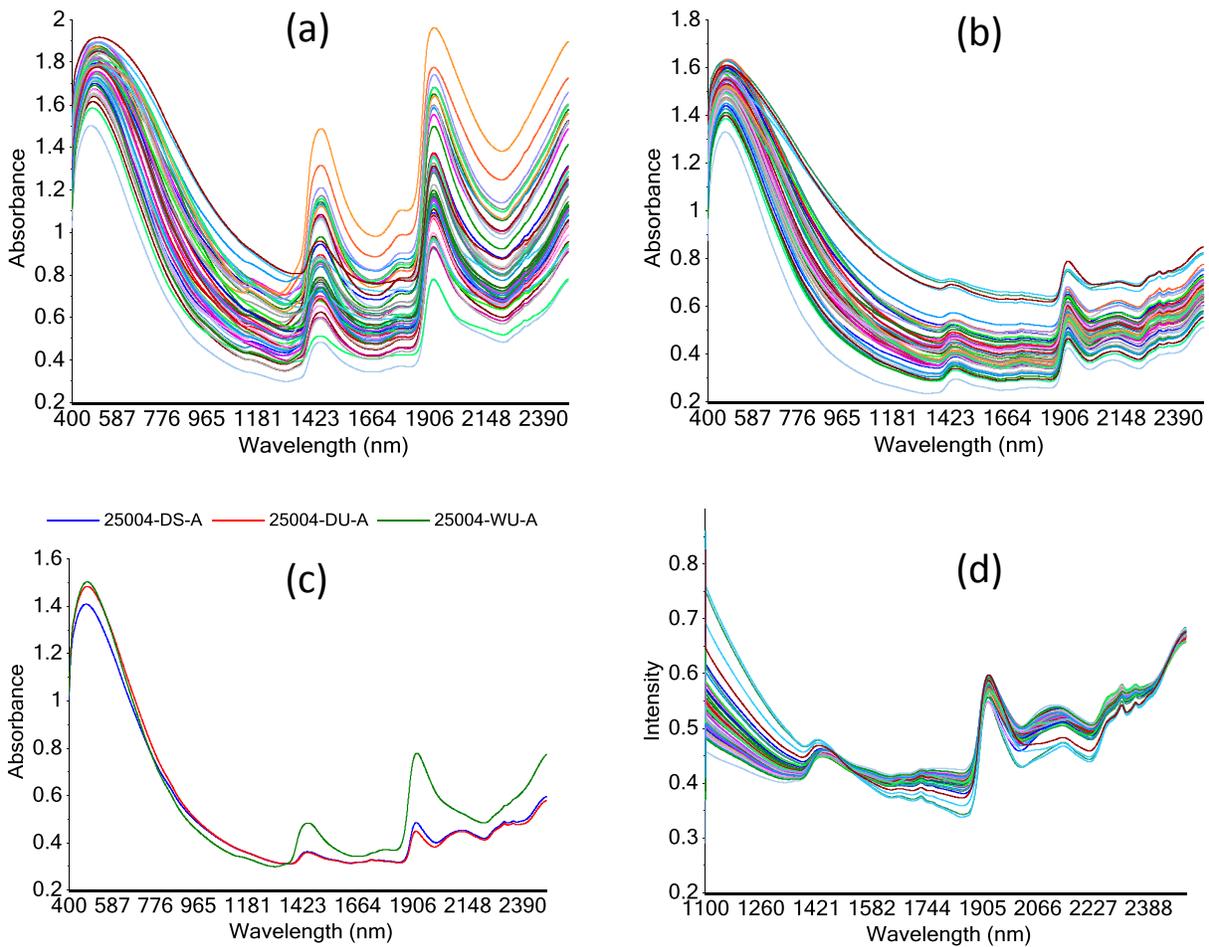


Figure 13-1: Spectra of peat samples. (a) The WU-A spectra of 53 peat samples; (b) The DS-A spectra of 53 peat samples; (c) The DS-A (blue line), DU-A (red line), and WU-A (green line) spectra of peat sample number 25004; (d) represents the 1100-2500 nm section of (b) after a full MSC transform.

Figure D-7 shows an explained X-variance plot, under full cross validation for the 400-2500 nm PCA models of the WU, DU, and DS datasets and Figure D-8 shows a similar plot for PCA models using the 1100-2500 nm. All spectra were transformed to the second derivative (SG2,2,14,14) prior to PCA. As expected, the first PC of the WU dataset accounts for much of the spectral variation provided by the water content of these samples as reflected in the loadings plot for this PC, Figure D-9 (a). The second PC of the NIR-region WU model also demonstrated high absolute loading values in the region around 1900 nm but the peaks were inverted. In contrast, the first PC of the DS (Figure D-9 (b)) and DU models demonstrates the highest loading values at the shortest wavelengths in order to account for variations in the curve of the visible absorbance peak seen in Figure 13-1 (b). The next few PCs of the DS and DU full-wavelength PCA models focus on the NIR region of the spectrum. Interestingly, while the DS dataset demonstrated greater values for explained X variance, for most PCs, compared with the DU dataset for the full-region models, the opposite is the case for the NIR-region models. This relationship for the NIR models is surprising and the opposite of what was seen for Miscanthus and bagasse samples. In general, however, a greater amount of X-variance was explained by each model for a given number of PCs compared to the Miscanthus/bagasse models with the same number of PCs.

The distribution of samples along the PC1 axis in a PC1 vs. PC2 scores plot for the WU NIR-region model, Figure D-9 (c), reflects to some degree the moisture contents of these samples determined at the time of NIR analysis. It can also be seen on this plot that there are two outlying samples (25016, 25025) with high scores for both components. The most outlying sample is sample 25016, the peat sample that was not milled and had the highest moisture content. These two samples also had the highest residual X-variance and leverage values in an influence plot of a 5 PC NIR-region model. Sample 25025 also had the highest leverage in the 6 PC NIR-region DU model but not in the corresponding DS model. In contrast, sample 25016 had normal values for residual X-variance and leverage in these PC models. Therefore it suggests that the main reason for the influence of sample 25016 in the WU model is due to its high moisture content rather than its physicochemical properties.

Figure D-9 (d) presents a PC1 vs. PC2 scores plot of the NIR-region models for the DU dataset. The samples are labelled according to the Low, Medium or High humification classifications provided by BNM (see Table D-1) with the unclassified samples, as well as those samples classified as being between two of these classes, labelled as unclassified ("UNCLASS"). The Low and Medium peats are located in the centre of the plot with no segregation between them according to the axes of the first 2 PCs. Some of the High peats are also located towards the centre; however, others show high scores along PC1. There

are three clear outlying samples – 25046, 25047, 25048, located in the bottom left corner. These are the samples of group “3” discussed in Section 13.3.1. These samples also demonstrated high leverage values in the models for the first several PCs.

13.3.3 Quantitative Calibrations

13.3.3.1 Chemical Constituents

Quantitative NIRS calibrations were developed for the DS, DU, and WU datasets for the following constituents: Glucose (i.e. GLU_SRS), Xylose, Galactose, Arabinose, Xylose, Rhamnose, Mannose, Total Sugars (total of the previous six sugars), Ash, EXTR_PD, AIA, AIR, KL, ASL, and Uronic Acids (UA). The results of these calibrations are provided, on a whole mass basis and using the SRS corrected data for the sugar values, in Table D-11 to Table D-15.

In contrast to the work on bagasse samples, it was considered that there were a sufficient number of samples to allow a validation set comprising samples not involved in the calibration. However, 53 samples is not a large enough number to trust that random selection of the samples for validation/calibration would be representative. Such a fair validation would involve samples that covered much of the range in concentration values seen for the samples in the calibration set.

On the basis that 25% of samples would comprise the validation set, with the remainder used for calibration and fitting (through cross validation), 13 samples were to be selected for validation for most constituents (with only 12 for EXTR_PD and 8 for UA since fewer results were obtained for these constituents). For the selection of validation samples for the DS and DU datasets, the reference data for each constituent was used as an input vector for K-medians clustering, using Euclidean distance, in order to assign each sample to one of 13 clusters. The samples with the maximum and minimum values for each constituent were excluded from this clustering, as were samples 25029 and 25031 since these had no DU scans. From each cluster one sample was randomly selected for inclusion in the validation set. For the selection of samples for validation in the WU calibrations, a two column matrix comprising the constituent data and moisture content was used for the clustering, with both vectors mean normalised. In this case samples 25029 and 25031 were included in the clustering but the samples with maximum

and minimum moisture contents were excluded, along with the samples with the maximum and minimum constituent values. Therefore it could be said that the validation samples were testing calibrations involving certain ranges in constituent values and moisture contents.

As with the bagasse calibrations, Haaland's criterion was used for selecting the number of PLS factors used in the model. Table D-11 to Table D-15 present the results for the calibration set (including cross-validation statistics) and validation set, as well as results when all of the samples were included in the calibration set. For many of the calibrations for lignocellulosic components, results are presented for models that included all the samples in either the calibration or validation sets or models that excluded sample 25038. When this sample was present it was, through random selection of the clusters, in the calibration set for all models except for the DS/DU mannose models where the sample was in the validation set. It was noted in many of these all-samples models that this sample was a reference- y vs. predicted y outlier, as demonstrated in Figure D-10 (a). This was confirmed from the t-test for the residual values for this sample which, for the DS calibrations, were 4.36, 2.65, 3.14, 3.93, 3.9 for glucose, xylose, rhamnose, galactose, and KL, respectively. In these calibrations no other samples consistently breached the t-test for the residual limit of 2.5. Interestingly, in the DU calibrations this sample was far less of an outlier (e.g. Figure D-10 (b)) but it was also outlying in many WU models (e.g. Figure D-10 (c)), although to a lesser degree than for the DS models.

In some cases the RMSEP increased following the removal of sample 25038, despite the RMSECV values falling for the calibration set and all-samples model. When this occurred it was typically a result of Haaland's criterion selecting an increased number of PLS factors for the model as a result of a more well-balanced calibration set. In the instance where 25038 was in the validation set, its exclusion made little difference to the validation statistics. This is unsurprising given that the sample was not an outlier for this constituent (mannose).

Table D-11 to Table D-15 show that, with the exception of the ash and moisture models, all of the models involve a Savitzky-Golay (SG) pretreatment using either the first or second derivative. Many other pretreatments were tested, but these SG transformations provided the superior RMSECVs. Interestingly, while SNVDT provided many of the best models for bagasse samples, this was not the case for the peat samples. It can be seen that for some constituents (e.g. xylose) the same pretreatment method and conditions were employed for each of the three datasets, whereas for others (e.g. arabinose) different pretreatments were used. In general, however, the predictive ability for most constituents was similar for models using SG1,1,10,10 or SG2,2,25,25 (the two most common

pretreatments used for these samples). For ash, the models using the SNV transform over the region 1100-2500 nm are provided in Table D-11 since these provided the lowest RMSECVs; however, the predictive ability using SG derivatives were close.

Regarding the wavelengths for calibration, no model incorporates the 400-1100 nm wavelength region. This was not found to provide any added value to the predictive ability of the models, and just brought forward more X-variance which the PLS models would need to account for. Calibrations involving the 1100-1800 nm region were tested for the WU models but these were not as accurate in cross-validation predictions as the models presented here, although their predictive abilities were reasonable.

Glucose and Total Sugars

For glucose the RMSEP ranged from 0.66% for the DS models (all samples included) to 1.15% for the WU model (excluding sample 25038). This corresponded to RER (RPD) values ranging from 19.1 (5.7) to 10.1 (3.0). Similar trends existed for the total sugars models. The RER_{pred} and RER_{CV} were not less than 10 for any of the glucose or total sugar models. That means that these calibrations are suitable for quality control (see Section 6.11.2). Furthermore, RER_{pred} values of over 15, indicating a calibration suitable for quantification, were possible for glucose and total sugars for the DS models, and RER_{CV} values of over 15 were possible for both constituents for the DU models.

For the DS model the first factor explained the majority of the Y-variance (86% in the calibration set comprising all samples) but less of the X-variance (18%). In contrast, in the DU model the first factor explained only 49% of the Y-variance and 45% of the X-variance, while for the WU model only 11% of the Y-variance, but 89% of the X-variance, was explained by the first factor. These trends between the different dataset models, in terms of the relative proportions of X- and y- variances explained, were consistent throughout models for the other constituents. These differences meant that comparison of X-loadings plots between different datasets was difficult, as illustrated in Figure D-15 (a) which presents the Factor 1 X-loadings plots for the different glucose models (all based on a SG2,2,25,25 transform prior to regression). However, regression coefficients plots determined from models using the optimum number of PLS factors, Figure D-15 (b), tended to be similar, particularly for the DU and WU models.

There was no discernible structure to the y-residuals with respect to predicted y for any of the datasets/models. In the DS models using the number of factors shown in Table D-11, samples 25046, 25047, and 25048 (those included in group “3” in Section 13.3.1) had Hotelling T^2 values greater than

the limit. For the DS 53 samples model, these samples were principally responsible for factor 3 where they demonstrated large positive scores (Figure D-10 (d)) whilst all other samples occupied a relatively narrow range along this factor. In contrast, for the DS calibration, sample 25038 demonstrated low Hotelling T^2 and Q-residual values, but a high residual Y variance. This indicates that the sample was a “reference y vs. predicted y” outlier and not a spectral outlier.

Similar trends were noticed regarding the relative influence of the samples for the DU calibrations; however, for the WU models the samples that exerted the most influence on the various models tended to be those with the larger moisture contents. For instance, samples 25016 (the non-milled peat), 25025, and 25034 in Figure D-11 (a) demonstrate large hotelling T^2 scores and, in some cases, large Q-residual scores, for the WU glucose model where all samples were in the calibration set. In this plot, samples of group “3” have leverage and Q-residual values much closer to the average.

Xylose Content

For this constituent excellent RER values were obtained for all datasets, with values close to 20 for the WU models. This is primarily a result of the significantly expanded concentration range that samples from group “3” provide to the models, as shown in Figure D-11 (b). For a DS calibration that excluded these samples the RMSECV was similar, 0.218%, but the RER_{CV} fell to 8.911. Clearly the distribution of samples along the concentration range for this constituent is far from ideal and for an improved calibration more samples should be sought to fill in the gaps. Sample 25038 was a predicted y vs. reference y outlier for this constituent and its removal tended to improve the models.

Arabinose Content

Interestingly, for this constituent superior RMSEP values were obtained for the DU models compared with the DS models, despite the same samples being in the validation sets for each of these models. However, when all samples were included in the calibration set, the RMSECV of the DS models were superior to the other datasets. Halaland’s criterion selected differing numbers of factors for the various models and this may be a reason for the unexpected differences between the datasets. The extra factor used in the calibration model for the DS model may have resulted in an overfitting that reduced predictive accuracy on validation. With all samples in the calibration set, 12 factors were chosen by the criterion for the DS dataset. This was the limit that the Author specified to the software. The “creeping up” of the explained variance, in cross validation, is shown for this DS calibration in Figure D-11 (c).

Figure D-11 (d) presents the reference y vs. predicted y plot for the DU model. The three datapoints with the highest arabinose concentrations are from group “3” and, as was the case with the xylose concentrations, there is a large gap between these and the majority of the other samples. In this case, however, there is one sample in between the two groups. According to the clustering method outlined for validation sample selection, this was the only sample in a cluster covering this concentration range, and so it was included in the validation set. Ideally a sample within this range should also be included in the calibration set so that this validation sample could be well modelled.

Galactose Content

This constituent demonstrated some of the greatest differences in RMSEP/RMSECV values between the various datasets. Results for the DS dataset were excellent with RER_{pred} and RER_{CV} values over 15. Indeed, the RER_{pred} was 20.9 and the RPD_{pred} 6.2 when sample 25038 was excluded from the calibration set; see Figure D-12 (a) for the predicted y vs. reference y plot for this model. As with the arabinose models, this constituent required a relatively large number of factors, particularly when all samples were included in the calibration set. Also, as with the arabinose models, the WU model for galactose, based on all samples in the calibration set, had fewer factors than the DS model. Since WU models typically require more factors than DS models to model properties to the same level of accuracy (as a result of the need to model the large added X-variance provided by the increased moisture content), it appears that there is a level of spectral detail, needed for the accurate DS galactose calibrations, that the WU spectra cannot provide to the same extent.

Rhamnose and Mannose Content

These are minor constituents in bagasse samples but present in larger proportions, with more variation, in the peat samples. This means that more accurate calibrations are possible, as demonstrated in Table D-12 and Table D-14. For the mannose DS and DU calibrations, the clustering algorithm for validation sample selection chose sample 25038, a reference y vs. predicted y outlier in the DS but not DU calibrations involving all samples. However, even when this sample was removed from the validation set the RMSEP for the DS model was greater than for the DU and WU models. The WU models required more factors than the DS/DU models, but provided improved RMSEPs. When all samples (excluding 25038) were included in a calibration, the RMSECVs were more consistent between models, however.

In contrast, for the rhamnose calibration the WU dataset demonstrated higher RMSEPs and RMSECVs, particularly in validation where the RER was less than 10. Figure D-12 (b) presents a predicted y vs. reference y for the mannose WU model, and Figure D-12 (c) presents the same plot for the rhamnose DS model, both plots include sample 25038.

Klason Lignin (KL), AIA, AIR, Ash and ASL Content

Klason lignin was the largest single constituent in all samples; the AIR values were higher, but these represent a combination of the AIA and KL values. For the WU and DS datasets sample 25038 was a clear outlier, as shown in Figure D-12 (d) for the DS model, and its removal helped to improve the predictive ability of most models, particularly so for AIR. The removal of this sample did not significantly change the AIA or Ash models, however (only the full sample-set models are provided for these constituents).

The RER_{pred} values for AIA, KL, and AIR are lower than for many other constituents, with values over 10 only obtained for the KL and AIR models for the DS dataset. However, the RER_{CV} values, where all samples are in the calibration set, are much higher, in all cases, and above 10 in all instances except for the WU KL model including sample 25038.

Regarding these three constituents, the lowest RER and R^2 values were obtained for the AIA. Figure D-13 (a) provides a reference y vs. predicted y plot for this constituent, and Figure D-13 (b) provides the same plot for DS-Ash. Clearly AIA is closely related to the total ash content and in both plots, as with the xylose plot, samples from group “3” provided an increase in the concentration range, helping to increase the RER and RPD values. These samples occupied a unique region in score space in F1 vs. F2 scores plots for AIA, AIR, KL, ash, and ASL. For ash, when these samples were excluded from the DS model, the RMSECV (for a model built on the remaining 48 samples) was reduced to 1.36%, but the RER_{CV} fell to 9.53. When these samples were excluded from an AIA DS model comprising all other samples the RMSECV was similar at 1.30% but the R^2_{CV} fell to 0.351. This particularly low value was a result of samples 25049, 25026, and 25042 which all had large residuals in cross validation. When these were removed the RMSECV fell to 0.88% and the R^2_{CV} increased to 0.682.

RER_{pred} values for KL were improved over the corresponding values for AIA, but lower than for many of the models based on the sugars. The difficulty in modelling the KL content may be because Klason lignin is not an accurate term for what is actually present in the peat. In contrast to other constituents, such as glucose and xylose, which are analysed by direct reference methods, KL is determined simply through

gravimetric methods after acid digestion. The degradation of plant matter over thousands of years will produce a whole spectrum of degradation products that may be insoluble to acid but highly different from the lignin observed in virgin plant material. The type and relative proportions of these products are likely to vary according to the degree of humification and the type of peat bog. Hence, a KL calibration may require the PLS model to account for numerous compounds with varying spectral contributions. Such a model will clearly be complex and require many more samples in order to cover the range in degradation products that may be present in peat. The improvement in RMSECV values for calibrations developed over all samples compared with the RMSEP values for the samples in the validation set suggests that the additional samples benefit the model; hence more samples may improve a future model validated on unknown samples. Nevertheless, the models as they stand have value in rough estimates of KL content which should aid in the screening of samples according to low, medium or high levels of humification, Figure D-1 (b). With an RMSEP of 3.2% for the WU model this classification should even be possible for the WU samples.

The ASL calibrations provided interesting results with lower RMSEPs for the DU and WU models compared with the DS model. Figure D-13 (d) presents a predicted y vs. reference y plot for the WU model. When all samples were included in the calibration set the RMSECVs were very similar across all datasets, with or without the removal of sample 25038. Furthermore, the DU model required only 3 factors compared with 8 for the DS model – the DS model at 3 factors had a RMSECV of 0.268%, 25% higher (in relative terms) than the DU RMSECV. These results for ASL are a significant improvement over those for the bagasse ASL models and over many of the publications available in the literature (see Appendix B). However, as with the KL content, what is being measured in the reference methods as ASL in these peat samples may not necessarily be true ASL. Loadings and regression coefficients plots were observed for this constituent in the various models but no specific wavelengths associated with ASL components were found in the literature to allow comparisons to be made. The F1 loadings plot was certainly quite different from that of KL; see Figure D-16 (a).

Extractives Content

These calibrations excluded samples 25024, 25025, and 25028 for which reliable EXTR_PD data were not available. The lowest RER values of all the constituents were obtained for the extractives content with a particularly low value of 6.3 for the WU validation set (R_{pred}^2 was 0.65). This relatively poor performance

of the WU calibration compared with the other models is consistent with what has been seen for the Miscanthus and bagasse samples.

The overall poor performance of the calibrations must be put in context with reference to the high SEL of 0.37%. It was noted that the peat samples had a relatively high (over 10%) equilibrium moisture content upon exposure to the air after oven drying. The samples also absorb moisture quickly upon removal from the oven. While good dessicator practice was employed to reduce the effect of the absorption of airborne water molecules, there will always be some variable effects as a result of this absorption and these will be magnified for those samples that absorb moisture most rapidly. The fact that the extractives content was determined based on mass loss, with moisture correction, is a basis for analytical, and hence calibration, inaccuracies.

Also, as with the KL content, those components that are determined (in total) gravimetrically will vary according to the types of peats and their relative degrees of humification. It is possible that there were an insufficient number of samples to model this variation adequately with PLS models. However, the relatively poor performance of these extractives calibrations should be put in context. The extractives are primarily removed/quantified in this study so that they do not interfere with the analytical hydrolysis methods. The quantification of extractives was adequate enough to correct lignocellulosic data from an extractives-free to a whole mass basis and still enable good calibrations to be developed for these components. Figure D-14 (a) presents a predicted y vs. reference y plot for the DU model.

Uronic Acid Content

Only a total of 34 samples were analysed for their uronic acid content. Despite this the R_{pred}^2 values for each dataset are reasonable, particularly given that uronic acids are hard to quantify accurately and reproducibly using the photometric method used in this study (see Section 3.2.4), and given that some previous NIR calibrations for these have been poor (see Appendix B). In all models sample 25053 was excluded since this was a clear outlier in reference y vs. predicted y plots; see Figure D-14 (b) for a plot for the DU model.

Comparisons Between Loading Plots

Given that the DS models explained the most Y-variance with the first factor, loading plots for selected constituents will be compared for this dataset (Figure D-16). The regression coefficient plots are also presented in Figure D-17. The spectra were transformed by SG2,2,25,25 prior to the PLS1 regression for

each constituent. Figure D-16 (a) shows that the AIR and KL F1 loadings plots are similar. The area around 1662 nm is of importance, and likely to represent the 1st overtone of a C-H stretch in the aromatic rings of lignin (Shenk et al., 2008). There is also a trough in the loadings plot around 1400 nm, associated with the first overtone of an O-H stretch in lignin (Shenk et al., 2008). There are large absolute loading values around 1900 nm, another region where lignin characteristically absorbs (Üner et al., 2011). This wavelength region demonstrated by far the largest absolute regression coefficient values for the DS (and DU) KL and AIR models; however, in the WU KL model the absorbance at 1662 nm had a greater regression coefficient value. This is to be expected since the large contribution of moisture to the absorbance around 1900 nm would mask much of the signal from lignin. The ASL loading plot differs significantly from the AIR/KL plots, but bears some resemblance to the glucose loading plot.

Regarding glucose, there are several key areas in the loading and regression coefficients plot. The areas around 1450 nm have been linked to the 1st overtone of O-H stretches in crystalline and semi-crystalline cellulose (Tsuchikawa and Siesler, 2003b, Tsuchikawa and Siesler, 2003a), as well as in amorphous polysaccharides (Mitsui et al., 2008). The largest absolute regression coefficient values occur around 2100 nm and 2326nm, which are characteristic absorption region for cellulose (Üner et al., 2011). The F1 loadings plots for mannose, rhamnose, and arabinose share similarities to that of glucose, at shorter wavelengths, but are more similar to each other. In contrast, the xylose and galactose loading plots have marked differences from the other sugars in some wavelength regions.

13.3.3.2 Moisture Content

Moisture Content of WU Samples

Figure D-18 presents regression statistics for the best model found for the (wet basis) moisture content (MC) of the WU peat samples. Unlike the models for the other constituents, SNVDT was the ideal spectral pretreatment here. The RMSEP (1.37%) is substantially less than the RMSEC (2.02%), and less than the RMSECV for a calibration involving all the samples (2.09%), and even less than the SEL (1.62%). The samples chosen for the validation set were based on the clustering algorithm separating the samples according to moisture contents. It appears that the random selection of the samples from these clusters provided a close to ideal set of samples for prediction. Nevertheless, even the RMSECV of 2.1% provides an RER of over 23 and an RPD of over 5, which are good.

Moisture Contents for Reference Analytical Methods

As with the bagasse samples (see Section 12.3.3.2) NIR calibrations were developed for the moisture content prior to: the removal of ethanol extractives (DS-E); the determinations of the moisture contents of the extracted samples (DS-E dishes); and the analytical hydrolysis method (E-H).

Table D-16 provides summary statistics for the moisture contents at these various stages. It can be seen that the peat samples have higher equilibrium moisture contents than most bagasse and Miscanthus samples. Regression statistics for the various calibrations are provided in Table D-17. The RMSEPs are generally higher than those for Miscanthus and bagasse calibrations, presumably due to the rapid absorption of moisture by peat samples as discussed in Section 13.3.3.1.

After evaluating the models with independent validation samples, these samples were included in the calibration set and new models developed using the number of PLS factors presented in Table D-17. These models were then used to predict the moisture contents of the samples that comprised the calibration/validation sets of the other models, as described in Section 12.3.3.2. Table D-18 lists the results of these regressions and Figure D-19 plots them. The relationships here are the opposite of those for the bagasse samples. While the bagasse DS-E model overestimated the moisture contents of the DS-E Dishes and E-H samples, the peat DS-E model underestimates them. This may be because not all of the residual ethanol had been lost in the air drying, meaning that the DS-E models, not calibrated for ethanol content, would not be able to reflect this. Also, the E-H samples are poorly predicted by the DS-E Dishes samples. The largest residuals in these predictions were seen for E-H samples of lower moisture contents. Such a poor prediction would be expected since the DS-E Dishes model would need to extrapolate in prediction to lower moisture contents than that of the samples in its calibration set.

13.3.4 Differences Between Replicate Scans

The models, based on all the samples except 25038 inside the calibration set, for KL, GLU_SRS, and RHA_SRS were applied to the spectral datasets comprising the replicate scans in order to determine any differences in predictions that may occur upon rescanning. This was done for the WU, DU, and DS scans and models. The results are provided in Table D-19 where “Bias” represents the average value of the predicted value for the first scan minus the predicted value for the second scan. “Av. Abs. Diff” is the average value taken of the absolute difference between the two scans. SD is the standard deviation of

the absolute differences, and Max and Min the maximum/minimum values of it. For each sample the absolute value of the difference was expressed as a percentage of the average predicted value for the two scans. The average and maximum and minimum of these percentages are also provided in Table D-19. In the worst case, the difference between scan 1 and 2 for the prediction of the rhamnose content of one DU sample was 44.6% of the average value of both predictions; a significant deviation.

It would be expected that the average absolute difference between replicate scans would fall moving from the WU to DU and DS samples since sample homogeneity, in terms of presentation to the window of the NIR cell, should increase. The average difference is less for the DS scans for glucose but, interestingly, it is greater for the minor constituent rhamnose. Regarding KL, while the average difference between replicate scans for the DS samples is less than the difference for the WU scans, it is similar to the DU average difference.

Table D-19 shows that the relative importance of the differences between the predicted compositional values for replicate scans decreases as the magnitude of the constituent increases. Hence, the Av (%) and Max (%) figures are lowest for KL, which is present in the greatest quantities, and highest for the minor constituent rhamnose. The magnitude of the absolute difference between the duplicate scans does not appear to be related to the magnitude of the predicted value, however, as shown by the scatter plot of difference vs. predicted glucose content for the WU dataset in Figure D-20.

Figure D-21 presents histograms for the absolute differences for predicted values between replicates for the glucose and KL models of the WU, DU, and DS datasets. These show that there are occasional outlying samples where there are large differences between duplicate scans. For example there is one sample in the DS glucose model that has a difference of approximately 2.5%, whereas the modal group would be a 0-0.25% difference. It is also of interest that the structure, for any particular dataset, of the glucose histogram is not the same as the KL histogram despite the same raw spectra being used.

These results reaffirm the importance for replicate scans in order to lessen any effects that variations in sample presentation to the NIR cell may have on the spectra. If the models are to be used to predict unknown samples in the future it will be important that more than one scan is taken and the differences between these scans noted in order to check for incorrect sample presentation methods. The removal of much of the moisture and reduction of mean particle size does not eliminate the need for this.

13.4 Summary

Analytical data and quantitative NIRS models for various lignocellulosic parameters, moisture, and ash have been presented based on a set of 53 peat samples. This set comprised of samples with high, medium, and low degrees of humification/degradation as well as three samples that differed substantially from representative samples of each of these classes. The computation of correlation coefficients between the various constituents indicates that there are strong links in many cases and reference analytical data may be used for prediction of the other properties of samples (Section 13.3.1). In contrast to the bagasse and Miscanthus samples, rhamnose and mannose are present in quantities, and with variances, sufficiently large to allow for the development of accurate NIRS models.

Regarding these models, Table 13-3 provides a summary of the RER values for the best model for each dataset, each constituent, and whether the RER_{CV} or RER_{pred} is used. The values are classified as to whether they are greater than 15 (a model suitable for quantitative prediction, "A"), between 10 and 15 (suitable for classification, "B"), or between 4 and 10 (suitable for screening, "C"). Table 13-3 shows that there are no models that have RER values less than 4 and for most constituents an RER over 15 is possible. This may not always be possible from the WU samples (those that involve the least sample preparation); however, the WU models do have value in classification purposes. Hence they can be used to see if further processing of the sample to the DS state, for a more accurate quantitative prediction, is warranted.

For most of the models there are relatively large differences between the RER_{CV} , based on models with all samples in the calibration set, and the RER_{pred} , based on the samples in the validation set, with the former being greater and, correspondingly, the RMSEPs are typically greater than the RMSECVs. This indicates that the samples included in the validation set brought additional useful spectral and constituent data when they were included in the global calibration set. This, coupled with the wide ranges of compositional values noted for the samples, suggests that peat is a complex feedstock to model for NIRS calibrations and that improved models could result from the inclusion of additional samples in calibrations. Clearly, however, inclusions should be targeted towards samples that would provide additional useful information. Hence, potential future samples should be tested with existing models, and their X-residuals and predicted constituent values (and deviation in prediction) examined before deciding whether their subsequent processing and reference analysis is warranted.

Table 13-3: Classification of the RER values for the best model for each constituent and dataset for the peat samples. The RER_{pred} , based on the samples in the validation set, and RER_{CV} , based on models where all the samples are in the calibration set, are presented. A: $RER > 15$; B: $10 < RER < 15$; C: $4 < RER < 10$

Constituent	DS		DU		WU	
	RER_{pred}	RER_{CV}	RER_{pred}	RER_{CV}	RER_{pred}	RER_{CV}
Glucose	A	A	B	A	B	B
Xylose	B	A	B	A	A	A
Rhamnose	B	B	B	B	C	B
Mannose	C	A	B	B	B	B
Arabinose	C	B	B	B	B	B
Galactose	A	A	B	B	C	B
Total Sugars	A	A	B	A	B	B
KL	B	A	C	B	C	B
ASL	C	B	B	B	B	B
AIR	B	A	C	A	C	B
AIA	C	C	C	C	C	C
Ash	B	B	C	B	C	B
EXTR_PD	C	C	C	C	C	C
UA	C	C	C	C	C	C
Moisture	-	-	-	-	A	A

It is possible that such a targeted analysis of future samples may allow the RER values for the WU models to rise to the A or B classes for many constituents. Nevertheless, the models developed on the 53 samples used in this study are good and enable a wider range of constituents to be predicted, and with a greater degree of accuracy, than any other publications that the Author found in the literature. On the suitability of peat as a feedstock for biorefining, however, the Author considers that the total carbohydrate contents (which only reach a maximum of 28%) are too low to allow for the processing of peat in hydrolysis-based technologies. Gasification and pyrolysis technologies could be alternatives; however, the large moisture contents of peat may be prohibitive in these cases. While the NIR calibrations developed in this chapter are therefore unlikely to have potential applications in online biorefining systems, they can have value in chemical studies on the fate of lignocellulosic constituents in various peat types and will allow for the more rapid and productive analysis of samples than normal reference methods would allow.

14 Qualitative Analysis of Miscanthus Samples

This Chapter will outline the use of NIRS to discriminate between the numerous Miscanthus samples that have been collected as part of this study. A literature review of Miscanthus is provided in Chapter 10 and of qualitative analysis methods in Chapter 7. Many of the Tables and Figures for this Chapter are presented in Appendix E.

14.1 Methodology

14.1.1 Collection of Miscanthus Samples

A €120k research project, written by the Author and funded by the Department of Agriculture, commenced in December 2007 and involved the collection and analysis of a variety of agricultural lignocellulosic feedstocks that have potential for use in biorefining technologies. A major focus was on the collection of Miscanthus samples. Plants were sampled from stands for each month during the period of October 2007 to April 2008 (or until the stand had been harvested). Seven different stands were used. These were chosen to reflect variations in stand age and the success in plant establishment: These sites are listed below:

- Adare-H: A plantation that was in its third year at a farm of Joe Hogan in Adare, Co. Limerick.
- Adare-C: A plantation in its second year at a farm of Joe Hogan in Adare, Co. Limerick.
- Shannagolden: A plantation in its second year at a farm in Shanagolden, Co. Limerick.
- Langton: A first-year crop at a farm of Paul Langton in Co. Kilkenny.
- Clonmel: A first-year crop at a farm of Ann Kehoe in Co. Tipperary.
- Carlow-F: A crop that was in its thirteenth year at the Teagasc Oak Park Research Centre in Carlow, Co. Carlow.
- Carlow-G: A crop that was in its thirteenth year (different site) at the Teagasc Oak Park Research Centre in Carlow, Co. Carlow.

The dates at which samples were collected from each site are presented in Figure 14-1. If more than one plant was sampled then this is represented by stacked points on the chart.

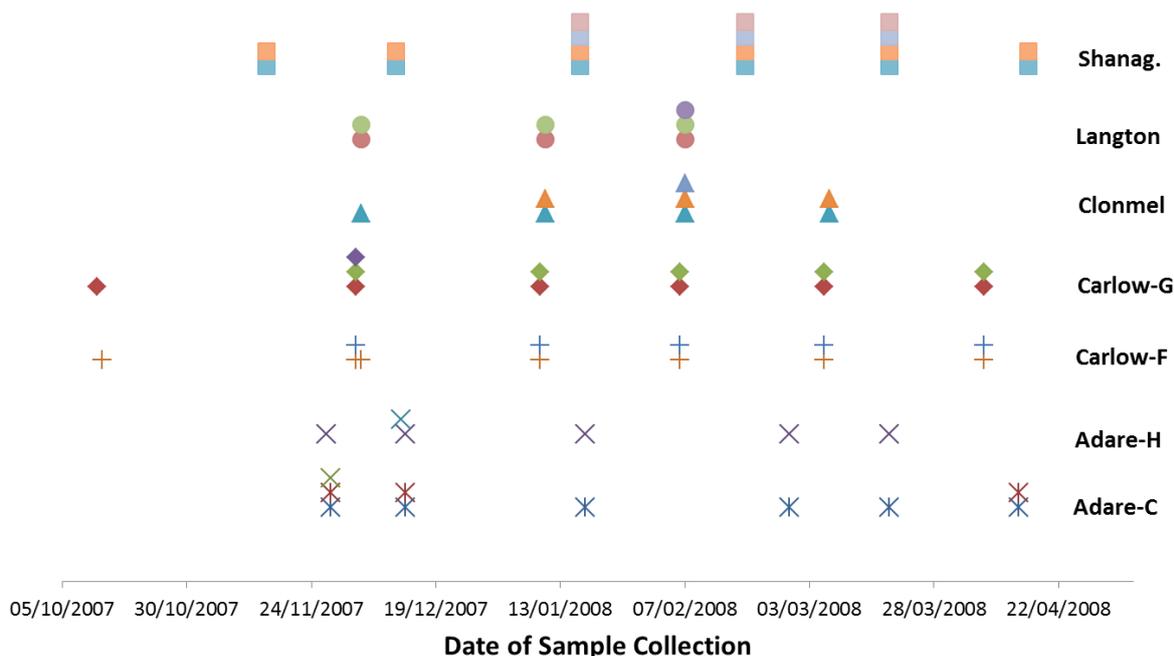


Figure 14-1: Chart representing the number of *Miscanthus* plants collected, according to date and site location, during the period October 2007 to April 2008, as part of the Department of Agriculture project.

The plant sampling methodology involved the random selection, at a location within the field, of a whole plant. All of the plantations listed above featured the *Miscanthus x giganteus* variety and a plantation density of 1 plant per metre squared. It was easy to determine, according to the distribution of the stems, the boundaries of each plant. Separate fractions of the plant were collected. The general methodology involved is outlined below:

1. The “live” leaf blades (“K” fraction) were considered to be those blades that were more than 60% green (determined upon visual inspection). These were separated from the sheath at the ligule and placed in airtight plastic bags.
2. The “live” leaf sheaths (“M” fraction) were the sheaths attached to the live leaf blades and they were removed from the plant and collected in separate bags.
3. The “dead” leaf blades (“F” fraction) were those that were judged, upon visual inspection, to be less than 60% green. These were removed and collected separately.
4. The “dead” leaf sheaths (“H” fraction) were then also collected.
5. In some instances there were inflorescences at the top of the plant. These “flowers” (the “FL” fraction) were collected separately.

6. At this stage the remainder of the plant that had not been sampled consisted of the stems. These were cut at a height of 5 cm above the ground. Then each stem was cut at distances of 1 metre. The first 1 metre section was labelled X1, the second X2, etc. In order to fit the stem sections into the airtight bags it was necessary to cut them into billets of approximately 15 cm in length. Care was taken so that the stems were only cut at internode, rather than node, sections.
7. The samples were then brought to the Carbolea laboratories and stored in a freezer.
8. At a later point the stem sections were removed from the freezer and allowed to equilibrate to room temperature (the bags were still sealed to prevent moisture loss). The node sections were then cut from these stems resulting in the formation of two separate samples for each stem section. For example, the X1 fraction would provide an X1N sample corresponding to the nodes collected from that sample and an X1T sample corresponding to the internodes sections collected from that sample.

A total of 77 *Miscanthus x giganteus* plants were collected from these seven sites between October 2007 and April 2008. The separation of these plants into the anatomical fractions described above resulted in the production of 479 samples.

On 9/2/10 a total of 8 WP (whole plant) samples were collected from the Adare-H site and 8 WP samples from the Adare-C site. These samples involved either the collection of a set of whole biomass stems (including the leaf blades, leaf sheaths and flowers) or the collection of a certain metre fraction of these plants (WP1 = first metre, WP2 = second metre etc.). A further 13 WP samples were collected from the Carlow-G site on 17/2/10; three of these WP samples were from a stand of the *Miscanthus x sinensis* variety. On 29/9/10 a further 8 WP samples were collected from the Adare-H site and 7 WP samples from the Adare-C site. The Adare-C site was returned to on 27/10/10 for the collection of a further 4 WP samples and the Adare-H site was returned to on 10/11/10 for the collection of a further 9 WP samples. On 6/10/10 seven WP samples were collected from the Shanagolden site with three more WP samples collected from this site on 27/10/10.

On 15/10/09 samples of several *Miscanthus* varieties other than *giganteus* were collected from an experimental plot in the Teagasc Oak Park Research Facility. These samples comprised two *sinensis* plants and 7 other varieties. The F, H, K, and M fractions of these plants were collected separately, as described above. However, the stem sections were not cut to separate the internode and node sections – instead the X1, X2 etc. stem sections were retained for analysis. On 9/2/10 a further 13

stem sections, this time of the *Miscanthus x giganteus* variety, were collected from the Adare-H (7 samples) and the Adare-C (6 samples) sites.

In November 2010, it was decided that more leaf samples were needed so a total of 8 F samples, 6 H samples, and 12 K samples were collected from the Adare-H and Adare-C sites.

In addition to the samples that were collected by the Author, 16 samples of harvested *Miscanthus* cultivars other than *M. x giganteus* were sent to the Carbolea laboratories from Dr. Eppel-Hotz in Germany (see Section 16.2). Since these were received dry, no wet NIR analysis (i.e. WU scans) was possible for these samples. These samples were labelled "HP" to differentiate them from the other plant fractions that were collected.

Table E-1 provides a summary of the samples collected/received as part of this study.

14.1.2 Processing of Samples

The general methodology outlined in Section 11 for the processing of biomass samples was carried out for selected *Miscanthus* samples. Most of the samples listed in Table E-1 were scanned in the WU state after comminution, usually using the Retsch SM2000 chipper and a 2 cm sieve aperture, to a suitable particle size that would allow presentation to the coarse sample cell of the NIR unit. The average was taken of the three scans collected per sample. In total, see Table E-2, 628 average spectra were collected.

Selected samples were then put through the rest of the sample preparation protocol for the production of DS and DF samples. This resulted in the collection of average DU, DV, DG, DH, DS, DT and DF spectra for each of these samples. The number of samples for which each of these average spectra were collected are summarised, according to plant fraction and *Miscanthus* variety, in Table E-2 to Table E-10.

The samples that did not go through this protocol were returned to the freezer after the collection of their wet spectra. In order to differentiate between the wet spectra of unprocessed samples and the wet spectra of (subsequently) processed samples the former were labelled WC and the latter WU. For WC samples the only means of predicting their chemical composition would be based on calibrations developed on WU spectra.

Figure 14-2 (a) provides a photograph of a WU stem section and Figure 14-2 (b) provides a photograph of an internode section after it had passed through the chipper and just prior to the collection of its WU spectra. It can be seen that the sample is somewhat “grassy” in nature with some fine strands produced after chipping. However there are also some larger particle sizes and semi-intact stem sections preserved. Random hand selection was used for sample presentation to the NIR cell meaning that both the “grassy” parts and the intact stem fractions could be presented to the cell window. Figure 14-2 (c) is a photograph of another WU sample prior to the collection of its spectra. This sample had a lower moisture content than the sample in Figure 14-2 (b) which meant that the sample was less likely to stick to the walls of the chipper and more likely to pass quickly through it and so the degree of comminution was less. This can be seen in the greater proportion of intact stem fragments, compared to “grassy” material, in this sample. Figure 14-2 (d) presents a WU sample from a later harvest (April 2008). This sample had a lower moisture content than the sample in Figure 14-2 (c) and this is reflected in an even greater abundance of stem fragments and much less of the “grassy” material. All WU samples, irrespective of their moisture content and subsequent particle size distribution, were retained within the same dataset.



Figure 14-2: Photographs of Miscanthus stems. (a) Billets; (b) an internode WU sample after processing through the chipper; (c) same as (b) but the sample had a lower moisture content; (d) same as (c) but from a lower moisture content sample and harvested later in the harvest-window.

Figure 14-3 (a) provides a photograph of a WU node sample after comminution. It is clearly physically different from the internode samples in Figure 14-2. Since the nodes were cut separate from the stems they existed in a much smaller form than the internodes when presented to the chipping device. The same 20 mm sieve aperture was used in the chipper for the nodes. That meant that the comminution involved was minimal as the samples quickly passed through the unit. The samples were then presented to the NIR cell, but care was taken so that there were no voids in the vicinity of the cell window. Figure 14-3 (b) shows a sample of a Miscanthus live leaf blade sample after comminution and Figure 14-3 (c) shows a sample of a dead sheath sample after comminution. Again there are clear differences between the physical forms of the two samples. The dead sheaths were very hard to process in the chipping device and tended to pass directly through. This often meant that these were still too large to present directly to the NIR cell, and so further passes through the chipper were required. Figure 14-3 (d) shows a picture of a WP sample that had been processed through the chipper. This sample contained all of the fractions of the plant when it was presented to the unit and so is the most heterogeneous fraction.



Figure 14-3: Photographs of some Miscanthus fractions. (a) WU node; (b) WU live leaf blade; (c) WU dead sheath; and (d) a WU WP sample.

14.2 Comparison of Spectra

Figure E-1 presents the spectra for 566 WC/WU samples (scans of whole stem sections and WP fractions excluded). The spectra are colour-coded according to the plant fraction. Figure E-2 also shows these samples, this time over the 1100-2500 nm region after a full MSC transform has been applied. As with the spectra of the wet bagasse samples (Section 12) the influence of water on these spectra are clear with the high absorbance values, particularly around 1450 nm and 1930 nm. The moisture contents of the various plant fractions tended to decrease with increasing time from the period of plant senescence. Since Figure E-1 and Figure E-2 contain the spectra of all samples collected from senescence to the end of the harvest window, this trend cannot be delineated. However, it appears that the node sections, in having some of the highest absorbance values in the NIR regions associated with absorbances by the molecular bonds in water, may have among the highest water contents. In contrast, the dead sheaths and the dead leaf fractions tend to have much lower absorbances in these regions.

Figure E-3 presents the raw WU spectra for some of the plant fractions obtained from a single plant collected from the Shanagolden site on 15/11/07. There were four internode and four node fractions obtained. However, for clarity only the spectra resulting from the first metre section are provided. Figure E-4 provides the raw WU spectra of a similarly sized plant that was collected from the same location on 19/3/08, approximately 4 months later. There are no spectra for the K and M fractions for this sample since these fractions were not present at this point.

An interesting feature evident in Figure E-1 and Figure E-3 is a peak in absorbance values around 673 nm. This peak occurs for the K, M and XT sections and to a much lesser extent for the XN section in Figure E-3. It is only present as a shoulder in the F, H, and FL sections in Figure E-3. In Figure E-4 there is no peak in this region; there is only a small shoulder in the X1T and X1N spectra, while there is no visually apparent effect for the F and H spectra. Clearly this peak is associated with chlorophyll and its lower values in the “dead” sections and at a later point in the harvest window reflects the gradual loss of this component from the standing biomass after senescence. For the early-harvest spectra this peak is consistently the greatest for the live leaf samples and offers a means for the visual identification of these samples using only the spectra. However, if longer wavelengths and

scatter correction techniques are used (e.g. Figure E-2) there are no longer any clear visual differences between the plant fractions.

Figure E-5 provides the six different average spectra obtained from a single sample (the X1T internode sample in Figure E-3) at the various stages of sample preparation. The blue line represents the WU scan and shows the increased absorbance due to water. The red line represents the DU scan and all of the lines below it represent the trend seen with regard to increased sample scattering (and hence reduced absorbance) in response to a reduction in mean particle size.

It can be seen that the DT scan scattered less than the DS scan. This is an indication that the new method used for presenting the sample to the NIR cell in the DT scans, see Section 11.1, did result in an increased average particle size at the cell window compared with the method used for the DS scans. The absorbances for the DG scans are lower, reflecting the contribution of the DF component to this fraction. The DF fraction had the lowest mean particle size and therefore scattered the most, reflected by this fraction having the lowest absorbance values in Figure E-5. Interestingly, however, it can be seen that the DG scan is closer to the DF scan than it is to the DS/DT scans. This is despite the relative proportion of the DS fraction in the DG sample being greater than that of the DF fraction. Hence, the bias towards smaller particle sizes being presented to the NIR cell window under the DG method is apparent here.

It was general practice, upon development of the DT, DH, and DV scan protocols, to use these methods instead of the corresponding DS, DG, and DU scans. At a later point all of the available DS/DT samples were rescanned with the DT/DS methods so that comparisons could be made between the two. However, rescanning DU/DV samples was not possible since all of the unground samples had subsequently been processed to the DS/DF state. Rescanning DG samples using the DH protocol (and vice versa) could be possible for samples where enough of the DS and DF fraction remain following wet chemical analysis procedures. This would be an extremely time consuming method involving the proportionate mixing of DS and DF samples to obtain the correctly weighted DG fraction which could then be scanned in either the DG or DH method. This mixed sample would then need to be sieved again in order to retrieve the DS and DF fractions.

Experiments were made with several DG samples that had not been sieved. The spectra were collected using the DG method and the DH method and comparisons were made. It was found, as with the DS/DT scans, that the mean absorbances were greater for the DH scans compared against the DG scans. A test was also made on a DU fraction that had not been ground. It was scanned in the DU and DV methods. In this case the particle size induced baseline shift was more marked between

the two methods of spectra collection. The effect was comparable to the differences between the DU and DW bagasse scans (see Section 12.3.2.2).

Figure E-6 presents the 400-2500 nm DT absorbance spectra for various plant fractions and Figure E-7 presents these spectra for the 1100-2500 nm region after the MSC transform. The chlorophyll “hump” is still present in the visible region and the removal of most of the water from the samples has allowed more detail to be seen in the spectra, particularly over the 1800-2000 nm region. However, there are few readily identifiable consistent differences between plant fractions. Discrimination techniques will therefore need to rely on chemometrics.

14.3 PCA of *Miscanthus* Spectra

Figure E-8 shows an explained X-variance plot, under cross validation, for the 400-2500 nm PCA models of the various datasets (DV and DH not included due to a limited number of scans for these sets) and Figure E-9 shows a similar plot for PCA models based on the 1100-2500 nm region. All spectra were transformed to the second derivative (SG-2,2,14,14) prior to the PCA.

As with the PCA of bagasse samples, more of the X-variance is explained in the early PCs of the WC/WU dataset (WCA) than with the other datasets – a result of the massive influence that moisture content has on variation between spectra among these samples. The first PC of the WCA dataset, Figure E-10 (a) has high loadings associated with water absorption in wavelength regions around 1450 and 1930 nm. However there is also a high loading around 673 nm, reflecting the chlorophyll “hump” seen in many samples. This region had much higher relative loading values in PC1 loading plots for the other datasets, e.g. Figure E-10 (b) for the DG dataset, since these models did not need to account for as much variation in moisture content.

The loadings plot for PC2 in the WCA model looked similar to the PC1 plot with high loading values around 673 nm and in the regions where NIR radiation is absorbed by the molecular bonds of water. However, the peaks in these moisture regions were inflected. These two first PCs explained 83% of the total X-variance for the 400-2500 nm WCA model and the scores plot for these are presented in Figure E-10 (c). It can be seen that there is some degree of segregation between leaves (blue squares) and stem section samples (red stars). The green triangles in Figure E-10 (c) represent the whole plant (WP) samples that were collected. Those WP samples located in the area of low scores for both PCs represent the samples that were collected immediately after senescence (between

29/9/10 and 10/11/10). Hence these samples were collected much earlier than most of the other samples and were significantly “greener” in colour and also wetter and contained a higher proportion of extractives than other samples. In a similar plot of PC1 vs. PC2 scores for the NIR-only model, Figure E-10 (d), these early collected samples were no longer outlying.

As with the bagasse PCAs, the NIR-only models, Figure E-9, explain more X-variance with fewer PCs. For example the WCA model has 88.8% of all X-variance explained in the first PC. Figure E-9 also shows more significant differences between the explained X-variance for the dry scans compared with the full-region models. As before, the DU scans demonstrate the least explained variance, presumably as a reflection of the heterogeneous particle sizes of this set, and the associated errors upon sample repacks. However, the explained variance of the DT dataset is greater under the early PCs than that of the DS set and similar to that of the DF set. It is possible that the improved sample presentation of the DT method (see Section 11.1) may be responsible for this in a similar way to the DW scans displaying less variance than the DU scans for the bagasse samples (see Section 12.3.2.2). The explained variance of the DG set is in between that of the DS and DT sets for the early PCs, despite there being a higher degree of particle size variation within this set.

Only a total of 9 flower samples were collected in this study, and the majority of these (6) came from *Miscanthus* varieties other than *giganteus*. These samples were significantly different, both physically and chemically (see Section 16.1) from the other plant fractions, and it is therefore unsurprising that many of these samples tended to be present as outliers in influence plots. Flower samples were often present as outliers in the development of qualitative and quantitative models, and the removal of these samples tended to improve the predictive accuracy of these models.

Separate PCA models were constructed, using the DS scans, for the leaf and stem fractions. The PC1 vs. PC2 scores plot for the leaf samples full-wavelength-region model is presented in Figure E-11 (a). It shows that the dead sheaths (blue circles) and dead leaf blade samples (blue squares) occupy a much smaller region of the plot than the live leaf blades (green triangles) and live leaf sheaths (pink diamonds). The limited number of flower samples (brown inverted triangles) also have a relatively large spread across this plot. Since these first two PCs account for a total of 85% of the explained X-variance (in the calibration set) it can be said that these “green” samples are responsible for most of the variation in the spectra, an observation reflected in leverage plots. The relative spread of these samples was less in PCA models focussed only on the NIR region, however (Figure E-11 (b)).

It was also observed that the samples of varieties other than *giganteus* had high leverage values. These samples also spread a wide area of the PC1 vs. PC2 scores plot, Figure E-11 (c). This is despite

there being only 67 “other varieties” leaf section samples compared with 227 samples from the *giganteus* variety. Many of the “other” varieties sampled were the only representation in that dataset of the respective plant fraction of that variety. Hence, the influence these samples (which also differ chemically, see Section 16.2) have is understandable.

With regard to the PCA model involving all of the stem samples, the other varieties had lower relative leverage values than with the leaf PCA model and less separation on a PC1 vs. PC2 scores plot. In this plot, Figure E-11 (d), there was no separation of the relative plant fractions, although the internode samples were spread more widely than the node samples. This increase in spread for the internode samples was more marked for the full wavelength model compared with the 1100-2500 nm model.

14.4 Discrimination Between Samples

Various chemometric techniques have been tested to determine if good discriminations can be made, according to the NIR spectra of the various datasets (WC/WU, DU, DG etc.), between the various *Miscanthus* plant fractions. Methods were also sought to discriminate *Miscanthus x giganteus* samples from the other varieties of the plant collected. Experiments were also conducted to determine if the samples could be discriminated on the basis of stand age, and whether the samples were collected in the “early” harvest window (taken to be between October and the end of December) or the “late” harvest window (taken to be any date from the 1st of March onwards). The samples that were collected between the 1st of January and the 28th of February were excluded from the Early/Late classification in order that there could be distinct differences between these two classes.

Two different tests were undertaken for the discrimination between plant fractions. The first test was to see if stem fractions (internode, node, and whole stem) could be discriminated from leaf fractions (live leaf blade, live sheath, dead leaf blade, dead sheath, flowers). For this test the HP and WP fractions were excluded. The second test involved trying to discriminate between the following fractions: live leaf blade, dead leaf blade, dead leaf sheath, internodes, and nodes. For this test the HP, WP, flowers, live leaf sheath, and WP fractions were excluded.

14.4.1 Clustering Methods

Initially various automated clustering methods were attempted in order to see the natural groupings of samples that occurred according to their spectra. Firstly, K-means clustering (KMC) and K-medians clustering (KDC) were compared, using a range of different distance measures, for the DF dataset with no spectral pretreatment. This dataset was selected for initial tests because the samples are the most homogeneous at this stage and segregation according to particle size or moisture effects would be unlikely.

The full spectral range was used and two clusters were chosen to see if leaves and stems could be differentiated. The results are provided in Table E-11 for the two clustering techniques - the percentages indicate the proportion of the stem or leaf samples that were classified in a cluster predominately of that fraction. It can be seen that the results for the two methods are similar for most distance measures, although the KDC results tend to be more consistent between the different distance measures. The Spearman's Rank and Kendall's Tau distance measures required a significant amount of computer processing time and were not considered suitable for further tests after KMC.

The stem/leaf classifications were examined according to the specific plant fraction, and it was observed that, for the best-performing distance measures, the problem leaf fraction in the clustering algorithm was the live leaf sheaths. For example, in the KDC city-block method, only 37.5% of the live sheath samples were classified in the same group as other leaf samples with the remainder being assigned to the stem cluster. If this same cluster analysis excluded the "M" samples then 100% of the leaves were correctly classified and 90.4% of the stems. The clustering of live sheath samples together with stem sections is understandable given that these are closely associated with the stem and have similar chemical properties (see Section 16.1).

It was decided that the KDC clustering technique and the Euclidean (EU), City-Block (CB) and Bray Curtis (BC) distance measures would be used in tests involving different wavelength regions and different spectral pretreatments. These are summarised in Table E-12. The 5 methods that resulted in the highest average percentage between the two clusters are highlighted in bold. It can be seen that using a limited wavelength region, of 400-750 nm or 400-1100 nm, for the cluster analysis provides similar results to using the full spectral region. However, avoiding the visible region and only using 1100-2500 nm region results in poorer classifications in some instances, although not when using the CB distance measure.

Derivative treatments do not improve the predictive accuracy compared with the best examples, for each distance measure, of no pretreatments. Furthermore, the 400-1100 nm region compares less well to the 400-2500 nm region when derivatives are used. Scatter correction techniques such as SNV and MSC tend to perform poorly since these are only effective when the treatments focus on the 1100-2500 nm region; the visible wavelengths are also needed to improve the cluster distributions. A clustering method using 1100-2500 nm and the MSC transform on this region puts all the stem samples in the same cluster. However, this method also puts many of the leaf samples in this cluster.

Attempts were made for the grouping of all DF samples in 7 clusters to see if these clusters would be associated with plant fraction; however, this was not the case. Samples of each plant fraction were typically spread across several clusters. Hence, it would appear that automated clustering methods are not suitable for segregating *Miscanthus* fractions to this level of detail.

Four of the better performing combinations of distance measure, pretreatment, and wavelength region from Table E-12 were used to test, for the DF dataset, the following hierarchical clustering methods:

- Hierarchical Single Linkage (HSL).
- Hierarchical Complete Linkage (HCL).
- Hierarchical Average Linkage (HAL).
- Hierarchical Median Linkage (HML).
- Ward's Method (WRD).

Table E-13 shows that the HSL method produced elongated clusters and that it is of no benefit for these classification purposes. The HCL, HAL and HML methods all generally perform similarly. The Ward's method provides superior results to the other linkage methods in two of the four cases but does not provide the greatest success rates. These come from using the Bray Curtis distance measure, the full spectral region (with no pretreatment) and the HCL/HML linkage methods. Table E-13 also shows that the linkage methods, when applied to the derivative-treated set, resulted in poor classifications.

Tests were then made for the most appropriate clustering conditions to discriminate between stems and leaves fractions for the other datasets. The classifications associated with these experiments are provided in Table E-14 which shows that, in most cases, the success rates were lower than those

experienced with the DF dataset. This is as would be expected as factors such as particle size heterogeneity, sample presentation consistency, and moisture content become more important.

Interestingly it can be seen that the datasets of the scans utilising the improved sample presentation method (DT, DH, DV) tend to have improved success rates over their counterparts where the old method was used (DS, DG, DU). In the case of the DH and DV datasets there were substantially less samples than the DG and DU datasets, respectively. However, the DS and DT sets were more comparable in size and sample distribution. It appears that the improved consistency of the sample, as presented to the NIR cell window, associated with the new method helps to reduce the importance of particle size variation in automatic clustering methods. That means that the relative importance of plant fraction increases in classification. It is interesting to note that the WC/WU dataset performs better than the DU dataset, in both the KDC and HCL methods, indicating that either moisture content is an aid in classification, or that the sample presentation method associated with WC/WU scans was superior to that of DU scans.

In summary, while the automated clustering methods provide some degree of useful classification for some datasets this only occurs on the basis of stem vs. leaf sample discrimination and it does not offer any finer level of segregation. Furthermore, a substantial number of samples are misclassified in this method (no dataset achieves a 100% success rate for both groups) and this method also requires the reclassification of all samples once new samples are introduced.

The focus in discrimination methods will now shift towards supervised procedures, where the group associations of samples in the calibration set are provided to the model. This allows the classification of unknown samples on an individual-sample or group basis and should allow more discriminatory tests, with finer levels of detail, to be evaluated.

14.4.2 Discrimination Between Early/Late Harvest Samples

14.4.2.1 PLS-DA

PLS-DA models were developed for the DF dataset using 75% of the samples in the calibration set and 25% in the validation set. Table E-15 provides summary validation statistics for some of these models. The results are presented in a matrix form for each model, with the columns representing the actual class of the samples and the rows the class predicted by the model. Two separate

matrices are presented. The first, "Score" in Table E-15, provides the results based purely on the predicted y-value obtained for each sample. This means that, in the "Early" prediction, a value greater than zero would indicate an "Early" sample with a score under zero indicating a "Late" sample. The second results block, "Dev." in Table E-15, is a 3 x 2 matrix which considers the deviation in prediction as well as the predicted y-value (see Section 6.6). If this deviation crosses zero then the sample is not classified in either group ("None" in Table E-15).

Models were developed just for *Miscanthus x giganteus* samples as well as for a global dataset comprising all *Miscanthus* varieties. It can be seen in Table E-15 that the prediction results for the models comprising all the different varieties tended to be poorer than those developed just on *giganteus*. This is particularly the case when the deviation in prediction is considered. It was found that those samples that were misclassified (or not classified) tended to be from varieties other than *giganteus*. Excluding all other varieties from a model is reasonable, given that all commercially grown *Miscanthus* in Ireland is of the *giganteus* variety.

Model *4 in Table E-15 is based only on *giganteus* samples and achieved a 100% success rate in prediction when the deviation in prediction was not considered. However, this success rate fell, particularly for the "Late" samples when the deviation was considered. The best model considering the deviation was *5 which involved no spectral pretreatments and the full spectral region in calibration.

Table E-16 presents the PLS-DA models developed for the other datasets, using only the *Miscanthus x giganteus* samples. Calibrations were not attempted for the DH or DV datasets since there were insufficient "Late" samples in these groups. A relatively large number of factors, with the optimum number determined via cross validation within the calibration set, are required to achieve the optimum model for most of these datasets. However, the final results are good, particularly when the deviation in prediction is not considered. The results for the combined WC/WU dataset are particularly promising, with only 2 misclassifications for the Early samples and none for the Late samples. The situation is not as good when the deviation in prediction is considered. However, a total of 90 samples were used in this validation set and this is a sufficient number to indicate that the simple prediction (without deviation) can be trusted for classification of unknown samples.

The discrimination between Early and Late samples in the WC/WU dataset is not simply a case of separating samples according to their moisture contents. Although moisture content does fall for the plant as a whole as time from senescence increases, many of the spectra in the dataset are from individual plant sections, and these may have significantly different moisture contents. For example,

a dead sheath fraction of a plant harvested in December may have a lower moisture content than an internode sample collected in March. In a regression coefficients plot for the 14 factor WC/WU model, it was observed that the visible region was of importance while the absolute values of the regression coefficients in the regions where moisture absorbs NIR radiation were less. The ability of the visible and short-NIR region to discriminate between early and late harvested samples is shown by the reasonable predictive accuracy of model *2 in Table E-15. This model only uses the 400-1100 nm region. Regarding the visible region, the presence/absence of the chlorophyll “hump” discussed in Section 14.2 may help to select for early/late harvest samples. Interestingly, however, the NIR only models, also presented in Table E-15, also allow for good discrimination.

14.4.2.2 LDA and SIMCA

LDA and QDA methods were employed on the same WC/WU calibration set as used for PLS-DA and tested on the same validation set. The results of models using different numbers of principal components (10 or 20), LDA or QDA, or using the scores from PLS-DA analysis are provided in Table E-17. It can be seen that increasing the number of PCs improved the predictive accuracy slightly for the validation sets and that the QDA method tended to improve the fitting of the calibration set but to the detriment of the validation set. The classification results from using LDA/QDA on the PLS scores from PLS-DA are similar to those from using the y -values from PLS-DA as a classification criterion, although one extra sample was misclassified.

SIMCA was also employed as a tool to discriminate between Early and Late samples but the results were poor. Attempts were made to improve the SIMCA modelling, by adjusting the value for α and by adjusting the number of PCs used to explain each dataset, but the resulting classifications of the samples in the validation set were still substantially inferior to those of the PLS-DA method, with most Late samples also being classified as Early.

14.4.3 Discrimination Between Miscanthus Varieties

14.4.3.1 PLS-DA

As shown in Table E-1 to Table E-10, there were significantly more spectra collected for the *giganteus* variety than for the *sinensis* or “other” varieties. In order to make the

calibration/validation sets more balanced, for each dataset a lower number of *giganteus* samples than the total available were randomly selected (see Table E-18 for details) and brought forward to a new dataset that also comprised all of the scans of all other varieties. From this set 75% of the samples were selected for calibration and 25% for validation.

Preliminary trials were conducted on the DF set. Initially a 3-variable PLS-DA was attempted with scores for *giganteus*, *sinensis*, and “Other” varieties. However the results from this calibration were poor. It was decided that the “HP” samples would be excluded from the calibration since there was little variation in the chemical composition of these samples (see Section 16.2) and the origin of these samples was unclear (see Section 14.1.1). It was then judged that there were insufficient samples in the *sinensis* and “other” categories to allow separate discrimination between these categories, particularly given the needs for separate calibration and validation sets. Hence, models were developed on the basis of discriminating between *giganteus* samples and samples of all other varieties (including *sinensis*).

The results of the PLS-DA according to this classification are provided for most of the datasets in Table E-18. There were insufficient samples covering the “Other” varieties to allow calibrations to be developed for the DH and DV datasets. It can be seen that, using the predicted y-value only, the results are reasonable although there are some selected instances of false-positives for *giganteus* and non-*giganteus* sample identification in each dataset. The WC/WU is the most important dataset with regards to potential applications in online biorefining systems, and the performance of the calibration for this set was good although it required significantly more factors for classification than the other models. Most of the “None” classifications for this [WC/WU] dataset in the “with-deviation” scenario are for the “Not-*giganteus*” class. Most of those samples came from the “other” rather than *sinensis* category.

Figure E-12 (a) presents a Factor 1 vs. Factor 2 scores plot for the DF model. On it the *giganteus* (GIG) and non-*giganteus* (NG) samples from the calibration set are plotted. It can be seen that, while these two factors alone do not fully separate the two groups (these only explain 60% of the total y-variance), there is some segregation apparent. The regression coefficients plot for the 5 factor model, Figure E-12 (b), shows that the visible region is important in the discrimination. This was observed in the evaluation of regression models; those that excluded this region demonstrated poorer levels of segregation between the two groups.

14.4.3.2 SIMCA and LDA

SIMCA analysis was undertaken for variety discrimination using the WC/WU dataset, where the spectra were transformed with SG-2,2,25,25. Initially the optimum number of PCs used to describe the X-variance in each set (*giganteus* or non-*giganteus*) selected by the software were used in the SIMCA classification. However, this led to most samples being classified in both groups, even when different values for α were chosen. Increasing the number of PCs improved the situation somewhat, but an excessive number of PCs could not be chosen for the not-*giganteus* calibration set given that it only had 38 samples and too many PCs could lead to an overfitting of this small calibration set. In a SIMCA classification, using 15 PCs to describe both models, the sensitivity was 100% for both groups but the specificity was only 94.1% for the non-*giganteus* samples, and only 33.3% for the *giganteus* samples, using $\alpha = 0.25$. Only one *giganteus* sample was classified as both types, whereas 10 of the 15 non-*giganteus* samples were classified as both types. Using smaller values for α resulted in more samples being classified as both types.

Table E-19 provides classification results for the calibration and prediction sets of the WC/WU dataset using LDA/QDA methods with either 10 or 20 PCs, or where the scores for the first 13 factors of the PLS-DA were used as the input variables. Increasing the number of PCs used improved, in most cases but not for the LDA calibration set, the classification of samples. The results for the validation set for the LDA/QDA methods using 20 PCs are superior to those of the PLS-DA method (one less sample is misclassified). When the PLS-DA scores are instead used as input variables it can be seen that the QDA method correctly classifies all samples in the calibration and validation sets. This method is therefore the most accurate of those tested for this discrimination.

14.4.4 Discrimination Between Plant Fractions

14.4.4.1 PLS-DA

Firstly, models were developed to discriminate between leaf and stem samples for the various datasets and using the *Miscanthus x giganteus* samples only. Since there was only one flower sample for this variety it was excluded from the models. The best models developed for each dataset are presented in Table E-20. It can be seen that, with the exception of the WC/WU dataset where one

stem sample was misclassified as a leaf, the predictive accuracy when only the predicted y-values were considered was 100%.

When the deviation in prediction was considered there were single occurrences of no-classification in some of the datasets with three samples not classified in the WC/WU dataset. Nevertheless, for the WC/WU dataset, the most important model in terms of potential for application in an online bio-refining system, the predictive accuracy is still acceptable particularly given the large number of samples in the validation set. That model, however, does require more factors than the models for the dry samples. Indeed, it can be seen in a F1 vs. F2 scores plot for the DF model, Figure E-13 (a), that the stem and leaf samples are mostly separated (82% of y-variance explained) in one factor.

Attempts then were made to discriminate between the following fractions: dead leaf blades, dead leaf sheaths, live leaf blades, internodes, and nodes. Flowers and live leaf sheaths were not included in this classification since there were insufficient samples for each of these classes (see Table E-1 to Table E-10). This discrimination involved a 5-variable PLS2 regression with these 5 indicator variables showing whether a sample was (a value of 1) or was not (a value of -1) classified as each fraction. The validation results of these PLS-DA models are provided in Table E-21 and Table E-22.

The results for most datasets, when using the predicted y-values as a basis for classification, were very good. For example, only one sample in the WC/WU validation set was misclassified (a dead leaf sample was predicted to be a dead sheath sample). The classifications when the deviation in prediction was considered were less accurate in some, but not all, cases. For the wet nodes and wet internodes all samples were correctly classified even when considering the deviation in prediction. This is not surprising given that these fractions were physically very different from each other, and from the other fractions, after processing in the chipper prior to taking the WC/WU scan (see Figure 14-2 and Figure 14-3 (a)). These fractions also tended to have higher moisture contents.

Figure E-13 (b) presents a F1 vs. F2 scores plot for the samples in the calibration set of the WC/WU PLS-DA model. These two factors only explained 46% of the total y-variance but there are clear formations of segregations between most of the groups.

14.4.4.2 LDA

The PLS-DA results for discrimination according to the two classes of stem/leaf were considered so good that testing of LDA/QDA was not warranted. However, LDA/QDA were examined for

discriminating between the five different plant fractions for the WC/WU dataset, Table E-23. In every case the spectra were first pretreated with SG2,2,20,20 and the datapoints over the 400-2500 nm region were used. The calibration/validation sets are the same as those used for the WC/WU model in Table E-22. Prior to LDA and QDA the matrices were decomposed to 20 PC's in order to avoid collinearity problems. LDA and QDA were also carried out on the PLS-DA scores obtained from the calibration and validation sets in the model outlined in Table E-22.

Table E-23 shows that, even taking the best validation results (LDA or LDA on the PLS-DA scores), the predictive accuracy is not superior to that of a PLS-DA where only the predicted y-values are used. The QDA methods provide closer fitting for the calibration sets but the predictive ability in validation is poorer. There are only minor differences between using LDA/QDA on the PC scores of the matrices and using LDA/QDA on the PLS-DA scores; two less samples were misclassified in the QDA validation set when the PLS-DA scores were used. The LDA and PLS-DA methods therefore achieve comparable results to each other, and either could be used with reasonable predictive accuracy for the WC/WU dataset. For less qualified persons using the software, the LDA method may be preferable since the predicted classes are directly provided by the program whereas the PLS-DA method requires transferring the matrix of predicted Y-values to an Excel spreadsheet.

14.4.4.3 SIMCA

The use of SIMCA for the classification of plant fractions was evaluated for the WC/WU dataset since this comprised the most samples for each fraction. That meant that more accurate PC models could be built from a calibration set and a separate validation set could also be used for testing the models. The results, using $\alpha = 0.05$, are provided in Table E-24, which presents a matrix where the columns represent the actual fractions and the rows represent the number of samples that were classified in each fraction. For example, although all node samples were correctly identified as nodes (a sensitivity of 100%), a number of node samples were also classified as dead leaves (10), dead sheaths (12) and internodes (14), meaning that the specificity for the node samples was only 56.6%. Also, 18 internode samples were identified as nodes.

The value for α was varied and the effects upon the sensitivity, Figure E-14 (a), and specificity, Figure E-14 (b), values for each fraction observed. It can be seen that there is no optimum value that will enable high sensitivity and high specificity for each fraction.

The number of PCs for each model was determined automatically by the software according to the explained X-variance under cross-validation. Since the first few PCs tend to explain the majority of this variance in wet samples (see Section 14.3) it suggests that insufficient of the relevant variance, for sample discrimination, was being modelled. Attempts were made to increase the number of PCs used in the SIMCA classification of each group but, although specificity increased, the results were not as good as those from PLS-DA. SIMCA was also attempted for the DS dataset, for which the influence of moisture would be significantly less. However, once again, high values for both specificity and sensitivity were not possible for all fractions under any α level.

SIMCA discrimination was tested for the leaf vs. stem section classification. As before, the number of PCs used for each model in the classification and the value for α were adjusted in order to find the optimum conditions for maximising both sensitivity and specificity. Those optimum conditions, 20 PCs for both models and $\alpha = 0.1$, provided a sensitivity of 92% for leaves and 94.2% for stems, and a specificity of 100% for both groups. The specificity of the stem classification was very sensitive to any reductions in the value of α , while increasing α to 0.25 reduced the sensitivities of leaf and stems to 80% and 81.4%, respectively (i.e. more samples were classified in neither group). Two leaf samples and two stem samples would not be classified to their own groups under any setting of α since their leverage values breached the critical limit. Both of these leaf samples were live sheaths. Hence, providing more live sheaths to the PCA model would help to reduce the leverage of these samples, and should allow their classification. A discrimination power plot (comparing both models) over 400-2500 nm was complex with dozens of peaks having discrimination power values greater than 3. However, the highest value was observed at 1351 nm.

14.4.5 Discrimination According to Stand Age

PLS-DA

The productivity of a *Miscanthus* plantation during the first year of growth is significantly less than in subsequent years. This was observed for the plants that were collected in terms of a lower total dry mass as well as a greater leaf:stem ratio (see Section 16.4). Discriminant analysis was undertaken to see if sections of these first-year plants could be distinguished from plants harvested from older plantations. This was done using the WC/WU scans since this was the largest dataset for the one-year plants (83 average scans).

A further 312 scans of plant fractions collected from *Miscanthus* stands of 2, 3, or 13 years were provided to the group for calibration/validation sample selection. All of these spectra came from *Miscanthus x giganteus* samples collected from the sites listed in Section 14.1.1. From this group a calibration set of 295 samples (65 from a 1 year plantation and 230 from plantations older than one year) and a validation set of 100 samples (18:82 ratio for 1-year:older) were randomly selected. The matrix was pretreated with SG2,2,20,20. A large number of factors (20) were selected for this model, based on the cross-validation tests within the calibration set. The predictions that resulted when the PLS-DA calibration was applied to the validation set are provided in Table E-25. LDA/QDA were also evaluated for the discrimination but the results were less accurate.

It can be seen from Table E-25 that, when only considering the PLS-DA scores, all of the samples from “older” plantations were correctly identified. However, four of the 1-year plantation samples were classified as coming from the older plantations. All of these misclassified samples were stem sections. A separate calibration, where stem sections were excluded, see Table E-25, was developed. Regarding this model the predictive accuracy for the validation set, considering only the predicted PLS-DA y values, was 100% for the 1-year samples although 2 “older” samples were misclassified. This model, which was the most accurate found for the leaves subset, involved no spectral pretreatment but required 15 factors. A PLS-DA model on leaves spectra pretreated by SG2,2,20,20 also misclassified one of the “1-year” samples in the validation set but only required 7 factors. A calibration, involving stem and leaf sections, was also made for the DT dataset, which contained much fewer “1-year” samples. The results are provided in Table E-25 where it can be seen that one 1-year sample and 2 older samples were misclassified when using the PLS-DA y -values.

The results from these calibrations are promising and suggest that NIRS has potential for the online discrimination of wet first-year *Miscanthus* samples. Such a discrimination would probably be much easier if chopped whole plant samples were provided to the scanning cell since the higher leaf:stem ratio of the first-year crops should provide spectral differences that should be significant within a few PCs/factors. However, the samples scanned in this study had been separated according to their plant fraction. Therefore discrimination methods required spotting spectral variations that any physiochemical differences between the two categories may provide.

The stems of first-year plants were, on average, much shorter and thinner than those of plants from older stands. However, all stem sections of the samples used in this PLS-DA analysis had been obtained from separate metre sections of the plant. Since stem thickness decreases with plant height, and since chemical properties also vary (see Section 16.1) a third metre stem section of a crop growing on an “older” plantation may have many similar physiochemical characteristics to the

first metre stem section of a plant growing in its first year, yet these samples need to be discriminated by the PLS-DA model.

While some stem sections of 1-year plants were misclassified in the validation set as coming from “older” plantations not all of these were. Furthermore, no stems sections of “older” plants were classified as “1-year”. Hence, the model is providing some discriminatory power regarding these stems and the large number of factors required hints at the complexity of this. The regression coefficients plot of the high-factor models was complex. However, the 7-factor model on the leaves subset indicated that the visible region was of importance. The improved predictive accuracy of the leaves PLS-DA model for 1-year samples probably occurs because the difference between the two categories is more marked regarding the leaves (the leaves are much smaller and more frail in the first year of growth and it was also noted that there were differences in colour).

A PLS-DA model was developed to see if the samples obtained from plants growing on the 13-year old plantations in Oak Park could be discriminated from all other samples, but when the model was applied to the validation set all the 13-year samples were misclassified as “younger” samples.

14.5 Summary

Various techniques have been evaluated for the discrimination of *Miscanthus* samples according to: plant fraction, age, variety, and time of harvest. The unsupervised clustering methods provide some degree of segregation between stem samples and leaf samples but this depends on the dataset in question with more homogeneous samples performing better. These methods are beaten, in terms of accuracy, by supervised classification methods which require the user to specify the classes of the samples in the calibration set. These methods (PLS-DA, LDA/QDA, SIMCA) allow prediction of class status for samples on an individual basis, and therefore are more relevant to potential online applications. They also allow discrimination according to plant variety, stand age (one year or older) and allow more plant fractions to be separated.

The SIMCA method, however, tended to achieve the lowest accuracies of these three supervised techniques. This is not surprising given that the method involves fitting unknown samples to separate PCA models built for each class. These models focus on explaining the maximum amount of X-variance and so the major PCs will often be similar between different classes. This is particularly true for the variety/harvest-date/stand-age discrimination since each PCA model needs to describe

the significant X-variance provided by the different plant fractions whereas the spectral variation according to the class status will be significantly less. In contrast, PLS-DA models the covariance of X and Y where Y is a column or matrix containing the relevant indicator variables. Hence, it models the between class variability which is usually more relevant for discrimination. The result is that PLS-DA performs much better as a discriminatory tool in this study.

LDA/QDA methods which rely on decomposing the spectra with PCs generally correctly predict more samples than SIMCA, but the results are usually not quite as accurate as when PLS-DA is used. The reason for this is again the reliance of PCA, although this time only one model is built compared to one for each class in SIMCA. When the PLS-DA scores are used as the matrix for LDA/QDA methods the validation results are comparable to using PLS-DA alone in some cases and superior in others. In particular, the use of QDA on the PLS-DA scores for *giganteus*/non-*giganteus* variety discrimination allows for a 100% success rate in both validation and calibration sets where this was not possible with PLS-DA alone. Hence, this combined method, while requiring more work than a simple one-step classification, is of value.

Regarding the discrimination between *Miscanthus* varieties, a total of 7 “Other” plants were sampled, each one representing a different *Miscanthus* variety/breed. A more detailed study centred on variety discrimination should focus on the collection of a smaller number of varieties and involve the collection of more samples of each. It would make sense to select those varieties that are grown commercially (e.g. *sinensis*, *sacchariflorus*, *giganteus* etc.), or at least grown in greater quantities than the block field trials that took place in Oak Park. With the development of these deeper datasets it may be possible, based on the compositional differences between the different varieties (see Section 16.2), to discriminate between each variety.

The collection, separation, and processing of samples (Section 14.1) was designed on the basis of providing samples that would differ substantially in their chemical and spectral characteristics and so allow accurate quantitative NIRS calibrations to be developed (Section 15) and also allow detailed studies to be made regarding the chemical compositions of various anatomical fractions within the same plant (Section 16.1). Qualitative discrimination methods were not used as a basis for sample selection. If a study focussed on qualitative discrimination were to be designed then it would involve more whole plant (WP) samples rather than separating plants by their anatomical fractions. This would also have more relevance to potential applications in online biorefining systems where WP samples are more likely to be received.

An Early/Late harvest calibration based on WP samples would probably be based upon the differences in mean chlorophyll content and in the stem:leaf ratio (or in the relative proportions of all plant fractions) between these two periods. Similarly, the discriminations according to stand age and plant variety would use the differences in the contributions of the plant fractions to the total mass balance as well as the differences in various chemical components.

This study, in separating different plant fractions and then trying to discriminate, for some methods, not on the basis of these fractions but on other criteria (stand age, harvest period, variety), has attempted a much more difficult classification than would be necessary for WP samples. The fact that these have been reasonably successful is highly promising and it appears very likely that the results based on whole-plant samples would be even more accurate given the accuracy with which plant fractions can be discriminated. However, the development of WP calibrations is not needed for good discrimination, the models outlined in this chapter have value in themselves and will be applied for the classification of unknown samples in the Carbolea laboratories.

15 Development of NIRS Quantitative Calibrations for Miscanthus Samples

This Chapter will discuss the development of quantitative NIRS models for the contents of various lignocellulosic components, ash, moisture, carbon (C), and nitrogen (N) for selected Miscanthus samples from Table E-1 to Table E-10 . These samples, and their means of collection and processing, are outlined in Section 14.1. The models developed in this chapter will allow quantitative predictions for samples that have not been analysed by reference methods (see Section 16.4). Section 10 discusses the background to Miscanthus and provides secondary analytical data for this feedstock. The methods used for the reference analysis of these Miscanthus samples are outlined in Section 11. Many of the Figures and Tables referenced to in this chapter are provided in Appendix F and the reader is referred to Appendix A for a summary of the abbreviations used in this chapter.

15.1 Development of NIRS Models

Since the time involved would have been excessive, not all of the samples in Table E-1 were analysed using reference methods. Furthermore, due to changes in the means of spectra collection (moving from the DS to DT, DG to DH, and DU to DV system, see Section 11.1) not all of the samples that were analysed with reference methods are represented in each particle-size/sample-presentation method dataset. Towards the end of the research all of the samples for which there was a sufficient amount of material remaining were scanned in the DS/DT method if these had not been scanned in this method before. However, this was not done for the DG/DH or DU/DV samples meaning direct comparisons between these two methods, using the same samples, were not possible. Table F-1 provides a summary of the number of Miscanthus samples (of either the *giganteus*, *sinensis*, or “Other” varieties) that were analysed via reference methods to an acceptable method of precision for each constituent of interest and for which spectra for each of the relevant datasets exists.

The most important datasets are the DS/DT and WU datasets, the former for the development of the most accurate calibrations and the latter for the use of NIRS as a rapid primary analytical tool without the need for extensive sample processing. The WU method of spectra collection did not change throughout the project meaning that for every DS/DT spectrum there will be a corresponding

WU spectrum (with the exception of the samples received from Germany, see Section 16.2, which were dry on arrival). This allows the three datasets to be compared on a like-for-like basis.

The large number of samples in most of these datasets allows for the use of true random sample selection for the validation sets. As with the peat samples, 75% comprise the calibration set and the remainder the validation set. All of the samples for which precise reference analytical data were available were grouped together, using the reproducibility between the duplicates of the GLU_SRS (glucose) data as the selection criterion (see Section 11.7). This constituent was chosen since it can be considered to be the most important single constituent in terms of utilisation in many biorefining technologies and also because its precision provides a good indication of the quality of the whole analytical hydrolysis method; i.e. the quality of the results for the other sugars will reflect that of glucose. In some cases the KL/AIR results from the same hydrolysis batch may not be as precise as those for the sugars since these may be susceptible to errors in the gravimetric determination of these constituents; however, in such cases the hydrolysis was often repeated specifically for the determination of their values. The GLU_SRS data were also chosen because these are corrected to a whole-mass basis using the extractives content; hence GLU_SRS values will not exist unless there are also precise EXTR_PD data.

From this collection of samples, the samples that existed in each of the DS, DT and WU datasets were brought forward to another group which was then used for random sample selection for the validation set. That meant that the same samples are in the global validation sets for each dataset. Since reference analytical data may be absent for some other constituents, due to poor reproducibilities, the number of samples in the validation sets may differ between the different constituents but these will be the same between datasets for the same constituent. The calibration samples are classified as all the other samples that have precise analytical data for a specified constituent and dataset. Hence, the number of calibration samples will differ between the WU, DS and DT models since some only had DS or DT spectra collected in addition to their WU spectra. Nevertheless, as can be seen in Table F-1, the structure of the models will be similar with the majority of (precisely characterised) samples being in the calibration sets of all datasets.

For the DU, DG, and DF datasets, random selection was used for validation sample (25%) selection, again using the precision of the GLU_SRS data as a basis. Due to the limited number of samples with DH and DV spectra, cross-validation was used to test models for these datasets.

Models were developed based on calibration/validation sets with all the different *Miscanthus* varieties grouped together, and also on sets only comprising samples of the *giganteus* variety. The

same methodology for calibration/validation set sample selection was used in both cases. For the elemental analyses, samples with reproducible carbon data (rather than GLU_SRS data) were used as the basis group for calibration/validation sample selection. Since there were much fewer samples with elemental analysis data, models were also developed with all of the samples in the calibration set and the model tested with cross validation methods. Cross validation was the only method of validation for the uronic acids models due to the limited number of samples analysed.

15.1.1 Combination of DS and DF Data

While methods of sample comminution were developed with a target of maximising the DS fraction and minimising the amount of DF material, the DF fraction still represented an important proportion of the total mass balance of each sample (see Section 16.3). Methods were tested, as discussed in Section 15.2.2.5, to see if the combination of the DF data with the DS data to produce a weighted average composition (with the relative proportions of each fraction being the weights) could result in improved WU models. These models are labelled WU_{DG} . These needed to be compared on a like for like basis with the WU models based only on the DS data (i.e. WU_{DS}); hence, the WU_{DS} models only used the samples for which WU_{DG} data were available. These models were validated using full cross-validation since they only comprised a total of 36 samples (for the sugars data).

The second method of combining the DS and DF data involved using the quantitative models developed for the DF dataset to predict the compositions of all the DF samples that were scanned in the large XDS cell. These predictions were then combined in a weighted average with the reference analytical DS data. The models developed based on these samples are referred to as WU_{DGP} and are compared against WU_{DS} models that comprise the same samples. For these calibrations the samples for which WU_{DG} data (i.e. weighted averages involving real DF data) existed were excluded from the calibration set and instead comprised the independent validation set where the WU_{DG} constituent data were used to test the WU_{DGP} and WU_{DS} models. The theory was that if there were significant and consistent differences between the DF and DS data then the WU_{DS} models would show a much greater bias in validation than the WU_{DGP} models would. The total number of samples in the WU_{DGP} and WU_{DS} models were less than those in the normal WU models since for many samples there was insufficient DF material to allow this fraction to be scanned in the large XDS cell.

15.2 Results

15.2.1 Reference Analytical Data

Chapter 16 will discuss important trends regarding the differences in composition between different parts of the same plant, between plants from the same locations throughout the harvest period, and between the different varieties of *Miscanthus*. This section will discuss how the distribution of samples in the concentration ranges of relevant constituents will affect quantitative NIRS calibrations.

Table F-2 to Table F-10 present histograms, with associated statistics (including the SEL), for selected components based on a set comprising the DS samples of all *Miscanthus* varieties as well as on a set comprising only the *giganteus* DS samples and a set comprising the DF samples of all varieties. These histograms and associated statistics do not include samples of *Miscanthus* flowers since these were excluded from the NIRS models.

The compositional data of the *Miscanthus* samples share a basic similarity to that of the sugar bagasse samples, with glucose being the major constituent (39.1% on average over all DS samples), xylose the second largest (19.9%), closely followed by Klason lignin (18.2%). Arabinose and galactose are, like the bagasse samples, present in much smaller quantities but there appears to be sufficient variation across the samples to suggest NIRS modelling of these constituents may be possible. As also seen with bagasse, rhamnose and mannose are minor carbohydrates in these *Miscanthus* samples; however, the range in concentration values is greater than those seen for the bagasse samples. This is a consistent trend in these histograms - the range in composition for many constituents is increased over that seen for the sugarcane bagasse and, for some constituents, greater than the ranges seen for the peat samples. This extension in concentration ranges is a result of the sampling strategy which separated many of the plants collected according to their anatomical fractions (as discussed in Section 14.1.1). As a comparison, the group of whole plant (WP) samples collected only demonstrated a range in glucose content of 13.3%; lower than the 21% in Table F-2. This extension in concentration ranges allows for improved RER values to be achieved if RMSEP values are consistent. For example, an RMSEP of 1% for glucose would give an RER value of less than 15 for the WP model but give a value of 21 for the "ALL" DS model, a value indicating that the model is suitable for future quantitative predictions of unknown samples (AACC, 1999).

Regarding the ranges in concentration values, for some constituents (e.g. glucose, xylose, rhamnose, ash, AIA) these are increased when samples of other varieties of *Miscanthus* are included in a group with samples of *Miscanthus x giganteus* (the “ALL” histograms) compared with a group only containing samples of *Miscanthus x giganteus*. One of the most significant examples of this is for the xylose content (XYL_SRS, Table F-3). For the *giganteus* DS samples the range is 8.47% but this rises (by a relative 41.20%) to 11.96% for the “ALL” set. The range is extended in both directions, with sample of “Other” varieties having xylose contents greater-than and less-than the extremes in the *giganteus* set. Hence, if calibrations with similar precision (RMSEP) could be developed for both models, then improved RER and RPD values would result for the “ALL” models. Of course, the benefit of including extra varieties of differing compositions in the dataset will only result in an improved model if the added spectral complexity that may result from their inclusion could be accounted for by the PLS factors and inflated standard error values would not result.

The DF samples usually demonstrate a lower range in concentration values than the other sets. This must be taken in context given that much fewer DF samples were analysed. Given this lower concentrations range, the smaller size of the DF dataset, and the wide variety of anatomical fractions that populate it, it appears that the development of accurate NIRS models may be more difficult compared against developing models for the DS samples. The DF samples do demonstrate an increased range in concentration for 95%-ethanol soluble extractives, however, and the average extractives content of these samples is also significantly greater than that of the “ALL” and “GIG” samples.

The DF histograms also appear significantly different in structure from the “ALL” and “GIG” histograms. This is a result of the methods by which DF samples were selected for reference analysis (see Section 15.2.2.4). Regarding the structure of the DS sets, the histograms vary in shape according to the constituent of interest. For some (e.g. xylose) the distribution is somewhat normal whereas in other cases (e.g. ash, extractives, AIA, rhamnose) there is a positive skew and in other cases (e.g. total sugars) there is a negative skew. For the glucose content there is a quasi-bimodal distribution with a local peak around 34% before a larger peak at around 42%. The first peak is associated with the majority of the live leaf blade samples (see Section 16.1) and the latter peak with the majority of the stem section samples. No constituents have histograms that conform to the ideal flat distribution for NIRS calibrations; however, the arabinose and KL histograms are perhaps closest.

Table F-7 and Table F-8 provide the histograms for the elemental analysis of the 53 DS samples and the 51 DF samples that were analysed. The relative variation of these elements differs greatly. If the standard deviation in concentration values is expressed as a percentage of the average value, then

values of 1.8%, 1.9%, 76.7%, and 180% are obtained for the C, H, N, and S contents of the DS samples of all Miscanthus varieties. The value for sulphur is somewhat misleading, however, given that all samples but one had sulphur contents of less than 0.2% and that the standard deviation is inflated by one node sample which had a sulphur content of 0.57%. In fact, in many instances the quantity of sulphur present in the sample was below the limit of detection of the analyser and a quantity of 0% was recorded. As with sulphur, the carbon and hydrogen concentrations of the samples analysed occupy a quite limited region of concentration space.

The SEL column in the Table associated with each histogram provides, in the row giving the standard deviation statistics, a value for the SEL and, in the other rows, this SEL value is expressed as a percentage of the average value. It is important to mention that for carbon, hydrogen, and sulphur the SEL is relatively high when compared with the standard deviation of concentrations (the SEL is 18.2% of the SD for carbon, 33.3% for hydrogen, and 22.2% for sulphur). In contrast, the SEL is only 10.7% of the SD for nitrogen. While the majority of samples have a nitrogen content of less than 0.5%, there are a significant number of samples with greater nitrogen contents and some that have N contents of over 2%. These high nitrogen concentration samples were all live leaf blade samples that were collected early in the harvest window.

Quality of the Analytical Data

Regarding the SELs of other constituents, these are, in most cases, superior to those obtained for the peat and bagasse samples. For the major sugars, glucose and xylose, the SEL is less than 1% of the mean. For galactose, arabinose, and rhamnose the SEL is proportionately greater but still less than 5% of the mean. For mannose, however, the precision of the analysis is significantly poorer; the SEL is 13.1% of the mean for the set comprising all DS samples. The reason for this poor precision for mannose is that the chromatography software was inconsistent in its placement of peak delimiters for this constituent in cases where its concentrations were particularly low. The chromatographic method employed (see Section 4.6.2.3) was designed to allow the resolution of all sugars at concentrations which did not exceed the linear response of the detector. Since glucose and xylose were the most abundant and important sugars the dilution factors were calculated to maximise the precision of these sugars.

The 5X dilution used for chromatography is perhaps too large for enabling precise mannose data for these Miscanthus samples. Indeed, in some cases, no peak for mannose was detected by the software. Occasionally this occurred for only one of the duplicates (i.e. one duplicate had no

mannose peak while the other did). Figure F-1 presents a chromatogram of a *Miscanthus* internode sample. Mannose is the last eluting peak in this chromatogram.

The SEL of mannose could probably be improved with a lower dilution factor, or through direct injection of the hydrolysate; however, this would be to the detriment of glucose and xylose analysis and peak resolution. Given that mannose is such a minor constituent there is no great need for a highly precise NIRS calibration for it, at least in relation to the utilisation of this feedstock in biorefining technologies. Hence the data obtained for mannose in this study are considered to be acceptable.

With the exception of AIA, all of the constituents determined by gravimetric methods (KL, ash, AIR and extractives) have good SEL values that are less than 5% of the mean. Acid insoluble ash (AIA) has an SEL value that is the lowest of all these constituents (0.15%). However, when expressed as a percentage of the mean it is the greatest (11.5%). As with mannose, however, this is not a major constituent in the feedstock, nor is it of direct relevance to biorefining platform-chemical yields, meaning that the slightly lower relative precision for this analyte is also acceptable.

Sugar Recovery Data

Summary statistics for the sugar recovery (SR) rates of the sugar recovery solutions (SRS) in the 86 hydrolysis batches are presented in Table F-11, and Figure F-2 presents these as a plot with increasing batch number. For early batches only two sugar recovery solutions (SRS) were used in each autoclave run, but from batch 387 onwards three SRS tubes were used. The standard deviation in recovery rates between these 2/3 tubes were determined for each of the constituent sugars and form the basis of the second batch of columns (“average standard deviation within a batch”) in Table H-1. The SRS used to determine these recovery factors had similar concentrations of sugars to that of the samples being analysed meaning that the mannose concentration was low. This explains the high standard deviation values and poor reproducibilities for this constituent. It and rhamnose have been excluded from Figure F-2 to improve the clarity of the other, more important, sugars. For these sugars the recovery rates are reasonably consistent from the first to the last batch. The recovery rates are very similar to those for the bagasse samples (Section 12.3.1) which is unsurprising given that the same SRS solutions were used.

Correlations between Constituents

Table F-12 presents the correlation coefficients between selected constituents for all of the *Miscanthus x giganteus* DS samples that had good analytical data (see Table F-1). It shows that,

while the absolute correlation coefficient values are generally not as large as those for the peat samples, there are some interesting relationships present. These include:

- Glucose ($r = -0.855$) and total sugars ($r = -0.865$) negatively correlated with the nitrogen content. This is understandable given that the nitrogen contents are highest and the glucose contents lowest in the leaves (particularly the “live” leaves). Nitrogen is required as part of the photosynthetic pathways of C_4 plants (Taub and Lerdau, 2000); hence its relative abundance in the leaves. Cellulose is proportionately less in the leaves than in other regions of the plant as discussed in Section 16.1.
- Glucose is strongly negatively correlated with ASL ($r = -0.905$). KL is also negatively correlated with ASL ($r = -0.793$). Again, this is likely to be linked to the increased presence of ASL and decreased presence of KL/cellulose in those regions of the plant more associated with the assimilation of primary metabolites than structural support.
- Arabinose, galactose, rhamnose, and ash are all negatively correlated with glucose and KL. This is also likely to be a result of the differing proportions of these sugars and constituents in the leaves as compared with stem sections.
- Much lower correlation coefficient values for xylose, in general, than for other constituents.

More details regarding the trends observed in the chemical compositions of the different plant fractions are discussed in Section 16.1.

15.2.1.1 PCA of Chemical Data

A principal component analysis (PCA) was conducted on a dataset comprising the constituent data for the DS particle size fraction of the *Miscanthus x giganteus* samples. The following constituents were included: EXTR_PD, ASH, KL, ASL, ARA_SRS, GAL_SRS, RHA_SRS, GLU_SRS, XYL_SRS, MAN_SRS, C and N. For a sample to be included it needed to have precise (see Section 11) analytical data for each of these constituents. The final set comprised a total of 33 samples. PCA models were built on the raw data and on the mean normalised data. Figure F-3 (a) shows the explained variance plot for up to 10 factors for the PCA on the original data and Figure F-3 (c) presents this plot for the mean normalised data. The blue line represents the explained variance in the calibration set and the red line the explained variance estimated via full cross validation. Figure F-3 (b) presents the PC1 vs. PC2 loadings plot for the raw data model and Figure F-3 (d) the same plot for the model based on the

mean normalised data. Generally, a variable with a high loading in the first model has a low loading in the second. This is because the constituents that contribute a greater degree to the overall mass balance of the samples (e.g. the xylose, glucose and lignin contents) show lesser relative variation (after mean normalisation) than the less concentrated constituents (e.g. nitrogen, mannose).

PCR Using Chemical Data

Using the 12-constituent PCA variables (after mean-normalisation), PCR was used to predict the concentration values for each variable in regressions involving all of the other variables. Summary statistics of these PCRs are provided in Table F-13 and the regression coefficients for each model plotted in Figure F-4 and Figure F-5. The RMSEC/RMSECV and offset values have been corrected, from the mean-normalised number provided by the model, to a whole-mass basis so that these can be compared with the corresponding values for the NIRS calibrations in Section 15.2.2.

As with the correlation coefficients in Table F-12, the R_{CV}^2 values in Table F-13 for most of the constituents are lower than those obtained for the peat PCR models. This is not surprising given that the Miscanthus samples in these models are from various sections of the plant and, correspondingly, have different concentrations-of and relationships-between each of these constituents. However, a reasonable R_{CV}^2 value is obtained with only one PC in a glucose content calibration. There are several important variables in this regression, as displayed in the regression coefficients plot in Figure F-5 (b). However, the largest absolute regression values are for the nitrogen content which, as Table F-12 shows, is negatively correlated with glucose.

An excellent R_{CV}^2 of 0.957 and an RMSECV of 0.239% was obtained in the ASL regression, using only 2 PCs. Nitrogen, extractives, ash, arabinose, galactose and rhamnose all have important positive regression coefficients in this model, while KL and glucose have negative regression coefficients. Correspondingly, the PCR models developed for these constituents follow a similar trend. These relationships correspond to the differences between the compositions of the leaf and stem sections, discussed in Section 13.3.1.

Extractives, ash, xylose, mannose, and carbon have relatively poor R_{CV}^2 values indicating that there is a lot of extra information related to their variance in composition that is not provided by the variances of the other constituents. Using the RMSECV of 1.02% for xylose would give an RER value for this constituent of less than 10 which is well below the ideal value of 15+ for accurate quantitative models. For NIRS models to achieve improved RMSECVs and RER values the PLS models will either need to find regions of the spectrum where the changes in the absorbance are directly related to the concentration of xylose in the sample, or find regions where there are indirect

relationships better than those outlined in Table F-13 (e.g. direct absorbances of other constituents that are more closely related to the xylose content). In contrast, a good ASL NIRS calibration may be possible based purely on indirect relationships between the spectra and ASL concentration.

15.2.2 Quantitative Calibrations

Spectral Transformations

Table F-15 to Table F-29 present regression statistics for the models, comprised of either all *Miscanthus* varieties or only *Miscanthus x giganteus* samples, for all datasets. For the DS, DT, DG, DU, and WU models an upper limit of 20 PLS factors was set, while for the smaller DH, DV and DF datasets the limit was 12 factors. As in the development of the bagasse and peat models, a variety of spectral pretreatments were tested to examine their impacts upon the precision and accuracy of the models. These included: SNV, SNVDT, MSC, EMSC, and a wide number of SG treatments varying in their derivative order (from 1 to 4) and total number of smoothing points (from 8 to 100). Table F-14 includes the regression statistics for the cross-validation of the calibration sets (all *Miscanthus* varieties included) of some of the models tested for the constituent glucose and the DT and WU datasets. For the datasets dealing with dry samples the differences were generally quite small between most of the spectral pretreatments employed. Indeed, models based on the raw spectra often performed reasonably well, as indicated by the RER value of 19 for a glucose model developed on the raw DT spectra (see Table F-14).

The situation was somewhat different for the WU spectra. The statistics in Table F-14 show that models based on MSC, SNV, or SNVDT transforms have significantly larger RMSECV's than the models based on SG treatments, and in some cases, greater RMSECVs than models developed on the raw spectra. It appears that there is some important information in the spectra that is being removed when these scatter correction techniques are applied, and derivatives appear a more prudent pretreatment choice.

As with the bagasse and peat calibrations, limited wavelength regions were tested for the WU models in order to see if excluding the region where water absorbs most strongly could allow for a model of reduced complexity that could be more targeted towards the prediction of the analytes of interest; possibly enabling improved predictions. Table F-14 shows the example of models developed on only the 1100-1800 nm region and also of a model developed on the combined 1100-1800 nm and 2100-2500 nm regions (i.e. excluding the 1801-2099 nm region) and compares these

with models developed over the 1100-2500 nm region. It can be seen that, regarding the models developed on the spectra treated by SG-1,1,10,10, the 1100-1800 nm model has an RMSECV that is 0.1% less than the 1100-2500 nm model, while the RMSECV for the 2-regions model is very similar to that for the 1100-2500 nm model.

While Table F-14 only presents the results for glucose, similar trends were observed for the other constituents. Based on these observations either the SG-1,1,10,10 or the SG-2,2,25,25 treatments (depending on the constituent of interest) were applied to the spectra prior to model development. The pretreatment method was kept consistent between datasets. Whether the wavelength range of 1100-1800 nm or the 1100-2500 nm region was used for WU model development depended on the constituent. There was no case of constituent or dataset for which the inclusion of the visible or short-wavelength NIR (<1100 nm) region improved the predictive ability of the model.

Number of PLS Factors Used

An important point to note is that, in Table F-14 and in many of the subsequent Tables, a relatively large number of factors were selected for the models. As with the bagasse and peat samples, Haaland's criterion was used to select the optimum number of factors. Due to the much larger number of samples in most of the Miscanthus datasets, compared with the bagasse and peat datasets, the F test that is part of Haaland's criterion (see Section 6.8) allows for improved confidence in the selection of the number of factors and results in an RMSECV that is closer to the minimum PRESS. It may be argued that the numbers of factors selected are large and greater than those in other studies (see Appendix B). However, there are no studies that the Author is aware of that cover both such a large number of samples and a wide variety of sample types. The physical (see Section 14.1.2) and chemical (see Section 16.1) differences between the different plant fractions are significant, especially in the WU state, and it makes sense that there will need to be sufficient PLS factors to explain these. Indeed, when tests were made with calibrations developed on each of the plant fractions separately much fewer PLS factors were selected. This was even the case for the calibrations developed on the chopped whole plant (WP) samples; Haaland's criterion selected 6 PLS factors for the WU model based on the 1100-1800 nm region and the SG-1,1,10,10 treatment, compared to 13 factors for the same model developed on samples of all plant fractions.

It should also be noted that, even for the calibrations developed on all fractions and all varieties, decent models were possible using a much lower number of factors than those selected under Haaland's criterion. This is illustrated in Figure F-6 with a plot of the RMSECV, with increasing numbers of PLS factors, for the glucose models for the DT and WU datasets. For example, the

RMSECV of the DT glucose model using 4 PLS factors was 1.39%, enough to give an RER value of over 15. However, what is most important is whether these extra PLS factors are overfitting the model to the calibration set to the detriment of the accurate prediction of future unknown samples. However, this is proven not to be the case, as shown in Table F-15 to Table F-29 where the RMSECV and RMSEP values are generally similar.

Samples Excluded

Appendix A describes the nomenclature used in the Tables in this chapter. These Tables also include a row listing the samples excluded from the model. The samples are represented by their NIR number. The classification of plant fractions under this numbering scheme took place as follows:

- 100-300 = Internodes
- 2000-2100 = Stem sections (nodes and internodes in one sample).
- 2500-2600 = Whole plant (WP)
- 5000-5200 = Nodes
- 10000-10100 = Dead leaf blades
- 12000-12100 = Dead leaf sheaths
- 14000-14100 = Live leaf blades
- 16000-16100 = Live leaf sheaths
- 18000-18020 = Flowers

Where any spectra of flowers existed in a dataset they were excluded from model development, except in the case of elemental analyses where a sufficient number of flower samples were analysed to allow for their inclusion in models. Table F-15 to Table F-29 show that, for a specific dataset, the samples excluded from the models tend to be consistent between constituents. The samples excluded were consistent outliers in predicted y vs. reference y plots for many constituents indicating that there was either an error in the collection of the spectra or in the labelling of the samples following spectra collection. Figure F-7 provides an example of these clear outliers for the chosen glucose WU model (see Table F-15). If a sample is mentioned as excluded, for a specific dataset, for a certain constituent but not for another constituent this is most likely because data were not available for that sample for the second constituent. No arbitrary deletion of samples to improve models on a case by case basis took place. The results presented in Table F-15 to Table F-29 are the final result of the iterative process of model development; if samples were predicted y vs.

reference y outliers in earlier versions of the model then these were put through reference analytical methods again, a technique that solved the vast majority of cases.

15.2.2.1 Calibrations for Lignocellulosic Components

The models developed will now be discussed according to constituent. Results for the DF models will also be presented in this section but will be discussed in more detail in Section 15.2.2.4.

Glucose and Total Sugars

Table F-15 provides the results for the best models found for the various datasets and for models based on all *Miscanthus* varieties or for models only based on the *giganteus* samples. In every case the spectra were treated with SG-1,1,10,10 prior to model development.

The RMSEP ranged from 0.862% for the DS model incorporating all *Miscanthus* varieties to 1.388% for the DU model based only on the *giganteus* samples. Table F-15 shows that the RERs for all datasets except DU and DF were excellent with values over 15. The results for the WU models are particularly good given the high moisture contents (see Section 15.2.2.6) of many of these samples. For most models there is not a great difference in the RMSEP values between the model based on all varieties and the model based only on *giganteus* samples, with the exception of the WU model where the RMSEP for the *giganteus* model is almost 0.5% less. This is interesting given that the same samples were in the validation sets of the DS, DT and WU models. Indeed the RMSEP of the WU *giganteus* model is the least of all the models. This probably occurs by chance, with low residual samples (for this constituent) being in the validation set; indeed, the RMSECVs for these two WU models are much closer. Figure 15-1 (b) provides a predicted vs. reference plot for the All-Varieties WU model.

In a Hotelling T^2 vs. Q-residual plot for the 14 factor All-Varieties WU model, node and leaf sections tended to have the greatest leverage/residuals whereas internode sections contributed less to the model and had lower X-residuals. In a plot of Y-residuals vs. predicted Y there was no apparent structure to the residuals with increasing glucose content; this was also the case for the models developed on the other datasets. For the WU All-Varieties model, the first PLS factor explained 98% of the X variance but only 4% of the Y-variance and demonstrated strong loadings in the 1400 nm region; clearly this factor was explaining the significant effect that moisture has on the spectra even

when the 1800-2500 nm region is excluded from model development. This was a consistent phenomenon in the WU models developed for many constituents.

Figure F-8 presents a plot for the regression coefficients of a 5 factor PLS model for the glucose content of WU samples. This is a much lower number of factors than were used in the model presented in Table F-15, however the coefficients plot for the 14 factor model plot was much noisier. The plot presented in Figure F-8 is easier to interpret and still represents a model of decent predictive ability (the RMSECV was less than 1.5%).

Since the WU model excluded wavelengths longer than 1800 nm the characteristic absorption regions for cellulose around 2100 nm and 2326nm (Üner et al., 2011) could not be used in the models. Instead the greatest absolute value for the regression coefficients occurs at around 1670 nm, this is a region that has been used in previous studies to predict the acid detergent fibre content of forages (McLellan et al., 1991, Norris et al., 1976). There is also an important region around 1690 nm, which has been associated with the 1st overtone of an O-H stretch in crystalline cellulose (Tsuchikawa and Siesler, 2003b, Tsuchikawa and Siesler, 2003a) .

Regarding the DS and DT models, it was expected that the DT sample presentation method would allow for improved regression statistics. However, this was not the case according to the independent validation set, although the RMSECVs are approximately 0.05% lower. In contrast with the WU models, the first PLS factor of the DS/DT models for many constituents explained a large amount of the Y-variance. In the instance of glucose, 47% was explained by F1 and 39% by F2; this total explained Y-variance of 86% was greater than the X-variance explained by these factors (81%). Figure 15-1 (a) provides a predicted vs. reference plot for the All-Varieties DT model.

The poor relative performances of the DU models are interesting, and a phenomenon that was repeated for many constituents. The failings of the DU sample presentation method have been described in Section 11.1 and it seems that the poor RMSEPs seen for the DU models may result from the chemistry of the material presented to the NIR cell window during DU spectra collection being different from the whole-mass chemistry of the sample when put through the reference analytical methods. There would be a bias towards the finer particles in the DU fraction whereas the coarser material would later be ground up for DG/DH/DS/DT scans and subsequent reference analysis. The RMSECV of the DV model is substantially less than that of the corresponding DU model; however, given that only 30 DV-scanned samples were analysed it is too early to conclude as to whether this new sample presentation method will improve NIRS models (as is also the case for the improved RMSECVs seen for the DH model compared with the DG model).

Similar trends to those seen for the glucose models are observed regarding the models for total sugars content, Table F-16. RER values of over 15 are possible for the DS, DT, DG, and WU models while comparatively large RMSEPs are seen for the DU models.

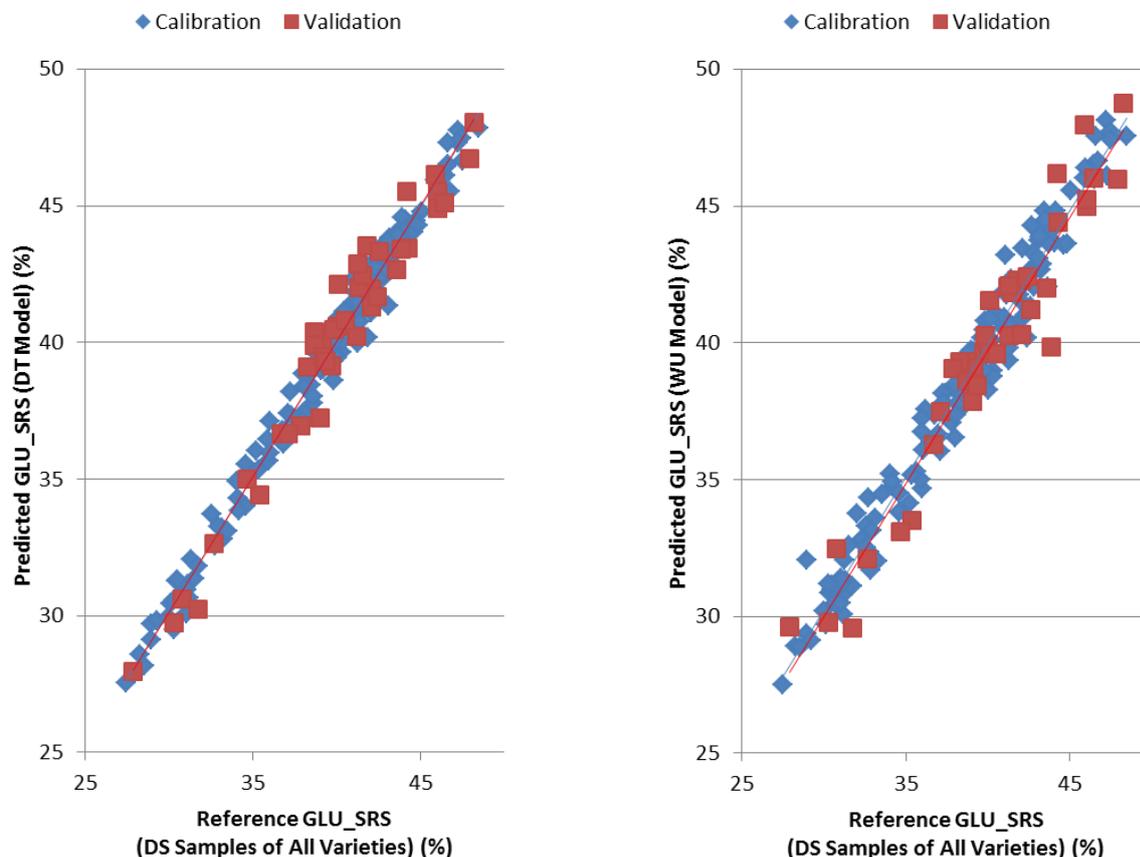


Figure 15-1: Glucose regression plots. (a) Predicted glucose vs. reference glucose for the DT model; (b) the same plot but for the WU model. Refer to Table F-15 for descriptions of these models.

Xylose Content

Refer to Table F-17 for the regression statistics for this constituent. For the DS, DG, DU, and WU datasets the models developed only on *giganteus* DS samples have greater RMSEPs than those developed on DS samples from all varieties. These differences are much greater when comparing the RER values for the “ALL” and “*giganteus*” models. This is a result of the concentration range being significantly extended when all varieties are included, as discussed in Section 13.3.1. Indeed, the RER value for the DT “ALL” model is also greater than that for the DT *giganteus* model despite the SEP for the “ALL” model being greater. Most importantly, an RER of 17 and an RMSEP of 0.53% is possible for the WU scans, indicating that, like glucose, an NIRS calibration suitable for the quantitative prediction of unknown samples can be developed on minimally processed wet samples.

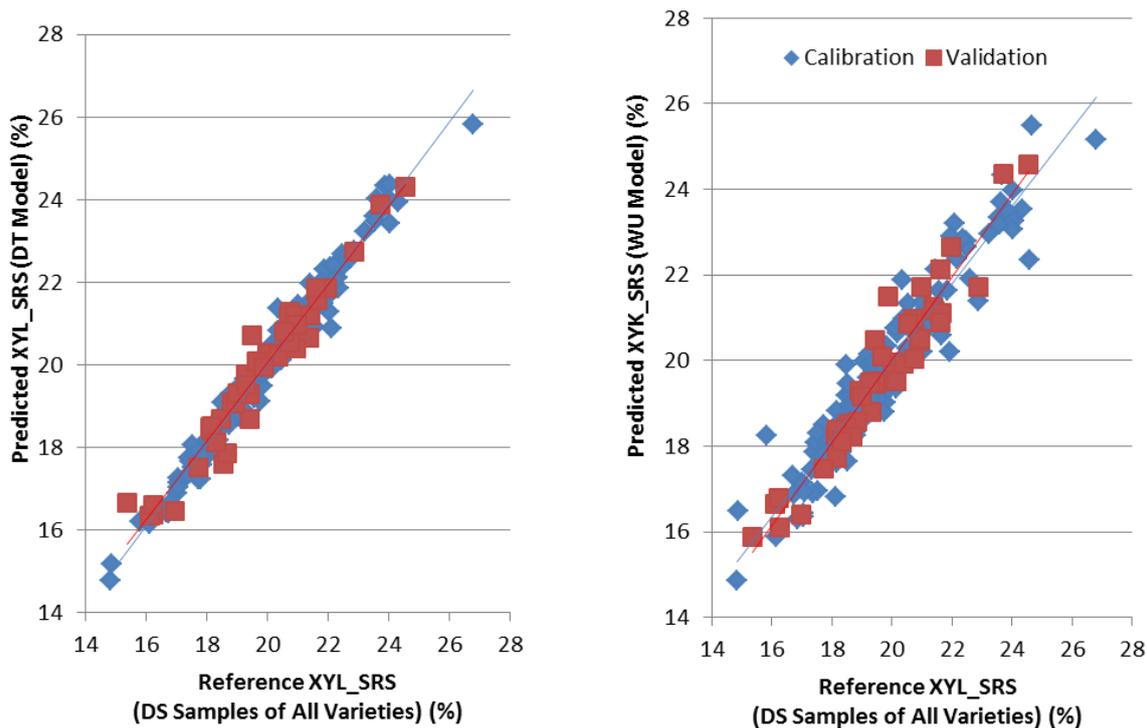


Figure 15-2: Xylose regression plots. (a) Predicted xylose vs. reference xylose for the DT model; (b) the same plot but for the WU model. Refer to Table F-17 for descriptions of these models.

For all xylose models the spectral pretreatment of SG-2,2,25,25 was used. It provided slightly improved RMSECV values compared with the first derivative. In a plot of y-residuals vs. predicted y value for the WU model it was noticed that the greatest residuals occurred for the whole plant and leaf samples, whilst the internode and node samples tended to have lower y-residuals. Figure 15-2 (a) provides a predicted vs. reference plot for the All-Varieties DT model and Figure 15-2 (b) the same plot for the all-varieties WU model. It can be seen that, for the calibration set, there is one outlying sample of high xylose concentration and two outlying samples of low xylose concentration. Extension of the validation set to these concentration ranges could improve the RER values further, as demonstrated by the excellent RER_{CV} value of 31 for the DS All-Varieties model.

Figure F-9 presents a plot for the regression coefficients of a 12 factor PLS model for the xylose content of WU samples. Given that it contains more PLS factors than the regression coefficients plot for glucose content, Figure F-8, it is a more complex plot but there are some wavelength regions that are of greater importance. In particular, the region between 1680 and 1730 nm has the largest absolute values for the regression coefficients. An absorption at 1724 nm has been attributed to the 1st overtone of a C-H stretch in hemicellulosic sugars (Tsuchikawa et al., 2005, Mitsui et al., 2008), while an absorption at 1678 nm has been attributed (Fackler et al., 2007) to the 1st overtone of a C-

H stretch in acetyl groups (the number of such groups is likely to be related to the hemicellulose content).

Arabinose Content

Table F-18 presents the regression statistics for the arabinose models. For all datasets the spectra were transformed with SG-2,2,25,25 prior to model development.

In contrast to the situation for the glucose, xylose, and total sugars models, the arabinose models developed on the DT spectra had lower RMSEPs than the models developed on the DS spectra. The effect of this was that, while the RER for the DS All-Varieties model was just below 15, it was above this threshold for the DT model. The RERs were lower and RMSEPs higher for all the other datasets (except the DH model). However, the RERs were greater than 10 (indicating a model suitable for classification) for all models except the *DF-giganteus* model. Interestingly, the DU models performed much better for this constituent than for the constituents presented in the previous Tables; these were comparable to the WU models. Also, with the exception of the DF dataset, this constituent is the first in which the *giganteus* models have lower RMSEPs than the All-Varieties models, for each constituent, although the differences between the two are small. Figure 15-3 (a) provides a predicted vs. reference plot for the all-varieties DT model and Figure 15-3 (b) the same plot for the all-varieties WU model. Sample 2513 (a WP sample) is outlying in the WU plot but not the DT plot.

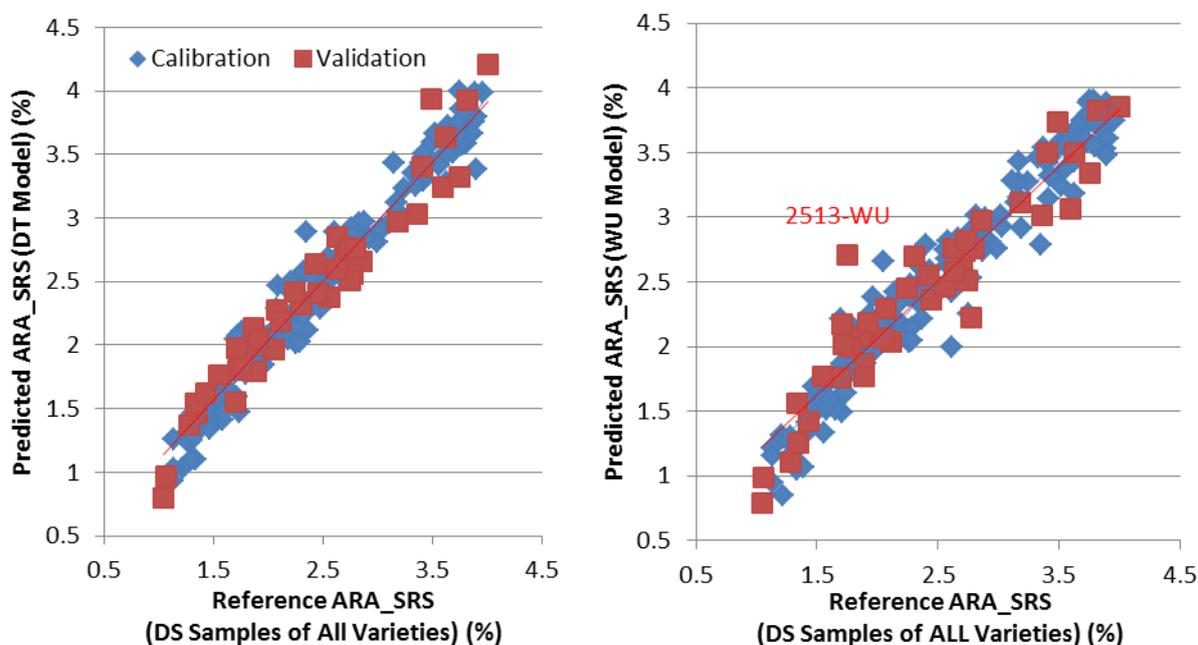


Figure 15-3: Arabinose regression plots. (a) Predicted arabinose vs. reference arabinose for the DT model; (b) the same plot but for the WU model. Refer to Table F-18 for descriptions of these models.

Galactose Content

Table F-19 presents the regression statistics for the galactose models. For all datasets the spectra were transformed with SG-2,2,25,25 prior to model development. This was the only constituent for which extra samples had to be removed from the calibration set. These samples (156 and 236) were both internode sections of the *Miscanthus x giganteus* variety and were clear calibration outliers, for all datasets, in predicted y vs. reference y plots

The RER values for this constituent are much poorer than those for the other sugars discussed so far, with values less than 10 for most datasets. This loss in accuracy and precision for galactose prediction was also seen for the bagasse and peat samples. Interestingly, the RMSEPs and RERs are similar between the various datasets, with very little difference between the DS and WU models for example. It is important to note that the RER_{CV} values are much greater than the RER_{pred} values, with values of over 10 in many cases (including the WU model). This suggests that the samples in the validation set had important spectral/chemical information not present in the calibration set and that the models could be improved with the strategic addition of further samples/spectra. As it stands, the RER_{pred} values obtained suggest that the models developed are only good for sample screening and do not have value in accurate quantitative predictions. However, given that this is a minor constituent and only of small relevance to biorefining processes, an RMSEP of around 0.12% can be considered acceptable. Figure 15-4 (a) provides a predicted vs. reference plot for the All-Varieties DT model and Figure 15-4 (b) the same plot for the WU model.

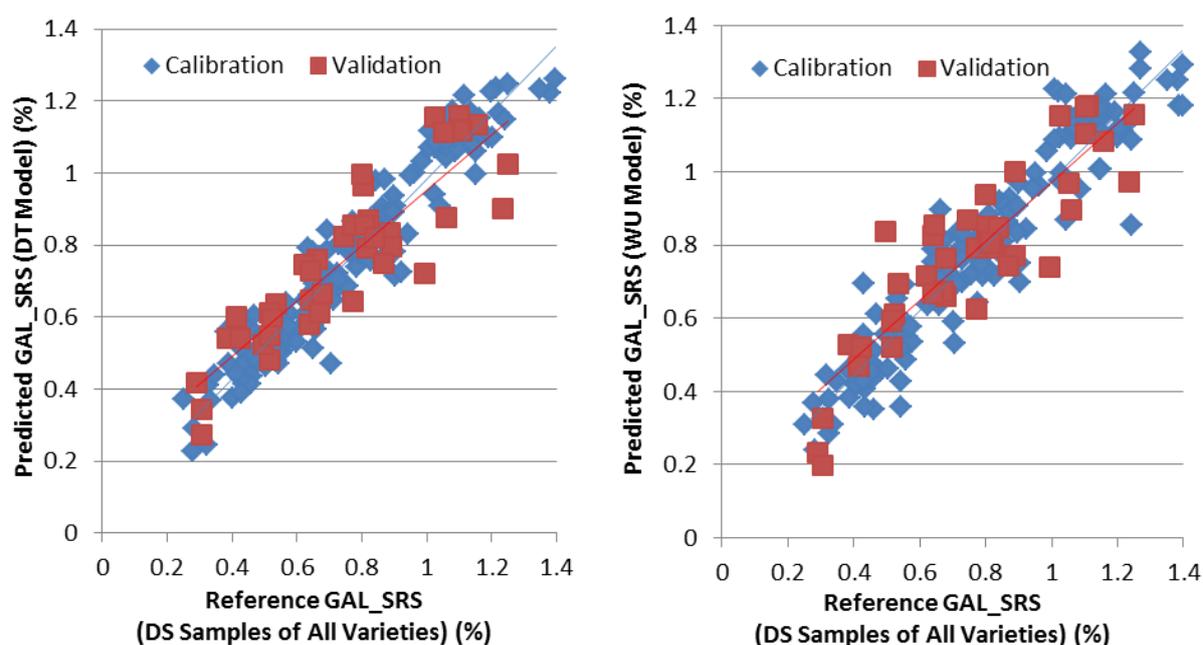


Figure 15-4: Galactose regression plots. (a) Predicted galactose vs. reference galactose for the DT model; (b) the same plot but for the WU model. Refer to Table F-19 for descriptions of these models.

Rhamnose Content

Table F-20 presents the regression statistics for the rhamnose models. For all datasets the spectra were transformed with SG-2,2,25,25 prior to model development.

The RER_{pred} values for this constituent are improved over those seen for galactose, despite this constituent being present in even smaller quantities. RER_{pred} values of over 10 are obtained for all of the models of the DS and DT datasets and for the *giganteus* model of the WU dataset and the All-Varieties model of the DG and DU datasets. The RMSEP values are under 0.05% for most models, indicating that NIRS is able to achieve very low levels of detection even for materials of high moisture content. The majority of samples had rhamnose contents below 0.4% and all of the samples with greater rhamnose contents were either dead leaf blades or live leaf blades. These higher rhamnose content samples tended to have somewhat higher residuals in cross-validation than the other samples.

Figure 15-5 (a) provides a predicted vs. reference plot for the All-Varieties DT model and Figure 15-5 (b) the same plot for the WU model.

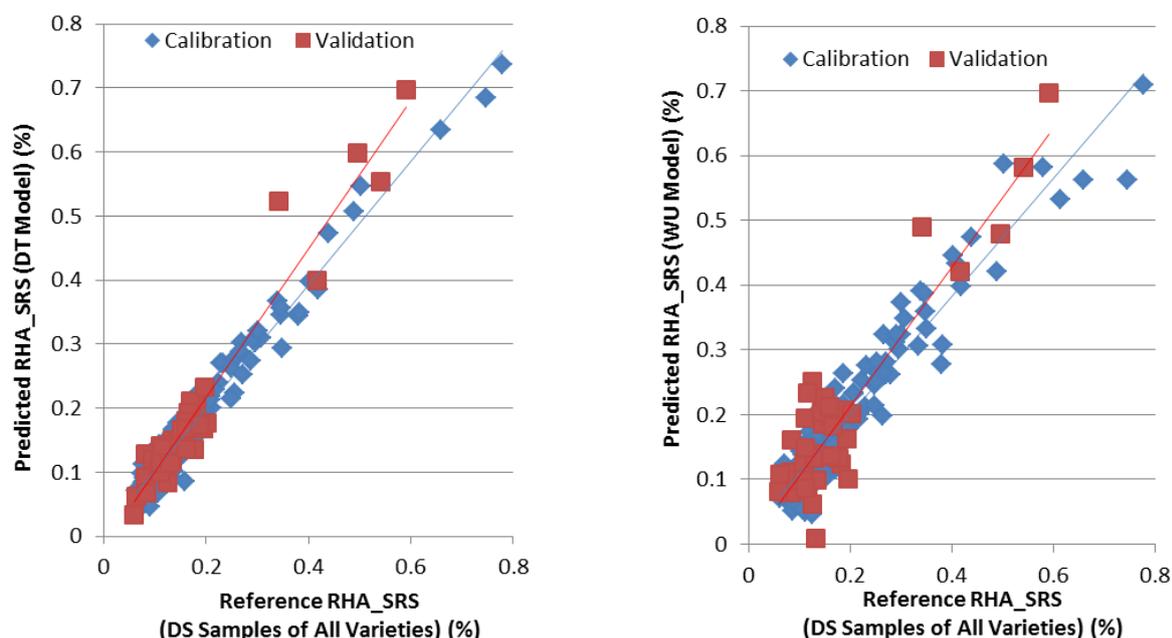


Figure 15-5: Rhamnose regression plots. (a) Predicted rhamnose vs. reference rhamnose for the DT model; (b) the same plot but for the WU model. Refer to Table F-20 for descriptions of these models.

Mannose Content

Table F-21 presents the regression statistics for the mannose models. For all datasets the spectra were transformed with SG-2,2,25,25 prior to model development. It can be seen that the models have higher RMSEPs than the corresponding rhamnose models for each dataset and that the

RER_{pred} values are even lower than for galactose. The issues associated with the chromatographic detection of this low concentration analyte, as discussed in Section 13.3.1, may be responsible to some degree for the poor predictive abilities of these models. Figure 15-6 (a) provides a predicted vs. reference plot for the all-varieties DT model and Figure 15-6 (b) the same plot for the WU model. It can be seen in both these Figures that there are some calibration and validation samples with a 0% reference mannose content but a greater than 0% content predicted according to the NIR model. Interestingly, in contrast to the models for most other constituents, the RER_{pred} values for the DF models are greater than those for many of the other datasets. This is a result of the range in concentration for the DF samples being greater than the DS samples, as illustrated by the histograms and statistics in Table F-3.

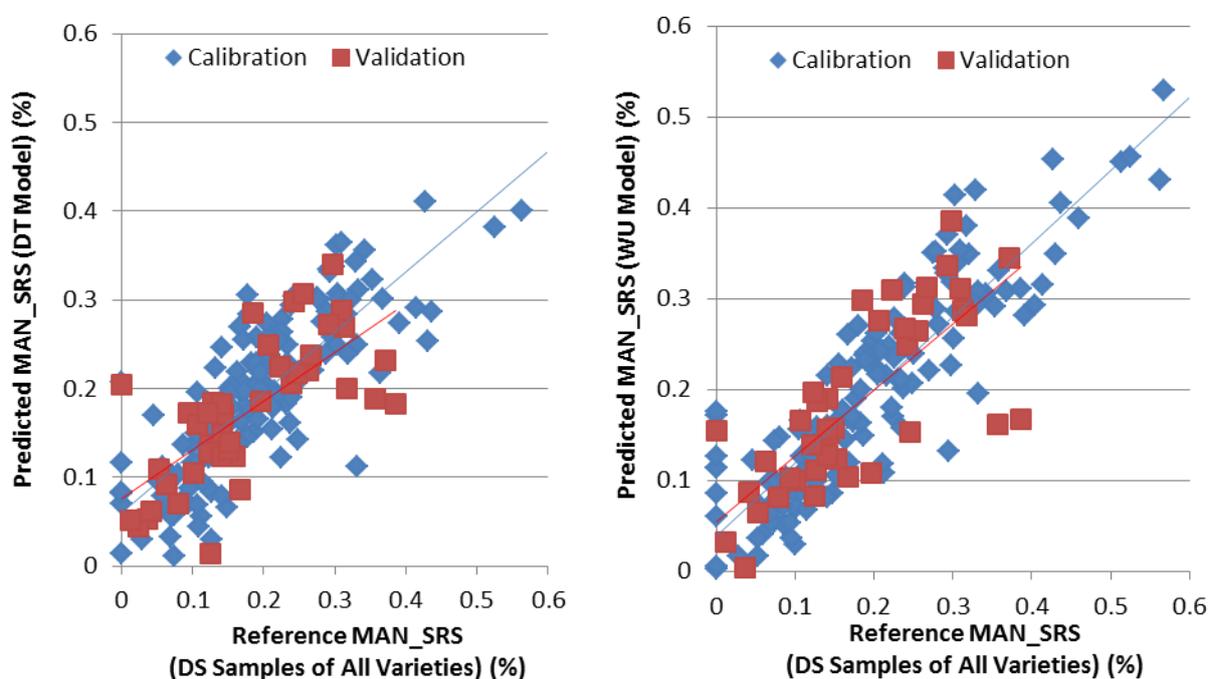


Figure 15-6: Mannose regression plots. (a) Predicted mannose vs. reference mannose for the DT model; (b) the same plot but for the WU model. Refer to Table F-21 for descriptions of these models.

Klason Lignin (KL) Content

Table F-22 presents the regression statistics for the KL models. For all datasets the spectra were transformed with SG-1,1,10,10 prior to model development. There are several interesting observations regarding these KL models:

- RER_{pred} values over 15 for all DS, DT, DU, and DF models and for the *giganteus* models of the WU and DG datasets.
- In contrast to most of the other constituents, the lowest RMSEPs are in the DF models.
- The DU All-Varieties model has a lower RMSEP and higher RER_{pred} than the DS/DT models.

- The *giganteus* models consistently demonstrate lower RMSEPs than the All-Varieties models.
- Despite the high predictive accuracy of the models, a relatively low number of PLS factors are used for most models, particularly when compared against the models for many of the carbohydrates.

Figure 15-7 (a) provides a predicted vs. reference plot for the All-Varieties DT model and Figure 15-7 (b) the same plot for the WU model. As with the arabinose content plot, Figure 15-3, sample 2513 is outlying in the WU model with a y -residual of -3.3%. When this sample was removed from the validation set the RER for the All-Varieties model increased to 13.6 and the RMSEP fell to 0.77%.

Figure F-10 presents a plot for the regression coefficients of an 8 factor PLS model for the KL content of WU samples. The stand-out region of importance in this plot is between 1650 and 1750 nm. This makes sense given the wavelengths quoted in the literature, that are within the 1100-1800 nm region, that are characteristic of lignin. For example, an absorbance at 1672 nm is said to represent the 1st overtone of a C-H stretch in the aromatic rings of lignin (Shenk et al., 2008). Also, an absorbance at 1724 nm has been attributed to the 1st overtone of a C-H stretch in the aliphatic lignin groups (Shenk et al., 2008).

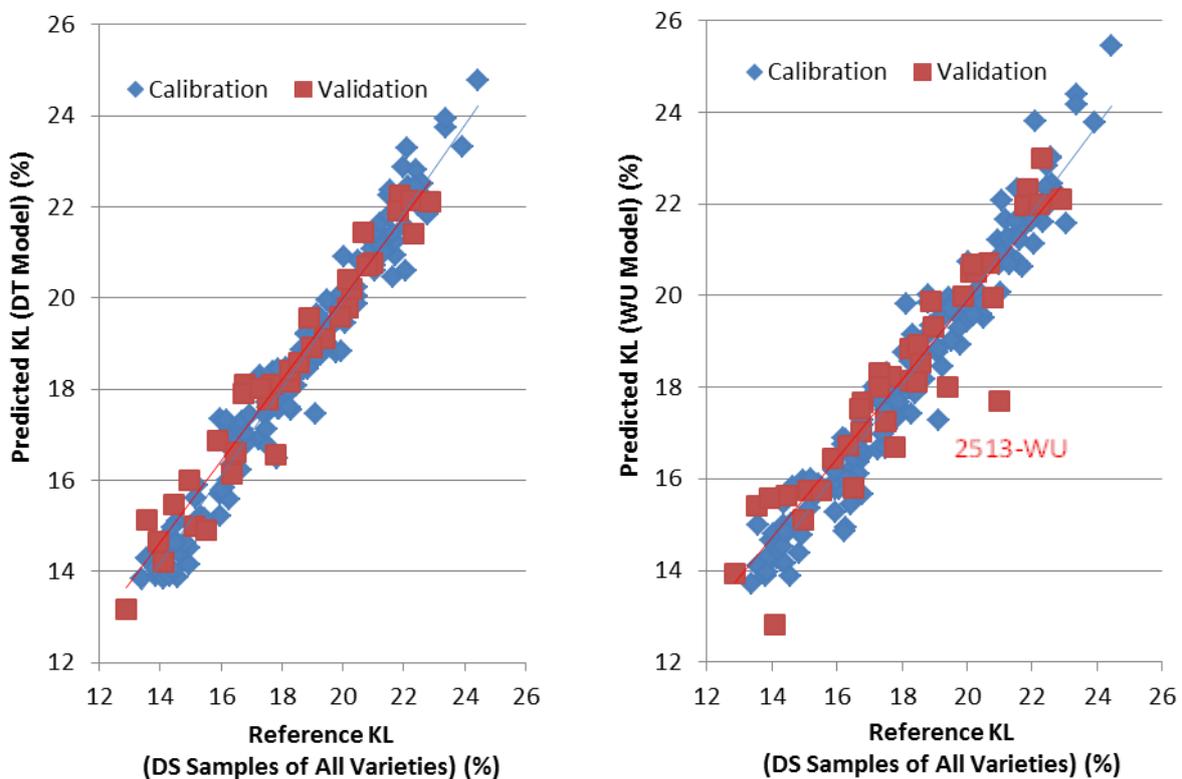


Figure 15-7: Klason lignin (KL) regression plots. (a) Predicted KL vs. reference KL for the DT model; (b) the same plot but for the WU model. Refer to Table F-22 for descriptions of these models.

Acid Soluble Lignin (ASL) Content

Table F-23 presents the regression statistics for the ASL models. For all datasets the spectra were transformed with SG-2,2,25,25 prior to model development. Some points of note regarding these calibrations are listed below:

- RER_{pred} values over 15 for all DS and DT models and for the *giganteus* models of the DU and DG datasets. RER_{pred} values over 10 for all other models except DF-All Varieties.
- The *giganteus* models consistently demonstrate lower RMSEPs than the All-Varieties models. Indeed the RMSEP for the DF *giganteus* model is less than half that of the DF All-Varieties model.
- The RMSEP of the *giganteus* DF model (0.15%) is close to the SEL for these samples (0.12%).

Figure 15-8 (a) provides a predicted vs. reference plot for the All-Varieties DT model and Figure 15-8 (b) the same plot for the WU model.

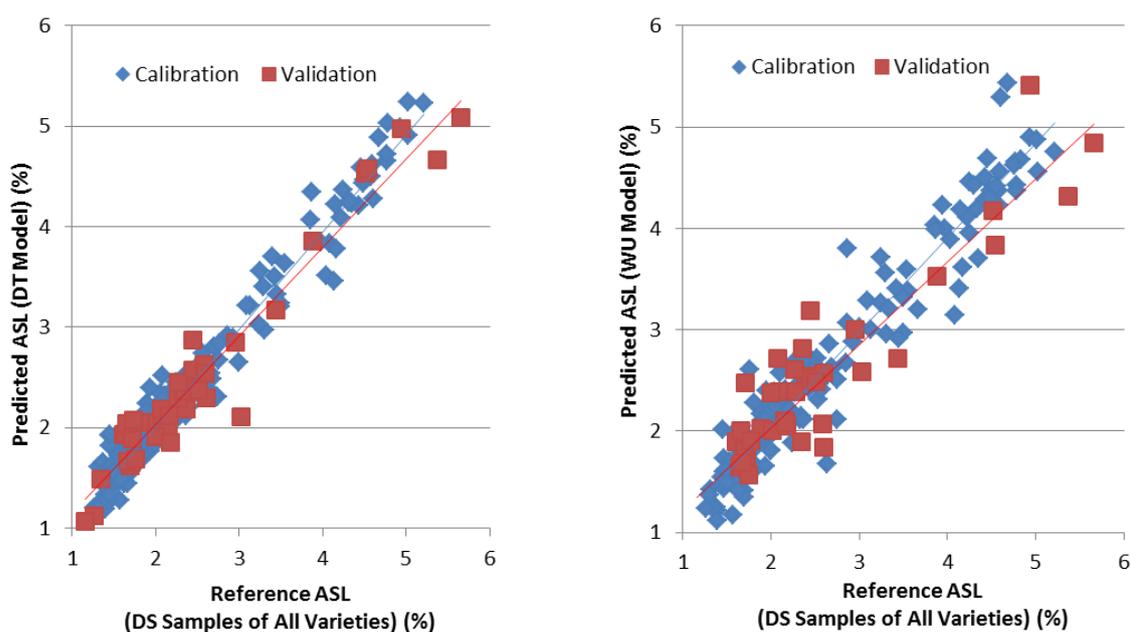


Figure 15-8: Acid soluble lignin (ASL) regression plots. (a) Predicted ASL vs. reference ASL for the DT model; (b) the same plot but for the WU model. Refer to Table F-23 for descriptions of these models.

Uronic Acids Content

Table F-24 presents the regression statistics for the uronic acids (UA) content models. For all datasets the spectra were transformed with SG-2,2,25,25 prior to model development. Given that only 32 samples were analysed for their UA content, cross-validation was used as a means of testing

the models rather than using an independent validation set. The results for these UA models are generally poor, with an RER of over 10 only obtained for the DS model. This needs to be taken in context regarding the low variation seen in the UA contents of the *Miscanthus* samples analysed (see Table F-6). Figure 15-9 (a) provides a predicted vs. reference plot for the all-varieties DS model and Figure 15-9 (b) the same plot for the WU model.

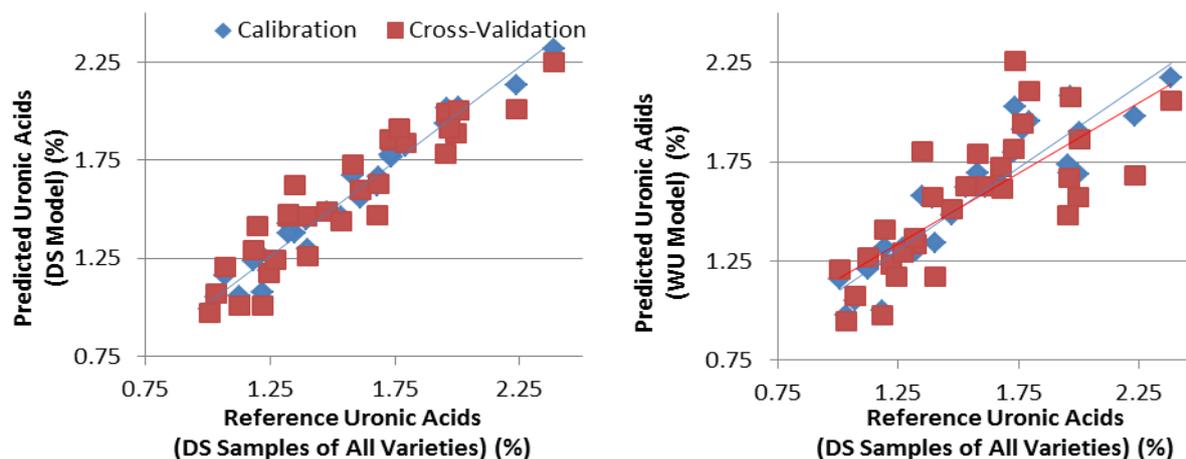


Figure 15-9: Uronic acids (UA) regression plots (a) Predicted UA vs. reference UA content for the DS model; (b) the same plot but for the WU model. Refer to Table F-24 for descriptions of these models.

15.2.2.2 Calibrations for Extractives and Ash Components

Extractives Content

Table F-25 presents regression statistics for models developed for 95% ethanol-soluble extractives content. The spectra were transformed with SG-2,2,25,25 prior to model development. Figure 15-10 (a) provides a predicted vs. reference plot for the all-varieties DT model and Figure 15-10 (b) the same plot for the WU model. Important notes regarding this calibration are listed below:

- With the exception of the DU dataset, models developed on the *giganteus* samples only have lower RMSEPs and higher RER_{pred} values than models developed on All-Varieties.
- RER_{pred} values of over 15 are only possible for the DS and DT *giganteus* models; however, an RER_{pred} value of over 10 was obtained for the WU *giganteus* model. Neither the bagasse or peat WU models achieved RER_{pred} values of over 10.
- The WU models are based on the 1100-2500 nm wavelength region whereas all the WU models for lignocellulosic components only use the 1100-1800 nm region. Incorporating these longer wavelengths allowed for improved RMSECVs.

- Incorporating the 400-1100 nm region in the models for various datasets was tested since it was theorised that including the region where chlorophyll absorbs (673 nm, see Section 14.2) may improve precision; however, that was not the case.

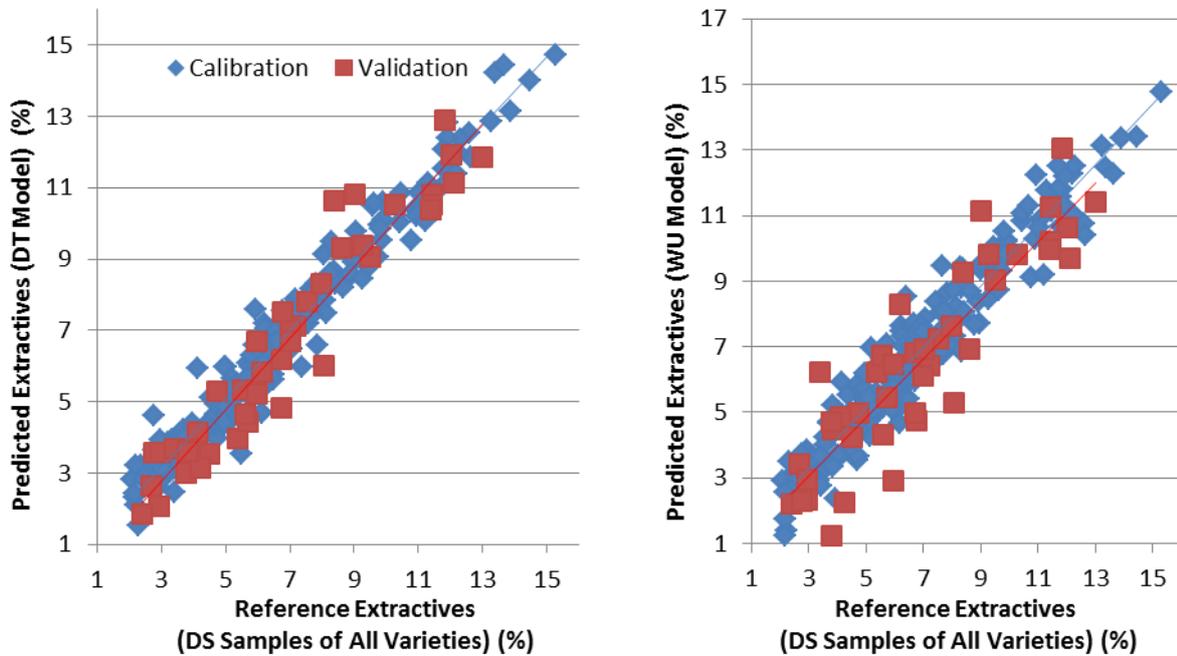


Figure 15-10: 95% ethanol-soluble extractives regression plots. (a) Predicted extractives vs. reference extractives content for the DT model; (b) the same plot but for the WU model. Refer to Table F-25 for descriptions of these models.

Ash Content

Table F-26 presents the regression statistics for the ash content models. For all datasets the spectra were transformed with SG-2,2,25,25 prior to model development. Figure 15-11 (a) provides a predicted vs. reference plot for the all-varieties DT model and Figure 15-11 (b) the same plot for the WU model. Important notes regarding these calibrations are listed below:

- RER_{pred} values of over 15 are possible for all DS and DT models, for the All-Varieties DG scans model, and for the *giganteus*-DF model. In all other models the RER_{pred} values are over 10, except for the WU-*giganteus* model (the RER value is 9.99).
- With the exception of the DG dataset, models developed on *giganteus* samples only have lower RMSEPs than models developed all All-Varieties.
- The WU models use the 1100-2500 nm region.

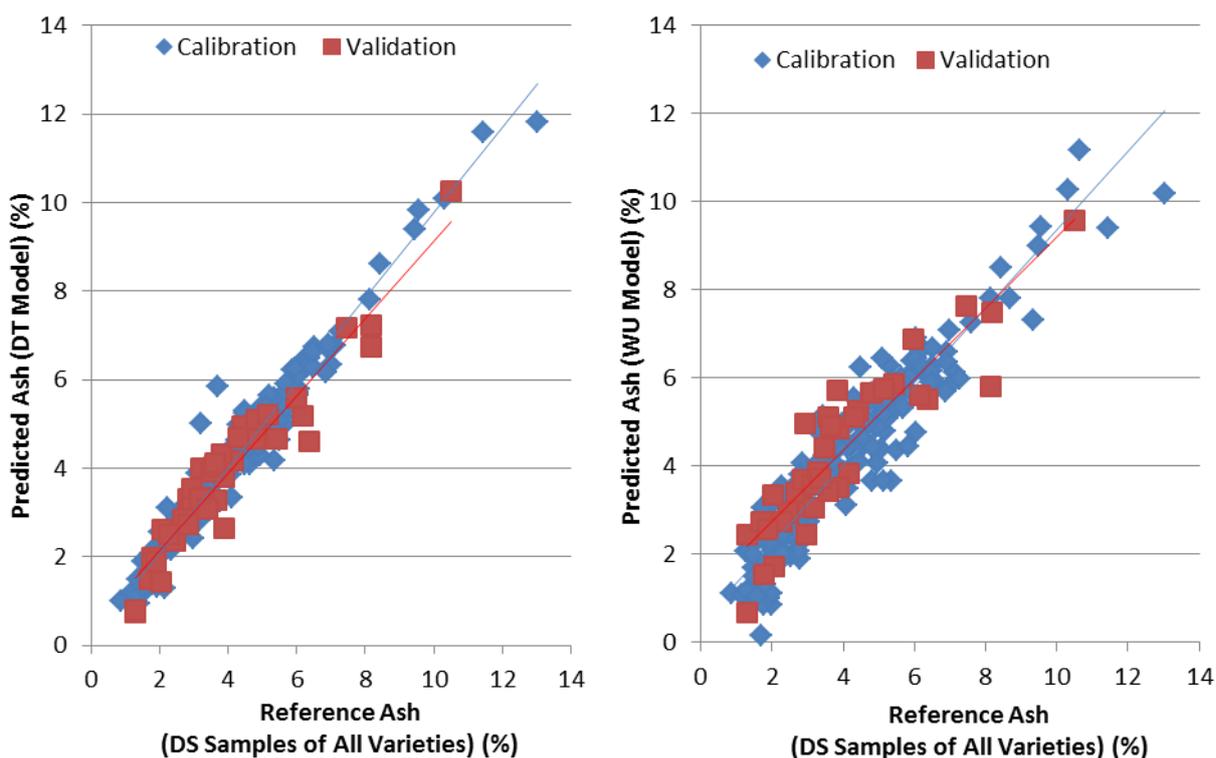


Figure 15-11: Ash regression plots. (a) Predicted ash vs. reference ash content for the DT model; (b) the same plot but for the WU model. Refer to Table F-26 for descriptions of these models.

Acid Insoluble Residue (AIR) Content

Table F-27 presents the regression statistics for the AIR models. For all datasets the spectra were transformed with SG-1,1,10,10 prior to model development. Figure 15-12 (a) provides a predicted vs. reference plot for the all-varieties DT model and Figure 15-12 (b) the same plot for the WU model. Important notes regarding these calibrations are listed below:

- With the exception of the DG All-Varieties model, the RMSEPs are greater than those for the corresponding models for KL (the RMSEP for the DF All-Varieties model is almost double that of the KL model).
- With the exception of the DG dataset, models developed on the *giganteus* samples only have lower RMSEPs than models developed all All-Varieties.
- RER_{pred} values of over 15 are only possible for the DS, DT, and DU models based on *giganteus* samples. The WU *giganteus* model has an RER_{pred} of over 10 but the value for the All-Varieties WU model is less than 10.
- The WU models use the 1100-2500 nm region.

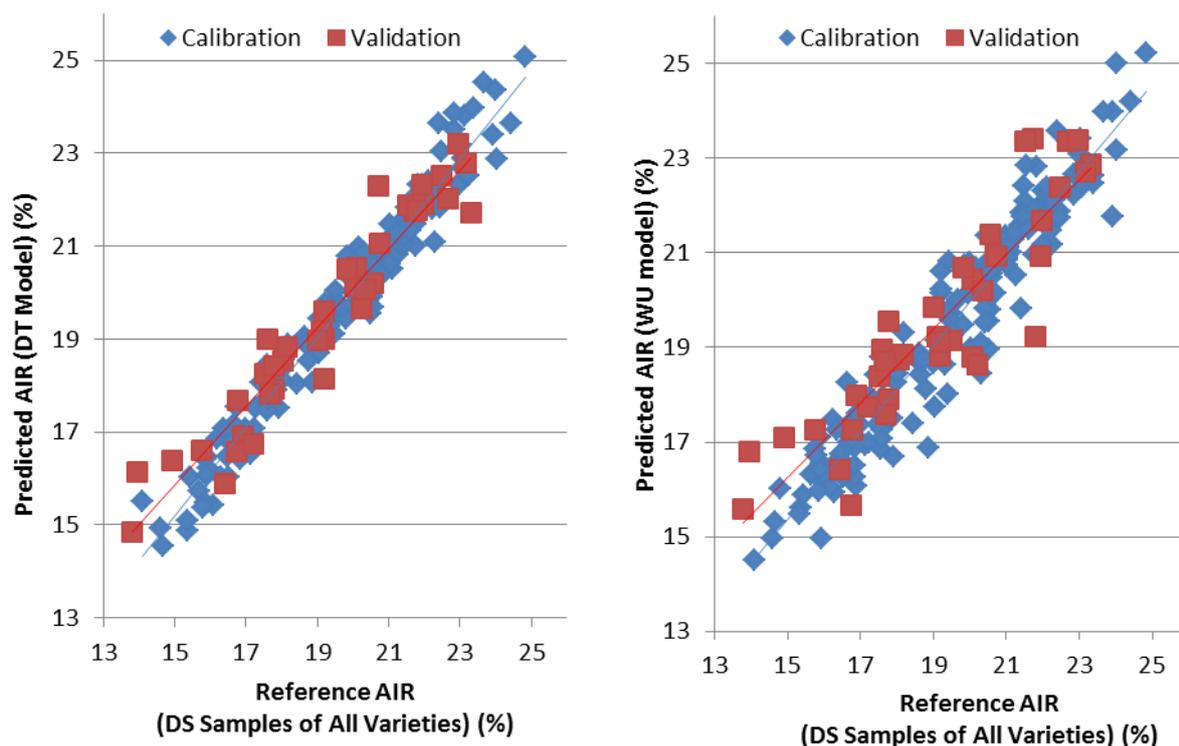


Figure 15-12: Acid insoluble residue (AIR) regression plots. (a) Predicted AIR vs. reference AIR content for the DT model; (b) the same plot but for the WU model. Refer to Table F-27 for descriptions of these models.

Acid Insoluble Ash (AIA) Content

Table F-28 presents the regression statistics for the AIA models. For all datasets the spectra were transformed with SG-1,1,10,10 prior to model development. Figure 15-13 (a) provides a predicted vs. reference plot for the all-varieties DT model and Figure 15-13 (b) the same plot for the WU model. Important notes regarding these calibrations are listed below:

- RER_{pred} values of over 15 are only possible for the DS and DT models based on *giganteus* samples and for the All-Varieties DG model.
- The RER_{pred} values for the WU models are over 4 but less than 10, suggesting the calibrations only have use in sample screening. These models use the 1100-2500 nm region.
- With the exception of the DU dataset, models developed on the *giganteus* samples only have lower RMSEPs than models developed on all All-Varieties.
- Figure 15-13 (a) and Figure 15-13 (b) show that the samples selected for the validation set do not span the whole concentration range. Improved models could result from filling the gaps in the concentration range with additional samples for model development and validation.

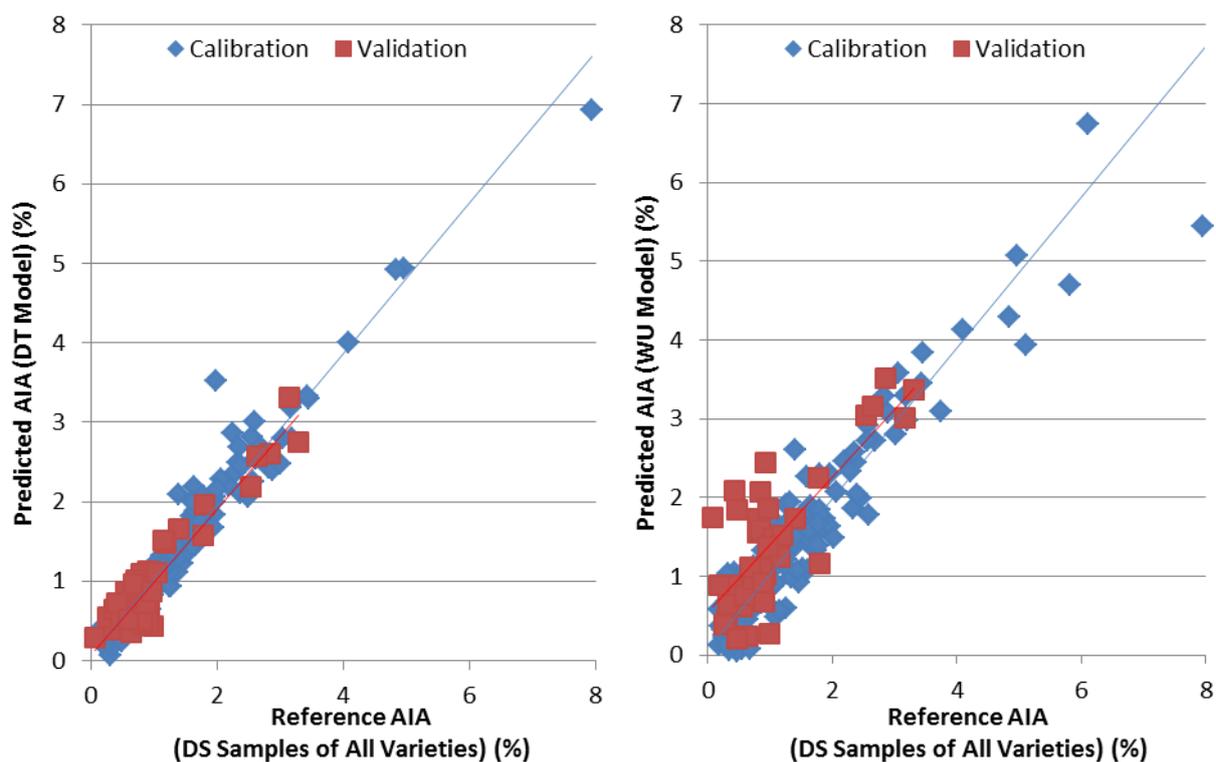


Figure 15-13: Acid insoluble ash (AIA) regression plots. (a) Predicted AIA vs. reference AIA content for the DT model; (b) the same plot but for the WU model. Refer to Table F-28 for descriptions of these models.

15.2.2.3 Elemental Analysis

Carbon Content

Table F-29 presents the regression statistics for the calibrations developed for carbon content. For all datasets the spectra were transformed with SG-2,2,25,25 prior to model development. It can be seen that the R_{pred}^2 and R_{CV}^2 values for this element are much poorer than for most of the other constituents presented in previous Tables. This is reflected in the relatively high standard errors and low RPD and RER values. No regression statistics are provided for the WU dataset since no models that provided R_{CV}^2 values over 0.5 could be found. Numerous spectral pretreatments and wavelength ranges other than those listed in Table F-29 were tested for the various models but none achieved a significant improvement in accuracy or precision. Figure 15-14 provides a predicted vs. reference plot for the all-varieties DT model.

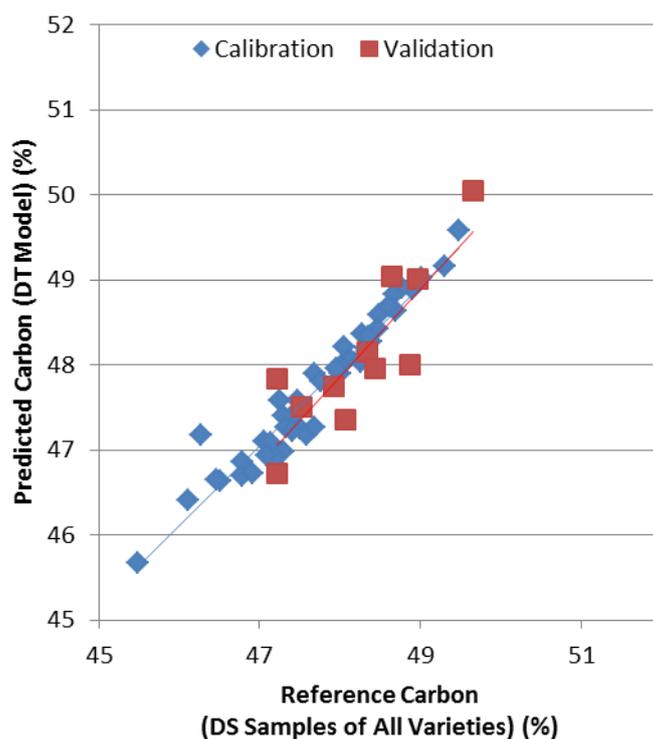


Figure 15-14: Predicted carbon vs. reference carbon content for the DT model. Refer to Table F-29 for descriptions of these models.

Nitrogen Content

Table F-29 presents the regression statistics for the nitrogen content models developed on samples comprising different *Miscanthus* varieties. For all datasets the spectra were transformed with SG-2,2,25,25 before PLSR. Important notes regarding these calibrations are listed below:

- Excellent RER_{pred} values of over 15 are possible for the DS, DT, and DG models, and with an RER_{CV} of over 15 possible for the DF dataset when all the samples are in the calibration. Figure 15-15 (a) provides a predicted vs. reference plot for the all-varieties DT model.
- The WU model gave better predictions when the 1100-2500 nm region was used compared with just the 1100-1800 nm region; however, even when all the samples were in the calibration set the RER was less than 10.
- It was considered that the relatively poor performance of the WU model could arise because more samples are needed to effectively model this highly heterogeneous and complex dataset than would be required for the DT scans.

- The DT model developed on all DT spectra with available nitrogen data was used to predict the nitrogen content of DT samples that had not been analysed for N. These predicted N contents were then linked to the corresponding WU scan of the same sample to form a new calibration set, WU_{pred} in Table F-29. The WU scans for which “real” N data existed were excluded from this calibration and formed the independent validation set.
- The WU_{pred} model has a lower RMSEP and higher RER; however, Figure 15-15 (b) shows in a predicted vs. reference plot for this model that the slope of the regression for validation is much lower than that of the calibration set. This may be because more leaf samples of the highest N contents were in the validation set than the calibration set.
- The improvement of the WU_{pred} model over the WU model suggests that good NIR prediction of nitrogen content is possible for wet and heterogeneous samples; however, ideally more samples with high nitrogen contents should be collected and analysed in order to produce a more accurate model based on real reference analytical data.

Figure F-11 presents a regression coefficients plot for the 5 factor nitrogen DT model where all samples are in the calibration set. It is a complex plot with many peaks. Regions of the plot of importance that match those presented in the literature include the region around 2174 nm which is associated with a strong N-H protein absorption feature, particularly in leaves (Martin and Aber, 1994, McLellan et al., 1991).

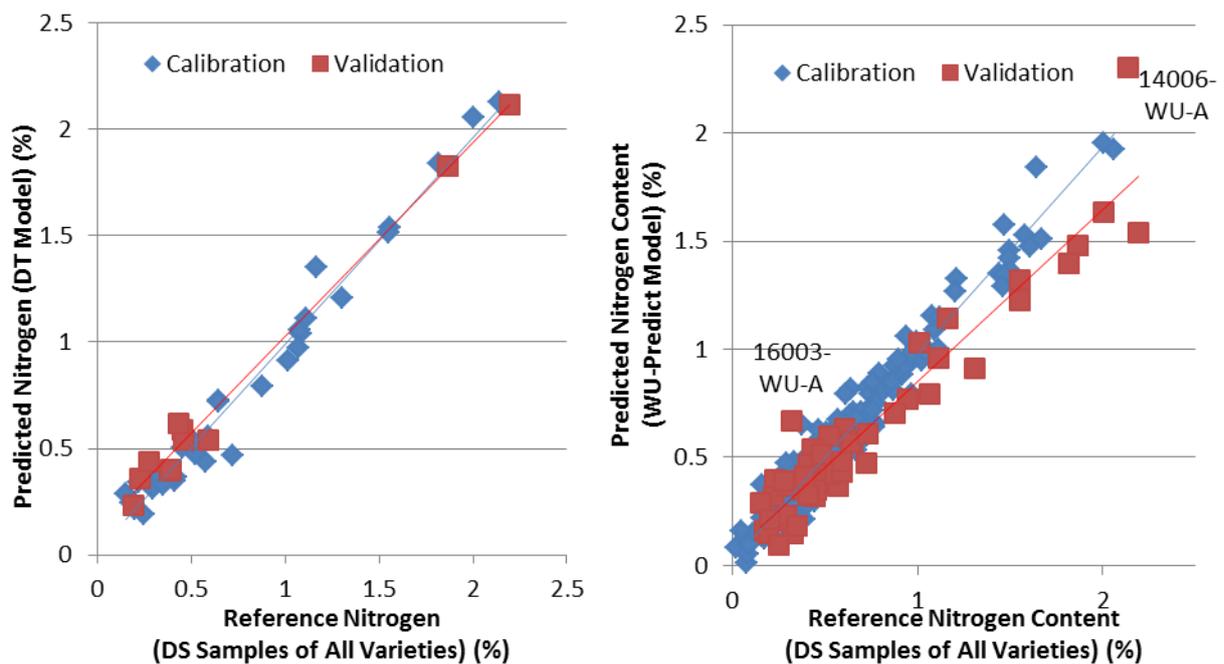


Figure 15-15: Nitrogen regression plots. (a) Predicted nitrogen vs. reference nitrogen content for the DT model; (b) the same plot but for the WU_{pred} model. Refer to Table F-29 for descriptions of these models.

Hydrogen and Sulphur Contents

No regression statistics are presented for the models developed on the various datasets for the hydrogen and sulphur contents. That is because, despite using various combinations of spectral pretreatments and wavelength regions for regression, no models with R_{CV}^2 value of over 0.5 could be developed. This is most likely due to the minimal variation seen in the hydrogen and sulphur contents of these samples, as discussed in Section 13.3.1.

15.2.2.4 Models Based on the DF Spectra

The regression statistics for the models developed on the DF samples are provided in Table F-15 to Table F-29. Much fewer DF samples were analysed compared with the DS samples (there are total sugars data for 52 DF samples and 191 DS samples, for example).

A targeted sample selection strategy was employed so that the DF samples that were analysed had a wide variety in concentrations and also covered all of the different fractions of the plant. Two preliminary DS models, one for the glucose content and the other for the KL content, were used to predict the glucose and KL compositions of the DF samples using the DF spectra. These predictions were sorted in ascending order for each constituent and samples were manually chosen along the concentration range ensuring that samples of each plant fraction were included in the dataset and that the DF samples that were selected had sufficient material to allow for replicate analysis.

The samples chosen were from the following plant fractions: Dead leaf blades (7 samples); dead leaf sheaths (5 samples); live leaf blades (7 samples); internodes (13 samples); nodes (10 samples); stem sections (9 samples). There was insufficient DF material of live leaf sheaths to allow for their analysis. This global set covered several *Miscanthus* varieties, as outlined in Table F-1. Random sample selection was used to select the 13 validation samples from this set with the remainder of the samples comprising the calibration set. Models were also developed on just the *Miscanthus x giganteus* samples with random selection again used to select the 11 samples for the validation set from the larger set (44 samples with sugars data). For the selection of the 51 samples for elemental analysis different criteria were employed; there was a greater focus on the analysis of leaf sections since these demonstrated the greatest variation in nitrogen content.

Since there were much fewer samples with DF data, models were also developed where all samples were present in the calibration set and the models were tested through cross validation. The statistics for these tests are placed in the lower parts of Table F-15 to Table F-29. In all DF models an upper limit of 12 PLS factors was set, with Haaland's criterion selecting the optimum number of factors within this range.

For many constituents the RMSEP values for the All-Varieties and *giganteus*-only DF models are greater than those for the DS/DT models. Indeed, in the case of the glucose, xylose, galactose, mannose, KL, and ash models the RMSEPs for the WU models were superior to those of the DF models. The relative poor performances of the DF models are most likely a result of the limited number of samples provided to the models. In nearly all cases the RMSECVs for models where all samples were in the calibration set were lower than the RMSEPs. The RER values were also highly sensitive to which samples were randomly selected for validation. Taking the example of the xylose calibration on all *Miscanthus* varieties, Table F-17, the RER_{CV} is 12.1 while the RER_{pred} is 4.7, despite the standard errors for both tests being similar (0.6%). The large difference in the RER value is a result of the samples in the validation set covering a much more limited region of the xylose concentration space than the calibration samples. It is therefore better to compare the DF models against the other models by using the standard errors rather than range/standard-deviation based criteria.

Table F-30 and Table F-31 present regression statistics for experiments where the DS/DT All-Varieties models listed in Table F-15 to Table F-28 for the various constituents are used to predict the compositions of the DF samples. Two scenarios are tested. In the first the number of PLS factors as determined by Haaland's criterion for these DS/DT models are used to predict the DF samples. In the second the R^2_{pred} values obtained on using these DS/DT models to predict the DF samples were examined over factors 1 to 20 and the factor which provided the greatest value was selected.

An optimal result in the prediction of these DF sample using the DS/DT models would be that there was a high degree of accuracy and precision in the analysis. If only the R^2_{pred} values in Table F-30 and Table F-31 are observed it would appear that the DS/DT models fit the DF data well, given that the R^2_{pred} values are often greater than the corresponding R^2_{pred} values for when the DF models are tested on their own independent datasets. However, while there does appear to be a close relationship between the predictions and real DF data, there is a significant bias in many of the models and the slope of the regression curve is typically much less than one.

Figure F-12 (a) presents a plot of predicted glucose content using the DS model (with the number of PLS factors selected according to Haaland's criterion) versus the reference analytical data for the DF sample. A regression line is plotted, as is a line representing a 1:1 relationship (the black line). It can be seen that the model is overestimating the glucose content of the DF samples, particularly at lower glucose concentrations. Figure F-12 (c) shows how the predicted glucose minus actual glucose y residuals vary according to the glucose content of the DF sample. Figure F-12 (b) presents a plot where each point represents a sample, its location on the x-axis being provided by the reference analytical glucose data for its DF fraction, and its location on the y-axis being provided by the reference analytical glucose data for its DS fraction. The slope of the trend line fitted to this correlation plot is much closer to 1 and with a much smaller bias than the plot of the NIR-predicted data vs. the reference DF data. Furthermore, in contrast to Figure F-12 (a), the greatest y minus x residuals appear to be present at higher glucose concentrations for the DF samples. At these concentrations the DS fractions tend to exhibit greater glucose contents than the DF fractions.

There are two possible explanations for the relationships seen for glucose:

- The NIR predictions are correct and there are errors in the reference analysis. As discussed in Section 3.1.3 the glucose liberated in hydrolysis may become degraded to a greater degree for samples of a very fine particle size. In order to test this theory the hydrolysates will need to be analysed for their furfural, hydroxymethylfurfural, and levulinic acid concentrations. If these are greater, as a proportion of the total glucose content, than those of the DS samples then the hypothesis may be correct. The Author plans to test this theory once a suitable analytical method for these degradation products has been developed on the HPAEC system. However, while this theory may explain a bias in predictions it does not explain the trend seen with decreasing y -residuals with increasing glucose content unless there is a link between glucose content and mean particle size of the DF fraction. This is possible if it is considered that the samples with higher glucose contents tend to be the more lignified fractions of the plant which were typically more resistant to comminution compared with the leaf fractions (which tended to have lower glucose contents).
- The DS/DT models are not suitable for accurate prediction of DF samples without a bias correction. Table F-30 shows that the RMSEP for glucose is 2.14% but the SEP, which involves a correction for the bias, is more reasonable at 1.58%. Adjusting the slope of the model may also provide improved predictive abilities.

Another constituent that demonstrated a large bias in the DS/DF regressions was the 95% ethanol-soluble extractives content. Figure F-13 (a) presents a plot of predicted extractives content using the DS model versus the reference analytical data for the DF sample. Figure F-13 (b) shows the DS extractives content vs. the DF extractives content and Figure F-13 (c) shows how the prediction y-residuals vary according to DF extractives content. In this case the NIR model is underestimating the extractives content, i.e. the bias value in Table F-31 is negative. Interestingly, Figure F-13 (a) and Figure F-13 (b) are much more similar to each other than Figure F-12 (a) and Figure F-12 (b) are. As discussed in Section 3.8, the efficiency of solvent extraction has been negatively correlated with particle size, which may be an explanation for the trends here. Also, it makes sense that the more easily broken material would have a higher extractives content. If this is the case, however, the NIR predictions do not seem to be reflecting it. It is possible that different types of extractives components are present in the finer particles than those present in the DS fractions, meaning that the DS calibrations will not have been trained on these specific extractives.

There was also a bias present in the DT model predictions of KL content. Figure F-14 (a) presents a plot of predicted KL content using the DT model versus the reference analytical data for the DF sample. Figure F-14 (b) shows the DS KL content vs. the DF KL content. The relationship is reasonably close to 1:1 with a slight shift towards higher KL contents for the DS samples. Figure F-14 (c) shows how the prediction y-residuals vary according to DF KL content. The correlation between y-residuals and the reference y-value is strongest for this constituent ($R^2 = 0.861$) with the regression line crossing from negative to positive residuals around 21% KL. Observing both Figure F-14 (a) and Figure F-14 (b), it is clear that the NIR model is predicting that the DF samples have higher KL values than their corresponding DS samples which seems counter-intuitive given that larger particle might be expected to be more fibrous. It may be the case that the regression coefficients for the DS model, while finely tuned for accurate prediction of DS samples, are not entirely suitable for accurate prediction of DF samples, although the precision (as demonstrated by the reduced SEP) is good.

Figure F-15 presents regression coefficients, over 1100-2500 nm, for the DF, DT, and DS KL models. Each model has 6 PLS factors and was developed on spectra pretreated by SG-1,1,10,10. The area around 1662 nm is of importance in all models, and is likely to represent the 1st overtone of a C-H stretch in the aromatic rings of lignin (Shenk et al., 2008). The largest absolute regression coefficient values are around 2260 nm, a region that has been identified as a lignin absorption band (Martin and Aber, 1994). The DT and DS regression coefficients plots follow each other very closely; however, the shape of the DF plot does differ in some areas as does the relative intensity of the major peaks. This

could explain the close correlation between the DS/DT predictions and the DF KL contents, as well as the fact that the relationship is not one to one.

Figure F-16 presents the regression coefficients, over the 1100-2500 nm region, for the DF, DT and DS glucose (GLU) models. Each model has 5 PLS factors in this instance and the models were developed on spectra pretreated by SG-1,1,10,10. As with the bagasse and peat regression coefficient plots there are peaks around 1450 nm, a region linked to the 1st overtone of O-H stretches in crystalline and semi-crystalline cellulose (Tsuchikawa and Siesler, 2003b, Tsuchikawa and Siesler, 2003a). Also, the largest absolute regression coefficient values occur around 2100 nm and 2326nm, which are characteristic absorption region for cellulose (Üner et al., 2011, Martin and Aber, 1994). As with the KL plot, the DF regression coefficients differ more greatly from the DS/DT coefficients than the DS and DT coefficients differ from each other.

15.2.2.5 Combination of DS and DF Data

Table F-32 presents the results for models developed for the glucose, xylose, acid soluble lignin, and KL contents of the WU scans using either the corresponding analytical data for the DS fraction of each sample (the standard WU model) or using the weighted average of the analytical data for the DS and DF fractions of each sample (the WU_{DG} model). The RMSECVs and other regression statistics are poorer in these models, compared with the models in Table F-15 to Table F-28, due to the much reduced number of samples available for model development. However, what is important is the relative difference between the two model types and whether including the DF data helps to build a more precise model. The data in Table F-32 show that, for every constituent, the RMSECV/SEC V values are lower and the RER values are higher for the WU_{DG} model. That suggests that improved regression statistics over those seen in Table F-15 to Table F-28 may be possible by including the proportionate influence of the DF fraction to the model. More importantly, as well as being more precise, the predicted values are likely to be more accurate regarding the real composition of the samples being scanned in the WU state.

The next step involved testing whether or not these weighted models were more accurate. This required more samples, however. Using the DF calibrations listed in Table F-15 to Table F-29, the DF compositions of the DF samples that had not been analysed by reference methods were predicted in order to create a new weighted average composition incorporating these and the real reference analytical DS data. The models developed on these data were termed WU_{DGP} and compared against

the standard WU models that are based on the DS data only, as discussed in Section 15.1.1. Regarding the calibrations for each constituent, the same number of samples were shared between the calibration sets for each model type. The samples used in the models in Table F-32 were excluded from the calibrations and instead formed the independent validation set. Since the data for these samples incorporated the “real” DF results from reference analysis, this set would test the suitability of the prediction of the unknown DF samples, as well as the accuracy of the standard WU and WU_{DGP} models for predicting the composition of the WU scans.

The regression statistics for these two types of models are presented in Table F-33 and Table F-34. It can be seen that, for all of the constituents except AIR, the RMSEP for the WU_{DGP} model is less than that for the WU model. In accordance with the relationships observed for when the DS/DT models predicted the compositions of the DF samples, Table F-30 and Table F-31, there is a significant bias for some of the WU models. These include a positive bias for glucose and total sugars and a negative bias for extractives. The WU_{DGP} models also have a bias for some constituents, however, with a negative bias for glucose and total sugars and a positive bias for extractives. As discussed in Section 15.2.2.4, the DF models, since they are only developed on a limited number of Miscanthus samples, are not fully refined. An ideal WU_{DGP} model would involve the use of DF calibrations developed on a greater number and variety of samples.

The RER values in Table F-33 and Table F-34 are dependent on the concentration distribution, for each constituent, of the samples in the validation set. The range in concentrations in this set are often less than the samples in the validation sets used for the models in Table F-15 to Table F-29. However, the RMSEPs and SEPs for the WU_{DGP} models can be compared with those for the WU models in Table F-15 to Table F-29. The WU_{DGP} models have lower RMSEPs for glucose, rhamnose, KL, extractives, ash, and AIA.

All DF scans taken as part of this research were taken from samples in the coarse sample cell. For this reason some DF samples could not be scanned since they would not cover the entirety of the cell window. However, given that the repack error of DF samples has been demonstrated to be low (see Section 15.2.3) the collection of DF spectra using a much smaller NIR cell appears feasible. This would allow for more DF samples to be scanned and their predicted compositions combined with the reference analytical data for the DS fraction, so expanding the size of the WU_{DGP} set. Coupled with refinement of the DF calibrations this could result in further improved models, in terms of precision and accuracy, than those seen in Table F-15 to Table F-29.

15.2.2.6 Moisture Contents

Moisture Content of WU Samples

The first attempt at developing a calibration for the moisture content (MC) of wet and unground (WU) samples took place using the scans and moisture contents that were taken after the internode and node sections had been chipped. These were used to develop separate NIR models for each of these fractions. These MCs were determined by weighing approximately 2-5 g of the WU material (depending on the density) into a 50 ml crucible. This was repeated so that each sample had a duplicate and the crucibles were then put in an oven at 105°C and weighed the following morning.

Figure F-17 presents a histogram of the distribution of MCs for the samples available for selection for the calibration and validation sets of the internode model. There is a wide range in MC for these samples with a close to normal distribution that has a negative skew. Table F-35 presents the regression statistics for this internodes model and Figure 15-16 (a) shows a predicted MC vs. reference MC plot. An RMSEP of 3.44% and an RER_{pred} of 15.46 was obtained.

Figure F-18 presents a histogram, with associated statistics, of the distribution of MCs for the node samples. It shows that there is less variation among these samples; that most nodes have MCs of between 55 and 70%; and that the nodes have a higher average MC content than the internode samples. Table F-35 presents the regression statistics for the nodes model and Figure 15-16 (b) shows a predicted MC vs. reference MC plot. This model provided a lower RMSEP than the internodes model (1.39%) as well as a higher RER_{pred} (18.04). In contrast to the other MC models, which used SG-1,1,10,10 as a spectral pretreatment, the optimal RMSECV values for the node model were based on an SNVDT transform of each spectrum.

The MC measurement method described above was more suitable for node sections since a representative subsample of these could easily fit inside the crucibles. However, it was not considered ideal for developing accurate MC calibrations for internode and leaf sections. This is because these often had larger fragments (see Figure 14-3) that were scanned in the NIR cell, and so contributed to the spectra collected, but were too large to fit inside the crucibles.

A second method was then tested for the development of NIRS calibrations for the MC of leaves and internode/whole-stem sections. This involved scanning the WU subsample inside the NIR coarse-cell and then placing the whole contents of this cell into an aluminium foil container with a lid that incorporated holes to allow for the moisture to vent when the container was placed in the oven to dry overnight. This was repeated for a second scan and second container meaning that the results

for each container could be associated with one particular scan. To avoid overfitting of the model only one of these scans could be included in global data set for selecting samples for the calibration and validation sets. Also, to check for errors in the reference analysis, only samples for which the difference in analysed MCs for the two containers was less than 1.5% could be included in this global set. Figure F-19 presents a histogram for the distribution of MCs across these samples. It can be seen that it is somewhat bimodal, the lower MC peak is mostly associated with samples of dead leaf blades and dead leaf sheaths while the higher MC peak is associated with internode/stem section and live leaf samples.

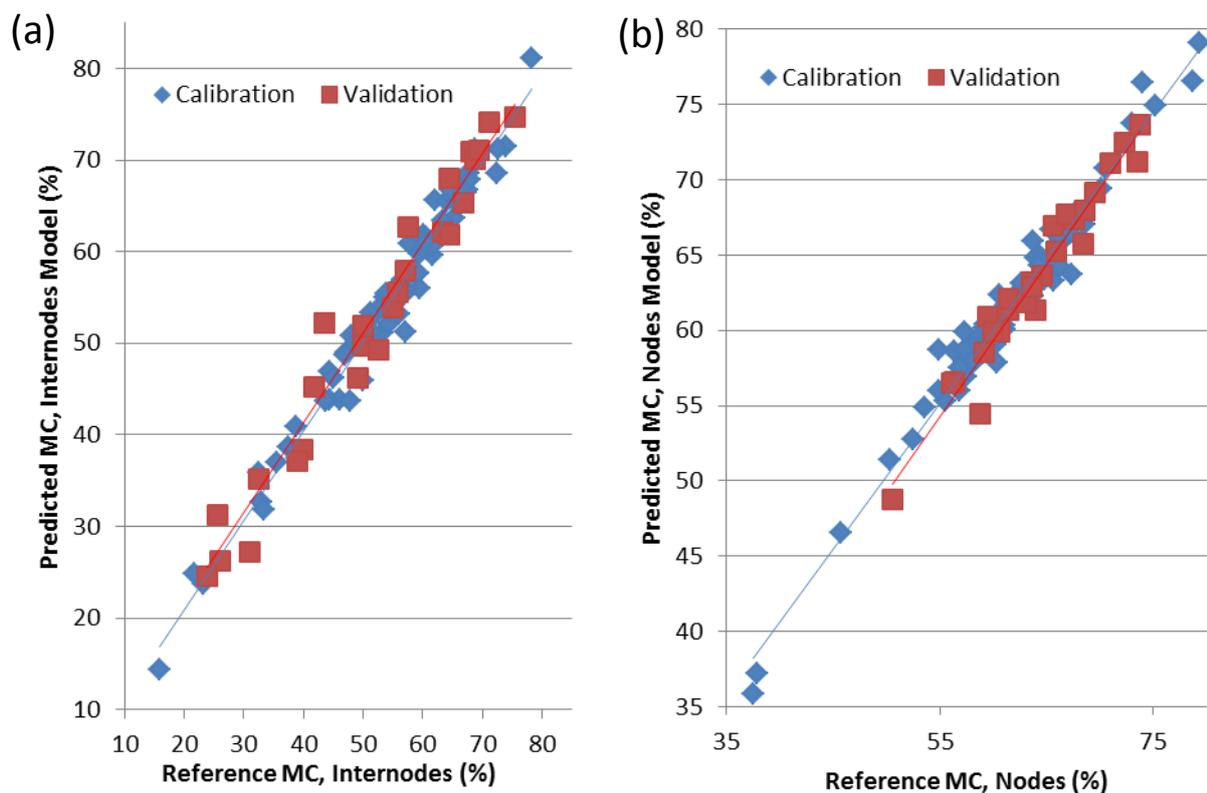


Figure 15-16: Predicted wet-basis moisture content (MC) vs. reference MC regression plots for: (a) the model based on internode samples; (b) the model based on node samples.

Table F-35 presents the regression statistics for this Stems and Leaves model and Figure 15-17 shows a predicted MC vs. reference MC plot. The RMSEP of 2.53% represents a significant improvement over the RMSEP value for the internodes model. Furthermore, this is the only MC model that achieved an RER_{pred} value of over 20. Since it has been developed on stems and leaves it is suitable for the prediction of a wide range of *Miscanthus* samples for which reference MC data do not exist. If separate node samples are to be predicted then the Node model remains the most appropriate calibration for these.

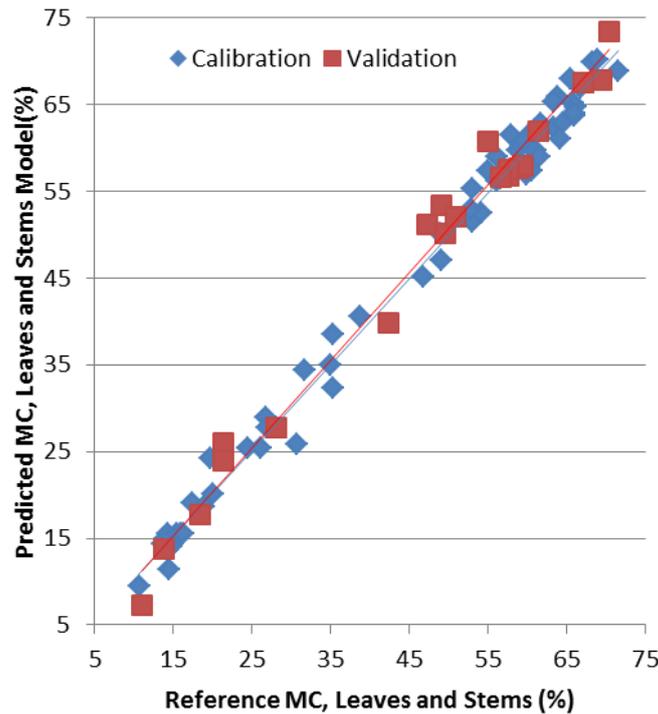


Figure 15-17: Predicted wet-basis moisture content (MC) vs. reference MC for the model based on samples of leaves and stems.

Moisture Contents for Reference Analytical Methods

As with the bagasse samples (see Section 12.3.3.4.3) NIR calibrations were developed for the moisture contents prior to: the removal of ethanol extractives (DS-E); the determinations of the moisture contents of the extracted samples (DS-E dishes); and the analytical hydrolysis method (E-H). Separate models were developed for these stages for the DS and DF samples.

Table F-36 provides summary statistics for the moisture contents at these various stages. It can be seen that, as with the peat and bagasse samples, the DS-E Dishes samples have a higher average moisture content and a wider range in MCs compared with the DS-E and E-H samples. The DF samples have less variation in moisture content, with a range of only 2.79% for the DF-E samples.

Table F-37 provides regression statistics for the various models that were developed. For all but one of these the samples for the calibration and validation sets were randomly selected. For the E-H DS model a hierarchical clustering method using average linkage and Euclidean distances was used to assign each of the samples to one of 25 clusters, based on their moisture contents. From each cluster one sample was randomly selected for the validation set and three for the calibration set. Sometimes a cluster contained less than 4 samples meaning that the total number of samples in the calibration set is less than 75. This clustering method of sample selection was chosen to see if it

could provide a better model where samples are more evenly distributed across the concentration range for both sets. However, the data in Table F-37 show that the standard method of random sample selection using all available samples provided better RMSECVs, RMSEPs and RERs. It is likely that these added samples are needed, not only for the information provided to the model regarding their moisture contents, but also due to the spectral variability they contribute to the model given that samples of various plant fractions need to be predicted.

With the exception of the “Dishes” models, the calibrations based on the DS samples have better regression statistics than the calibrations based on the DF samples. This may be due to the greater number of samples in the calibrations sets for the DS models, but it may also be partly due to the tendency of the DF samples to rapidly absorb moisture upon exposure to the air, bringing potentially higher degrees of analytical error into the models.

The DS-E model uses a very large number of PLS factors (15); however, there are many samples in the calibration set and the close fitting of the model developed on these to the, also large, independent validation set shows that the model can be applied to unknown samples (the RMSEP and RMSECV are very close).

The models developed have been used to predict the MCs of numerous samples prior to their extraction/hydrolysis in cases where insufficient material was available for reference moisture analysis. They are valuable tools in the standard practices now employed in the laboratories.

15.2.3 Differences Between Replicate Scans

For each dataset the models for GLU_SRS, KL, and Ash developed on *Miscanthus x giganteus* samples only were applied to the corresponding spectral dataset comprising the replicate scans in order to determine any differences in predictions that may occur upon rescanning. The results are provided in Table F-38 (see Section 13.3.4 for a description of the terms in this Table). The DG, DH, DS, DT, and DF scanning methods all involve the collection of two spectra per sample. The difference in prediction is then simply calculated by subtracting the second predicted value from the first predicted value. The WU, DU, and DV scanning methods, however, involved the collection of three spectra per sample. In order to obtain a difference value for each sample the maximum absolute difference between any two scans was selected for each sample. To calculate the bias for these scans the predicted value for the third scan was subtracted from the predicted value for the first scan. This was chosen since the time between the collection of these two spectra is the greatest;

hence, there is the most opportunity for the air-drying of the WU sample and any consistent bias between these two predictions may suggest a link to the slight decrease in moisture content.

It would be expected that the greatest repack errors would be associated with the samples of heterogeneous particle size (WU, DU, DV) and that, as mean particle size decreases, sample homogeneity increases and the repack error should fall. Furthermore, variations in the moisture content of the sample during the collection of spectra for the WU sample may introduce an added error to cell repacks. This theory is held up by the data in Table F-38; for each of the three constituents the repack error associated with the WU scans is greatest, while that for the DF fraction is the least.

Scanning methods DT, DH, and DV were introduced (see Section 11.1) in an attempt to improve the means of presentation of material to the cell over the DS, DG, and DU methods, respectively. As shown in Table E-8 and Table E-9 there are a similar number of samples in the DS/DT sets so comparison between the two is fair. Table F-38 shows that the DT method does reduce the average difference in predicted values between the duplicate scans for each constituent. The DH method does reduce the difference, compared to the DG method, for glucose but not for KL and ash. However, much fewer samples were scanned in the DH method so it is too early to judge. Furthermore, a much greater proportion of the DH scans were of WP samples which, as shall be discussed below for the WU scans, have the greatest repack error of all the plant fractions. The DV method reduces the average differences for the KL and ash predictions but not for the glucose predictions. Again, however, the DV set is much smaller than the DU set and more weighted towards WP samples.

Table F-38 also shows that, as with the peat samples, the repack error when expressed as a percentage of the average predicted value for each sample, is greater for the minor components. Indeed, in the worst case a WU scan resulted in the NIR model predicting that the sample had an ash content which was different from a replicate scan by over 3 times the average predicted ash content of that sample. In fact, the average repack error of the WU method was 25% of the actual ash content of each sample. These results reinforce the need for including multiple scans for each sample, particularly for the WU samples, so that any errors are smoothed out.

Given that the WU model is the most important model for the rapid prediction of the composition of unknown samples, a study was undertaken to see how the differences in predicted values between

replicate scans are distributed according to plant fraction. Table F-39 presents the results of this study, using the WU glucose model for prediction, and the distributions of the errors for each fraction are presented as histograms in Figure F-20.

The data in Table F-39 show that the lowest repack differences were associated with the live leaf blade samples (with an average difference in predicted glucose content between replicate scans of 0.73%) while the greatest differences are associated with the whole plant samples (1.31%). This is logical given that the heterogeneity of these WP samples means that a consistent proportion of stem and leaf fractions will not be presented to the NIR cell window with each replicate scan. The histograms also illustrate that there can be large differences in predicted glucose values between replicates (e.g. 3.98% for one WP sample). Careful monitoring of the predicted compositions of replicate scans for unknown samples will therefore be key in striving to obtain accurate predictions for these, or else the hard work involved in developing NIRS calibrations will be lost to simple errors associated with scanning the samples.

15.3 Summary

This chapter has presented and discussed the results obtained from the development of a large number of PLS models covering different spectral datasets, constituents, and plant variety groupings. The results for many of these models have been excellent with high RER values and low RMSEPs. The major focus in the research has been to demonstrate that good quantitative NIRS models can be developed for the major lignocellulosic constituents of relevance to biorefining technologies based on the spectra of wet *Miscanthus* samples.

Approximately 80% of the whole mass balance of most *Miscanthus* samples is represented by glucose (as cellulose), xylose (present in hemicellulose), and lignin. This research has demonstrated that RER values of over 15 are possible for each of these constituents using WU spectra only. These results suggest that NIRS can have major potential as an online tool in the nascent biorefining industry. The Author is aware of no published research, regarding any feedstock, that matches the levels of accuracy seen in these WU models. Indeed, the RMSEPs of the WU models are often superior to those in the literature for models based on dry homogenous material (see Appendix B).

The successes seen in the development of these models can be partially attributed to the strategy employed for sample collection and analysis. By separating the different anatomical fractions of

Miscanthus a much more diverse sample set is achieved. This provides important spectral and chemical variability to the models. Also, the time and effort involved in developing highly accurate and precise reference analytical methods have minimised the SELs for the constituents. This undoubtedly has a positive knock on effect regarding the RMSEPs obtained. An example of this is provided by comparing the models for rhamnose with the models for mannose. The relative precision of the reference analytical method for mannose was poorer than that for rhamnose and, correspondingly, the regression statistics for this constituent are not as good as those for rhamnose despite this constituent taking up a slightly higher proportion of the total mass balance of most samples.

Some other important points discussed in this chapter are summarised below.

- There is no consistent difference between models developed on all Miscanthus samples and those developed on only the Miscanthus *x giganteus* samples. For some constituents the RMSEP for the All-Varieties model is less but for other constituents it is greater.
- The DS/DT models can achieve high R_{pred}^2 values when predicting the compositions of DF samples; however, a bias and slope adjustment is needed to achieve a 1:1 relationship. The dynamics that take place in the hydrolysis/extraction of DF samples will need to be studied in more detail before definite conclusions can be made concerning these NIR predictions.
- Combining the DF and DS data improves the precision of the WU models.
- For some constituents the DU models are significantly poorer than models based on the other spectral datasets. The sample presentation method used in collecting DU scans may be responsible for this. More DV spectra will be required to test this hypothesis. A similar situation exists concerning the DG/DH models.
- Significant differences can be seen in the quantitative predictions of replicate scans when using the WU models. This relates to the heterogeneity of the material presented to the cell window. This phenomenon is likely to be less of an issue if NIRS models are to be applied in online facilities. This is because the final spectrum will be an average of the multiple spectra collected as feedstock passes the NIR cell window, as is the case in the online NIRS systems developed by BSES for sugar mills (see Section 12.1.4 and (O'Shea et al., 2010)).

In conclusion, this research has demonstrated the power of NIRS for the rapid and precise prediction of the important lignocellulosic properties of wet and minimally-processed Miscanthus samples. This can provide great improvements in productivity since time-consuming sample processing steps and reference analysis methods need not be employed. This will be outlined in Chapter 16 where predictions will be made for the compositions of a number of plants for which only WU spectra exist.

16 Lignocellulosic Properties of Miscanthus and Evaluation of the Crop as a Feedstock for Biorefining

16.1 Chemical Composition of Plant Fractions

Table G-1 presents compositional data for the various anatomical fractions of a plant that was sampled on October 12th 2007 from a 13-year old *Miscanthus x giganteus* stand in Oak Park, Carlow. The stem of the plant was close to 3 m in height meaning that there were three stem sections corresponding to each metre (0-1 m, 1-2 m, 2m +). At a later point these stem sections were separated into the internode and node samples. Each plant fraction was weighed while wet and its moisture content determined, allowing the relative proportion of stem/leaf fractions to be determined. According to total dry matter, 66.2% of the plant was stem sections (nodes or internodes) and 33.8% was leaf sections. These leaf sections were either: dead leaf blades (“F”), dead leaf sheaths (“H”), live leaf blades (“K”), or live leaf sheaths (“M”). There were no flowers on this plant. The only location in which flowering *giganteus* plants were observed was Shanagolden, a stand where *Miscanthus* was in its second year of production. The compositional data of flowers sampled from a plant at this location are also provided in Table G-1.

Data in Table G-1 labelled with an asterisk have been predicted using the WU NIRS models discussed in Section 15.2.2. All other data have been determined with reference analytical methods. Table G-1 also provides data for the whole plant (“WP”); this has been determined as the weighted average of the data of all the fractions of the plant. The data for the whole stem (“X”), and each metre of the stem (nodes and internodes combined, e.g. X1) have also been calculated as the weighted average of the relevant stem sections.

For all of the plants sampled, the internode sections contributed the majority of the dry matter to the total stem weight. For example, the first metre internode section (X1T) of the 2+ m plant described in Table G-1 contributed 91.2% to the total mass balance for the first metre section of the stem, with the remainder coming from the nodes. The X1T section also contributed 34.6% to the total dry matter of the plant, with 22.4% coming from X2T section and only 3.6% from the X3T section. Regarding the leaves, the K fraction contributed 15.1% of the total dry mass, the F fraction 11.3%, the M fraction 3.5% and the H fraction 3.9%. Note that these mass data are not provided in Table G-1, instead that Table provides the concentrations of various constituents by plant fraction.

There are several important observations regarding the 3 stem sections plant in Table G-1:

- Glucose is the largest constituent in all plant sections, this carbohydrate is clearly representative of the cellulose, which is the major polysaccharide in *Miscanthus* ((El Hage et al., 2010) as discussed in Section 10.5).
- Xylose is the second most abundant carbohydrate in the samples, 18.9% of the total mass balance of the whole plant (WP), followed by arabinose with 2.06%. The concentration of galactose is approximately 3 times less, in the whole plant, than that of arabinose, whilst mannose and rhamnose are minor constituents. This suggests that, as outlined in the literature (Le Ngoc Huyen et al., 2010) and also the case for sugarcane bagasse (Section 12.1.3), the hemicellulose in *Miscanthus* is principally an arabinoxylan with approximately one arabinose residue in the polysaccharide for every 10 xylose residues.
- This arabinose to xylose ratio (Ara:Xyl in Table G-1) varies according to the plant fraction, however. Arabinose is proportionately greatest in *Miscanthus* flowers and is also present in higher concentrations in the leaf blades compared with the stems, as noted in the literature (Section 10.5 and (Le Ngoc Huyen et al., 2010)). Arabinose is also present in higher concentrations in the nodes than in the corresponding internode section and the concentrations in both nodes and internodes increase with height up the stem, Figure G-1.
- The xylose concentration also increases with stem height, however, meaning that the Ara:Xyl ratio value is constant between the first and second metre stem sections; however, it almost doubles for the third metre section.
- The galactose concentration also increases with stem height and is consistently higher in the nodes than in the internodes.
- After cellulose and hemicellulose, lignin is the most abundant polymer in these plant fractions. It is present in low concentrations in the live leaf blades and upper sections of the plant and in greatest concentrations in the lower stem sections (Figure G-2). This is logical given that more structural support is required in these lower sections whereas the leaves and upper parts of the plant are more focussed towards photosynthesis. These results are consistent with those in the literature (see Section 10.5).
- Lignin is consistently at higher concentrations in the nodes than it is in the corresponding internode sections (with the difference being between 1 and 2%). This is also logical given the previously mentioned concept of structural support.
- Acid soluble lignin (ASL) follows an inverse relationship with KL; it is present in greater quantities in the leaves and upper sections of the plant than it is in the lower stem sections.

- The glucose content follows similar trends to the KL content. It is lowest in the live leaf blades and lower in the upper stem sections than it is in the lower stem sections (see Figure G-2). However, in contrast to KL, its content is consistently greater in the internode section than it is in the corresponding node section.
- Table G-1 also provides data for the hemicellulose to cellulose ratio (Hc:Cel). It is assumed that all of the glucose present came from cellulose and that hemicellulose is the sum of all other sugars. This ratio ranges from 0.42 for the X1T section to 0.74 for the K section and 0.76 for the X3N section. The ratio is 1.16 for the flowers sample showing that cellulose is not the major polysaccharide in this fraction.
- Ash content is higher in the leaf blades and upper stem sections and lowest in the lower parts of the stem (see Figure G-3). There is also consistently more ash in the nodes compared with the corresponding internode section. Ash and acid insoluble ash (AIA) contents are higher in the dead leaf fractions compared with the corresponding live leaf fractions, and the AIA also contributes a greater proportion of the total ash in these dead sections. The AIA:Ash ratio is also greater in the sheaths than the leaf blades for both the dead and live leaves with the ratios for these sheaths being greater than for any other fraction.
- The extractives content is greatest in the live leaf blades as would be expected given that the leaves' primary role is for photosynthesis and the assimilation of primary metabolites (Taub and Lerdau, 2000). The live leaf sheaths also have large extractives contents. With the exception of sample X3N, the extractives content does not change greatly with stem height.
- The nitrogen content is greatest in the live leaf blades; expected as nitrogen is an important part of the photosynthetic pathway of C₄ plants (Taub and Lerdau, 2000). The nitrogen contents of the stem sections generally increase with plant height, see Figure G-4. Nitrogen is also present at higher than average concentrations in the flowers.

Table G-1 also includes analytical data for the anatomical fractions of a much smaller plant. This plant was the same age as the taller plant and sampled, from the same stand, on December 4th 2007. The leaf fractions were 144.4% of the dry mass of the stem sections with most of the leaves being dead. Some important differences between this and the taller plant are listed below:

- Higher ash content in all sections.
- Lower stem lignin content than the X1 fraction of the 2+ m plant. The KL content is instead more similar to that of the X2 section of the taller plant. This is logical given that the stems

of the smaller plant were much thinner and were required to give less structural support than the lower stem sections of taller plants.

- Glucose content in the stem is also lower.
- Xylose and ASL contents of the X1 sections that are greater than those for the X1 sections of the taller plant and more similar to the upper sections of that plant.

Table G-1 provides a column for the total mass balance for the samples (“TOTAL”). This is the sum of the following contents: Extractives, ash, total sugars, KL, and ASL. AIA and AIR are not included because these are represented in the ash and KL concentrations. The nitrogen content is not included because it is possible that some of the nitrogen would end up in the KL. It can be seen the total mass closure for all samples is less than 100% and is significantly lower for the live leaf blade samples (where it is less than 90%) than it is for the other plant fractions. The remainder of the mass balance could come from uronic acids (which range between 1.01 and 1.96% of total mass for the 31 samples analysed, see Section 15.2.1), acetyl groups liberated from the acid hydrolysis of hemicellulose, and extractives components that are not soluble in 95% ethanol but will be present in the hydrolysate. As discussed in Section 3.4, a hot water extraction step may be employed instead of or as a precursor to ethanol extraction. Given the amount of laboratory work involved in the extraction of samples this was not considered to be practical for all samples. However, towards the end of the research a solvent controller was purchased for the ASE 200 (see Section 3.4.4). This allowed for experiments to be made comparing ethanol extraction vs. water extraction vs. a sequential water then ethanol extraction for selected samples. Samples covering a variety of ethanol-soluble extractives contents were selected for this analysis and the results are provided in Table G-2. A column shows the difference between the 95% ethanol soluble extractives and the total amount of ethanol extractives removed in the sequential extraction, and the final column in the Table expresses the extractives removed by ethanol extraction as a percentage of the total amount of extractives removed in the sequential extraction.

Table G-2 shows that the extractives removed by ethanol extraction, for the samples analysed, accounted for between 46 and 77% of the extractives removed in the sequential extraction of the sample. This proportion tends to be higher for the lower stem section samples than for the upper stem section and leaf samples. Importantly, for the live leaf samples the difference between the extractives removed in ethanol extraction versus those removed in the sequential extraction is large (close to 10% for some samples). This difference will bring the total mass balances for these samples much closer to 100%.

16.1.1 Statistical Tests

Table 16-1: Results for the ANOVA tests regarding whether there is a significant difference in the constituent value means between the plant fractions for three-stem-section-plants, collected between the months of October and December 2007.

Constituent	Test	Degrees of Freedom	F Value	Significance of Difference	Post-Hoc Test Used
Klason lignin	ANOVA	9,71	56.466	P <0.01	Gabriel's
Glucose	ANOVA	9,71	130.036	P <0.01	Gabriel's
Xylose	Welch	9,24.549	38.279	P <0.01	Games-Howell
Arabinose	Welch	9,24.421	438.444	P <0.01	Games-Howell
Galactose	Welch	9,24.980	77.164	P <0.01	Games-Howell
Rhamnose	Welch	9,27.812	13.089	P <0.01	Games-Howell
Mannose	Welch	9,25.375	37.121	P <0.01	Games-Howell
Hemicellulose:Cellulose Ratio	Welch	9,25.203	222.120	P <0.01	Games-Howell
Ethanol-Soluble Extractives	ANOVA	9,71	13.281	P <0.01	Gabriel's
Ash	ANOVA	9,71	7.526	P <0.01	Gabriel's
Acid Soluble Lignin (ASL)	Welch	9,26.087	234.590	P <0.01	Games-Howell
Nitrogen	ANOVA	9,71	28.346	P <0.01	Gabriel's

Using SPSS Version 18, One Way ANOVA was used to test for significant differences in the means between the different plant fractions for various constituents. In all cases the means were compared for the samples of *Miscanthus x giganteus* plants, with three stem sections, collected between the months of October and December 2007. Table 16-1 summarises the results of these tests. It shows whether an ANOVA test or a Welch F test was used (the former being used under conditions where the variances of the groups were the same, a condition tested with Levene's test) and provides the F value and degrees of freedom for each of these tests. It also specifies the significance level of the difference with either P < 0.01, P < 0.05, or no value (indicating that the difference between the means was not significant). It can be seen that there were significant (P < 0.01) differences for all constituents. However this is a test for any significant difference between all groups; the post-hoc tests allow comparisons to be made between pairs of groups. If Levene's test resulted in the acceptance of the null hypothesis that the variances of the groups are the same then Gabriel's test was used to test for significant differences between the two groups in each pair. If Levene's test resulted in a rejection of the null hypothesis then the Games-Howell test was used for post-hoc tests. The Figures for the means of each group are provided in Figure G-5 to Figure G-16 and the results of the post-hoc tests in Table G-3 to Table G-14.

16.2 Comparison Between Miscanthus Varieties

The Department of Agriculture project described in Section 14 only involved the collection and analysis of plants of the *Miscanthus x giganteus* variety. However, following the start of the DIBANET project (see Section 18) plants of the *Miscanthus x sinensis* variety were also collected and analysed as were 6 plants of different experimental varieties that were growing in a plot at the Teagasc Oak Park Research Centre in Carlow. These plants were sampled on October 15th 2009. The *sinensis* plants are pictured in Figure G-17 (a). It can be seen that these plants are much shorter than *Miscanthus x giganteus*. This is normal, in regions where *Miscanthus x giganteus* grows well *Miscanthus x sinensis* is less productive (Clifton-Brown and Lewandowski, 2002). *Sinensis* plants can have an advantage over *giganteus* plants in cold climates as this variety is more cold resistant, particularly in the establishment phase. The colours of the leaves and flowers of the *sinensis* plants were also more red than those of the *giganteus* plants.

Figure G-17 (b) contains a picture of one of the other varieties sampled; it is labelled MSXY1. It was approximately 1.5m in height. Figure G-17 (c) shows another variety, labelled MSXY2. It was of a similar height to MSXY1. Figure G-17 (d) shows another variety, labelled MSXY3. This plant, while not as tall as most *giganteus* plants, is taller than the other varieties in Figure G-17 and is also closer in appearance to *giganteus* plants. Figure G-18 (a) shows another variety, labelled MSXY5. This was the smallest of the plants sampled at this location. It was clearly quite different in appearance from *Miscanthus x giganteus*; the colour of the leaves was much more red, for example. Figure G-18 (b) shows another variety, labelled MSXY6. This was taller and more visually similar to *giganteus* plants. Figure G-18 (c) shows another variety, labelled MSXY7. All of the leaves and sheaths were considered to be “live” for this sample. Figure G-18 (d) contains a picture of *Miscanthus x giganteus* for comparison with the other varieties (the person present in this picture is several inches taller than the person present in the other photos).

The analytical data for these samples, using reference data where these were available (with NIRS predictions highlighted by an asterisk), are presented for the different plant fractions of these varieties in Table G-15, Table G-16, and Table G-17. Table G-15 also provides summary statistics for the reference analytical data for the *Miscanthus x giganteus* samples in order that differences between these and the other varieties may be observed. Where a constituent value for a given sample is greater than the maximum observed in all *giganteus* samples of that plant fraction it is highlighted in bold. Conversely, when a constituent value is lower than the *giganteus* minimum it is written in italics. Since the stems of the non-*giganteus* varieties were all sampled and analysed

whole (with no separation into nodes and internodes) the stem (“X”) data for the *giganteus* samples are computed as the weighted average of the compositions of all the internode and node sections in each plant. Some relevant observations regarding how the compositions of the non-*giganteus* samples differ from those of the *giganteus* samples are discussed below:

- Live leaf blades (K):
 - Samples MSSS5 and MSXY1 have higher glucose contents than any *giganteus* live leaf blades that were analysed via reference methods.
 - Samples MSSS5, MSXY1, MSXY2, and MSXY3 have higher rhamnose contents (0.50%, 0.61%, 0.60%, and 0.41%, respectively) than any of the *giganteus* live leaf blades
 - The live leaf blades of MSXY6 have a higher galactose content (1.24%) and higher AIR (17.6%) and AIA (3.2%) content than the *giganteus* samples.
 - MSXY1 has a higher arabinose content (3.9%).
 - MSXY7 has greater ash, rhamnose, AIR, KL and AIA contents than the maxima values for the live leaf blades of the *giganteus* samples, whilst it has lower glucose, xylose, and total sugars contents.
- Live leaf sheaths (M):
 - The live leaf sheath sample of MSSS5 is quite different from the live leaf sheath samples of *giganteus*, with arabinose, rhamnose, xylose, total sugars, and ASL contents that are greater than the *giganteus* maxima and ash, AIR, KL and AIA contents that are lower than the *giganteus* minima.
 - MSXY1 has lower ash, glucose, AIR, and KL contents than the minima in the *giganteus* samples and higher arabinose, rhamnose, xylose, total sugars, and ASL than the maxima in the *giganteus* samples.
 - MSXY2 has lower ash, galactose, rhamnose, AIR, and KL contents than the *giganteus* minima and higher xylose, total sugars, and ASL contents than the maxima.
 - MSXY3 has lower galactose, rhamnose, AIR and KL contents than the *giganteus* minima and higher xylose and ASL contents than the maxima.
 - MSXY6 has a lower rhamnose and xylose content than the *giganteus* minima and a higher ash and ASL content than the *giganteus* maxima.
 - MSXY7 has higher ash, arabinose, rhamnose and ASL contents than the *giganteus* maxima and lower xylose AIR and KL contents than the *giganteus* minima.
- Dead leaf blades (F):
 - MSXY1 and MSXY2 have higher rhamnose contents than the *giganteus* samples.
 - MSXY2 (2.87%) has a lower ASL content.

- MSXY5 has higher extractives (12.1%), rhamnose (0.59%) and ASL (5.37%) contents than *giganteus* samples but a lower xylose content.
- MSXY6 has higher ash (13.02%), rhamnose (0.44%), and AIA (7.95%) contents than the *giganteus* samples but a lower xylose content (15.81%).
- MSSS5 (*sinensis*) has higher rhamnose (0.75%) and KL (19.12%) values.
- Dead leaf sheaths (H):
 - These often differ substantially from the corresponding samples of *Miscanthus x giganteus*. The xylose contents of MSXY1 (23.66%), MSXY2 (23.55%), MSXY3 (24.02%), and MSXY5 (23.04%) are greater than any of the *giganteus* samples.
 - The arabinose content of MSXY1 (3.89%) is also greater while the AIR (16.72%) and KL (15.20%) contents are lower than any of the *giganteus* samples.
 - MSXY5 has lower galactose, mannose, AIR, KL, and AIA contents.
 - MSXY6, the arabinose (3.59%) and ASL (3%) contents are both greater than the *giganteus* samples, whilst the xylose content is lower (19.49%). The effect is that this sample has a relatively large, for sheaths, arabinose:xylose ratio of 0.18.
- Stems (X):
 - These often follow a similar trend to the H samples. For example, the stem sections of MSSS5, MSXY1, MSXY5, and MSXY2 have higher xylose contents than *giganteus* stem sections. The xylose content of the X2 section of MSXY2 is the greatest of all the stem samples.
- A whole plant *sinensis* sample was analysed (MSSS4 in Table G-15). This also has a high xylose content (24.54%). However, the arabinose content is also high, meaning that the arabinose:xylose ratio for this plant is 0.11, the same as the taller plant in Table G-1. These large concentrations of hemicellulosic sugars have an effect on the hemicellulose to cellulose ratio, which is 0.71 for MSSS4 compared with 0.52 for the taller plant in Table G-1.

Table G-17 provides results from the samples of harvested *Miscanthus* plots that were sent to the Carbolea labs by Dr. Angelika Eppel-Hotz in Germany. These samples are cross-breeds of different varieties that were collected from Dr. Dueter at Tinplant Biotechnik, a genomics company that was involved extensively in *Miscanthus* breeding. That breeding program was purchased by Mendel biotechnology in 2007. Seedlings of these cultivars were planted by Dr. Eppel Hotz at sites between 2001 and 2003 and some of the high yielding single plants were picked out from these plots, propagated by rhizomes, and planted in 2005 and 2006. The samples in Table G-17 represent the on-field standing biomass that was harvested in mid-March 2008. The standing stock consisted of stems and leaves and was primarily stems; however, Dr. Eppel Holtz did not quantify the ratio. For

comparison, Table G-17 also provides the average whole plant composition of 4 *Miscanthus x giganteus* plants of three-stem-sections that were sampled by the Author in March (2 from Shanagolden and 2 from Carlow).

Table G-17 shows that, as with many of the non-*giganteus* samples collected by the Author, the xylose contents of the German samples are greater than for the *giganteus* plants. Indeed, the minimum xylose content German sample has a greater xylose content than the maximum xylose content *giganteus* sample. The same is true for the arabinose content of the samples whilst the opposite is the case for the Klason lignin content (it is higher in the *giganteus* samples). Also, on average, the glucose contents of the German samples are lower than that of the *giganteus* samples. The net effect is that the hemicellulose to cellulose ratio increases from an average of 49% for the *giganteus* samples to an average of 59% for the German samples (ranging from 55% to 61%). This is a 20.4% increase in relative terms. The average total sugars content of the German samples is 1.84% greater than the average content for the *giganteus* samples. The effects that these changes in the relative proportions of carbohydrates would have on the yields in biorefining technologies would depend on the relative efficiencies of those processes (see Section 16.5). For example, Technology E (see Section 16.5) can achieve similar yields of ethanol from pentose and hexose sugars whilst technology C yields substantially less ethanol per tonne of pentose sugars than it does per tonne of hexose sugars. Using the average data for the German and *giganteus* samples, the German samples would yield an extra 10 litres of ethanol per dry tonne of Feedstock using technology E, but only an extra 3 litres per tonne using Technology C.

The summary statistics for the German samples show that there is relatively little variation between them, and the degree of variation is similar to the four *giganteus* samples. For instance the range in the glucose content for the *giganteus* samples is 3.63% whilst that for the German samples is 3.73%. Given that more than one plant made up the homogenised German samples compared with the selection of individual plants by the Author in his field sampling methodology, the data for each German sample may be considered to be more representative of that variety as a whole than a single plant of *giganteus* would. Nevertheless the differences between the varieties are still quite small. The greatest concentration value for each constituent is highlighted in bold whilst the lowest concentration value is highlighted in italics. Sample 9009 has the lowest glucose (41.03%) and xylose (20.78%) contents and also the lowest total sugar (65.33%) content indicating that this sample may provide the lowest yields in hydrolysis biorefining technologies. This sample also has the greatest arabinose to xylose ratio indicating that either the leaves to stem ratio is greater for this sample or that the xylan is more substituted with arabinose side groups. Sample 9005 has the highest glucose

(44.76%) and total sugar (69.26%) contents. Using biorefining Technology E (see Section 16.5) sample 9009 would yield 414 litres of ethanol per tonne whilst sample 9005 would yield 439 litres, an increase of 6%.

16.3 Comparisons Between DF and DS Samples

Many of the DF samples that were analysed via reference methods also had analytical data for the DS sections. Section 15.2.2.4 discusses the differences seen between the two particle size fractions for NIRS calibrations for various constituents. In this section the DS and DF data for the same sample are plotted and the samples classified according to the plant fraction to see if there are any trends. A black line, representing a one to one relationship, is included. The residual, which is determined as the DS content minus the DF content is also plotted against the DF content. Figure G-19 to Figure G-42 cover these two plot types for the following constituents: extractives, KL, total sugars, glucose, xylose, arabinose, galactose, rhamnose, mannose, acid insoluble ash, ASL, and ash. The datapoints are labelled according to whether the sample is of a live leaf blade (“K”), a dead leaf blade (“F”), a dead leaf sheath (“H”), a node (“N”) or an internode/full stem section (“T,X”). The full stem section samples are obtained by processing the X1/X2/X3 metre section of the plant directly through the chipper. This means that the DS/DF sample contains internodes and nodes; however, the majority (90%+) of the mass balance comes from the internode fraction, as discussed in Section 16.1. There are a total of 39 samples in most of the plots, these consist of 8 internode, 8 stem, 9 node, 6 dead leaf blades, 1 dead leaf sheath, and 7 live sheath samples. This number is less than the total of DF samples used in NIRS calibrations (Section 15.2.2.4) because some of the corresponding DS samples were not analysed. Only wet chemical analytical data were used for the comparisons.

Important observations regarding the trends seen in Figure G-19 to Figure G-42 are discussed below:

- The extractives residuals (Figure G-20) for all samples are less than zero. The residual tends to increase with increased DF extractives content
- The relationship with KL is much closer to the 1:1 line (Figure G-21). There is no clear pattern regarding the different plant fractions although the nodes (with the exception of one sample) all have positive residuals (Figure G-22).
- Regarding total sugars (Figure G-23), glucose (Figure G-25), and xylose (Figure G-27), the leaf samples (with the exception of the one H sample) tend to lie much closer to the 1:1 line than

the N/T,X samples. All the node samples have positive residuals for these constituents (with the exception of the glucose content of one sample). The internode/stem samples are more widely distributed but also tend towards positive residuals.

- The F and K samples tend to fit the 1:1 line for arabinose (Figure G-29) and galactose (Figure G-31) contents while node samples deviate and have relatively large negative residuals. For example, the galactose residual (Figure G-32) for one node sample is -0.55%, a value that is 40.0% of the DF galactose content.
- For rhamnose (Figure G-33) the situation is different, with the F and K samples deviating most from the 1:1 relationship. The magnitude of the positive residual increases with DF rhamnose content.
- The ASL residuals (Figure G-40) are also greater for the F and K fractions but the values do not seem to be related to concentration.
- There are no clear trends between the different fractions for the AIA (Figure G-38) or ash (Figure G-42) residuals.
- Some large residuals are seen for the dead sheath sample (“H”) (e.g. Figure G-30 for arabinose); however, no definite conclusions can be drawn from a sample set of one. More corresponding DS samples of this fraction need to be analysed to allow extra points to be plotted.

The trend for greater residuals to be seen in the internodes/nodes can be explained by the different tissue types that make up these samples. The pith (Moller et al., 2007) is a soft spongy material that is located in the centre of the stem whilst the xylem and bark are more woody. It is logical that there may be differences in the proportions of pith and xylem/bark that end up in the DS and DF fractions. Liu *et al.* (2010b) separated the stems of corn residues into their pith and rind (the more woody material) fractions. The analysis of these sections showed that the rind had higher lignin but lower arabinose and galactose contents than the pith.

The DS minus DF residual for each sample can be expressed as a percentage of the DF content for that sample. Figure G-43 takes the average of these percentages for each plant fraction and plots these for each constituent. It is a good result that the relative differences for the most important constituents (glucose, total sugars, xylose, KL) tend to be smaller since this indicates that NIRS calibration developed on just the DS samples would not differ too greatly from the true composition of the whole material (DS and DF samples combined). The greatest relative differences are (with the exception of the extractives) seen for minor constituents such as galactose, rhamnose, and mannose.

These larger differences may explain the greater discrepancies sometimes seen in the precision of the NIRS models for these constituents based on the whole samples (DG, DU, WU).

Figure I-44 presents a histogram, with associated statistics, for the percentage, for the nodes samples, of the total DG material that was DS (the remainder being DF). Table I-18 presents the same for the other plant fractions. The data show that the nodes samples have the lowest average DS percentage. Given the significant differences between DS and DF compositions seen for the nodes it may be the case that this fraction could be responsible for much of the inherent error in NIRS calibrations developed on DS data. Various improvements to the sample preparation methodology were developed over time, resulting in the average % DS material for later samples being greater than for the early samples.

16.4 Changes During the Harvest Window

16.4.1 Changes in the Relative Proportions of Stems and Leaves

Section 14.1.1 detailed the sampling protocol that took place, between the months of October 2007 and April 2008 over 7 *Miscanthus* stands. Section 14.1.1 also provides details for these sites and Figure 14-1 shows when each plant was sampled from each location. The plants were collected for two reasons:

1. To examine how the relative mass proportions of the leaves and stems changes over the harvest window; and
2. To provide samples of varying physiochemical compositions in order to expand the concentration ranges and spectral variation of samples for NIRS calibrations.

In order to satisfy the second requirement some plants that were different from the majority of the plants in the stand were selected. This selection was primarily based on plant height. Using the example of the Shanagolden site, the majority of the plants were over 2 m in height meaning that they had three stem sections (X1, X2, X3); however, in order to increase spectral variability some smaller plants (e.g. those with only one or two internode sections) were sampled from the field. These smaller plants typically were on the boundaries of the plantation and appeared similar to the typical plants seen in the first year plantations (Langton and Clonmel).

As described in Section 14.1.1, all the standing biomass of one plant was always sampled. The small plants had a much lower total stem mass than the taller plants that were selected from the same site. It is known (see Section 10.2.1) that the proportion of leaves will fall over time after senescence; hence observing differences between plants of differing total stem mass can only occur in a relatively short window where the effects of leaf loss over time do not become dominant. Figure G-45 plots the total leaves weight, expressed as a percentage of the total stem weight, for all plants collected from stands in the month of January 2008. Total dry leaf mass is defined as the sum of the dry weights of dead leaf blades, dead leaf sheaths, live leaf blades, and live leaf sheaths, whilst total dry stem mass is defined as the sum of the dry weights of all the internode and node sections for that plant. Figure G-45 shows that the percent leaves is related to the total stem weight of the plant. A 3rd order polynomial regression curve is fitted to the datapoints, providing an R^2 of 0.936. Hence, a more productive plant will grow taller and the relative mass proportion of leaves will fall as the stems increase in diameter and lignify.

Figure G-45 shows that an experiment cannot examine the effects of harvest time on the stem/leaves proportions if the heights/weights of the plants are also a varying factor. Therefore, for this examination only plants that had the same number of stem sections were compared over the harvest window.

Shanagolden Site

This stand was in its second year of production and the majority of the plants sampled had three stem sections. Figure G-46 plots the variation in the percent leaves over time for such plants. The first two plants were sampled on November 15th 2007. There was a significant amount of leaves on both these plants, with an average of 17% of the total dry mass of both plants being either live leaf blades or live leaf sheaths, and a further 17% of the total dry mass being either dead leaf blades or dead leaf sheaths. When the stand was visited one month later the proportion of live leaves had fallen substantially to an average of 3.3% of the total dry mass of the plant. Over this month the leaves had lost their colour, becoming “dead” according to the classification outlined in Section 14.1. However, it was noticed that leaf fall was not substantial at this point and most of the leaves were still present on the plant (the amount of dead leaf fractions, expressed as a percentage of total plant mass, on both plants rose to an average of 28%, from 17% the previous month).

When the stand was again visited in January 2008 it was readily apparent that many of the leaves had fallen from the plant. Two plants (of three stem sections) were sampled at this time and these had no live leaves/sheaths present and the contribution that dead leaf fractions made to the total mass balance of the plants fell to an average of 21%. Furthermore, the majority of the dead leaf mass was provided by the sheaths rather than the leaf blades. These sheaths are encased around the stem and less prone to breakage/displacement from the stem than the leaf blades. Over the subsequent months the relative proportion of total dead leaves continued to fall but at a substantially lower rate. The principal reason for the continued reduction was the continued fall over time of the leaf blades that still remained on the plant. The last plant from this stand was sampled on April 16th 2008. The amount of dead leaf blades that were still present on the plant at this point only contributed 0.84% of the total dry mass of the plant. In contrast, the proportion of total plant mass provided by the dead sheaths remained relatively constant over time. This is illustrated in Figure G-47; the values are lower at the first sample date since less sheaths could be characterised as “dead” at this time.

Typically in Ireland *Miscanthus* stands are harvested in March or April. If the plant sampled on April 16th 2008 is used to represent the stand at harvest then it can be assumed that the leaf fractions of the crop comprise 13.8% of the total mass balance, i.e. their mass is 16% of the total stem mass. The average percent leaves of the two plants collected in November 2007 was 53.0% of stem mass or 34.7% of total plant mass. If it is assumed that the dry stem mass is constant over the harvest window, then the extra leaf material present in the early harvest represent a potential extra 31.9% in total dry matter yield. With an April yield of 12 dry tonnes per hectare, this would translate to a November yield of 15.8 dry tonnes per hectare.

Some smaller plants were also sampled from Shanagolden over this harvest window. The percent leaves (expressed in terms of percent of total dry mass of stems) for these plants are plotted in Figure G-48. A regression line is added to the 2-stem sections plant since these were sampled on three different dates; however, no line is included for the one-stem-section plants since these were only sampled on two dates. It is clear that, as previously shown in Figure G-45, these smaller plants have much higher proportions of leaves. Indeed, the two one-stem-section plants that were sampled on February 19th 2008 had a leaf mass that was almost the same as the stem mass (96.2%, on average). In contrast, the two 3-stem-section plants that were also sampled on this date had an average leaves proportion of 20.8% of total stem mass. At this date, whilst there were no live leaf fractions, there was still a significant amount of dead leaf blades on the plants (an average of 20.31% of the total mass of the two plants).

Adare-H Site

This stand was in its third year of production. All of the plants sampled from this location had three stem sections. Figure G-49 plots the variation in the percent leaves over time. The plot is similar to that seen for the 3-stem-section plants at Shanagolden, Figure G-46. A third order polynomial regression line is fitted to the datapoints, the R^2 value here is 0.9698. The last samples were collected on April 13th 2008, and the leaf fractions were on average 13.0% of the total stem mass, compared with the leaves being 56.4% of total stem mass for the first sample (collected on November 27th 2007). The relative advantage of an early harvest in this instance is an increase in total dry matter yield of 38.4%. Therefore, if the late harvest provides a yield of 12 tonnes per hectare the early harvest would provide a yield of 16.6 dry tonnes per hectare.

Adare-C Site

This stand was in its second year of production and the majority of the plants growing in this plantation were over 2 m in height. Figure G-50 presents the variation in the percent leaves over time for plants of three stems sections. The percent leaves for two plants that only had two stem sections, collected on November 28th 2007, are also plotted. The second order polynomial regression line is only fitted to the three-stem-section plants datapoints. This stand was harvested earlier than the Adare-H and Shanagolden stands meaning that the last samples were collected on February 28th 2008. This is the probable reason why the percent leaves for the latest datapoint (17.0%) is greater than the percent leaves for the last datapoints for three-stem-section plants from other sites.

Carlow-F and Carlow-G Sites

These sites, as described in Section 14.1.1, had Miscanthus plants in their 13th year of growth. The sites had been established and maintained in the same way but were separated by a field of another crop. The results for both sites have been combined and are presented, for plants with three stem sections, in Figure G-51. Plants were sampled from these sites on the 12th and 13th of October 2007, an earlier sampling date than for the other locations. The datapoints from the two plants that were sampled on these dates provide extra interesting information about the dynamics seen in leaf:stem proportions in Miscanthus stands following senescence. There are only minor differences seen in the percent leaves between 13/10/07 and 3/12/07. That indicates, as also suggested by the data for the plants collected from the Shanagolden site over the first two dates, that leaves are not lost to a significant degree over this period. As with the Shanagolden site, however, the leaves did change in colour over this period meaning that the dead leaf blades were more abundant than the live leaf

blades in the December samples. A 5th order polynomial regression line fits the datapoints reasonably well and provides an R^2 of 0.9825.

The last samples were collected on April 7th 2008 and the leaf fractions were on average 15.0% of the total stem mass for these samples, compared with the leaves being an average of 53.4% of total stem mass for the first two samples. The relative advantage of an early harvest in this instance is an increase in total dry matter yield of 33.4%. Hence, if the late harvest provides a yield of 12 tonnes per hectare the early harvest would provide a yield of 16.0 tonnes per hectare.

Some smaller plants were also sampled from these Carlow sites. The percent Leaves (expressed in terms of % of total dry mass of stems) for the two-stem-section plants are plotted in Figure G-52. The relationship regarding stem to leaf proportions is less clear here and reflects the heterogeneous nature of the smaller plants that were sampled. A one stem-section plant was also sampled from the Carlow F site on December 4th 2007; its total dry leaves mass was 144.4% of the total dry stem mass, or 54.9% of the total dry mass of the plant.

Langton Site

This was a site with *Miscanthus* plants in their first year of growth. There was no clear modal height class of plants with some being less than 1 metre high and others taller than 1 m; however, no plant achieved a height greater than 2 m. It was also very clear that the plants had not established well in some parts of the field - the growth was patchy and the establishment success rate was said by the farmer to be poor. This was attributed to the use of an inefficient planting method, according to the farmer. The percent leaves for the 8 plants that were collected from this site are presented in Figure G-53. As seen in Figure G-48 the one-stem-section plants have a higher proportion of leaf material; the plant sampled on January 10th 2008 had a total dry mass for leaves that was 102% of the total dry stem mass. The last samples were collected on the 7th of March 2008 since the crop was cut shortly after that date.

Clonmel Site

This was also a first-year plantation with inconsistent heights across the various plants on the site. The percent leaves for the 7 plants that were collected from this site are presented in Figure I-54. These datapoints are combined with those from Figure G-53 to produce Figure G-55, which shows the trend in decreasing percent leaves for the two-stem-section and one-stem-section plants of these one-year plantations.

Combination of Data Across All Sites

Figure G-56 combines all of the leaf data of the three-stem sections plants for the 5 sites. A polynomial regression curve is fitted to this plot and, while the R^2 value is less than for the individual sites, it is reasonable considering that Figure G-56 covers stands of differing ages.

Figure G-57 combines the leaf data of the two-stem sections plants for the 6 sites. The fit is much poorer here, reflecting the heterogeneity of these plants. Figure G-58 combines the leaf data of the one-stem sections plants for the 4 sites. When these sites are grouped together the reduction in leaf content over time appears fairly linear over the given period.

16.4.2 Changes in Composition over Time

Examinations were conducted to see if there were changes in the chemical composition of the samples over the harvest window. These tests were done on certain plant fractions and on the plant as a whole (which involved the weighted average of the compositions of each fraction). Where wet-chemical data were available these were used, but in their absence the WU NIRS models described in Section 15.2.2 were used to predict the composition of each sample. Table G-19 presents the R^2 and Pearson correlation coefficient (r) values indicating the relationship (or absence of a relationship) between composition and date for whole plant samples at the Carlow, Adare-H, Adare-C and Shanagolden sites. Since the plants at the Adare-C and Shanagolden site were of the same stand age their data are also combined in Table G-19. For each of these sites only the plants of over 2m height (i.e. 3 stem sections) were used for comparison. Table G-19 also presents correlation coefficient values for the samples grouped according to their number of stem sections. The groups for 2 and one stem section plants include samples from the Langton and Carlow sites as well as samples from Carlow, Adare, and Shanagolden. The value for r was tested to see if it was significant at $p = 0.05$, using the t-test value for correlation (see Section 6.7) and determining from this the p value using the 2-tailed t-distribution with $(n-2)$ degrees of freedom. If this value was less than 0.05 then the correlation can be considered to be statistically significant at the 95% confidence interval. Such statistically significant correlations are highlighted in bold in Table G-19.

Some of the important observations regarding these data are discussed below:

- For most groups there is a statistically significant ($\alpha = 0.05$) negative correlation between extractives content and later dates for sample collection. This can be explained by the decrease in the relative proportion of leaves (see Section 16.4.1), a fraction that had higher extractives content than the stems (see Section 16.1).
- There are also statistically significant negative correlations for ash (Figure G-59 for 2-stem-section plants), arabinose, galactose, rhamnose, ASL (Figure G-60 for the Adare-C and Shanagolden sites), AIA, and nitrogen. These can also be explained by the higher concentrations of these constituents in the leaves, particularly the live leaf blades.
- In contrast, there are statistically significant positive correlations between constituent value and harvest date for KL (Figure G-62 for the Carlow sites), AIR, glucose (Figure G-61 for the Adare-C and Shanagolden sites) and total sugars. This is also rational given that the concentrations of these constituents is greater in the stems whose contribution to total mass balance increase over time.

Statistical Tests

Table 16-2: Results from ANOVA tests to determine if there is a significant difference in the constituent value means of the “Early” and “Late” WP samples.

Constituent	Test	Degrees of Freedom	F Value	Significance of Difference
Extractives	ANOVA	21	11.259	P < 0.01
Ash	ANOVA	21	17.258	P < 0.01
Arabinose	ANOVA	21	62.728	P < 0.01
Galactose	Welch	9.650	26.936	P < 0.01
Rhamnose	ANOVA	21	18.924	P < 0.01
Glucose	Welch	9.948	29.221	P < 0.01
Xylose	ANOVA	21	3.403	
Mannose	ANOVA	21	4.838	P < 0.05
Klason Lignin	ANOVA	21	33.518	P < 0.01
Acid Soluble Lignin	ANOVA	21	37.231	P < 0.01
Nitrogen	ANOVA	21	6.355	P < 0.05
Hemicellulose to cellulose ratio	ANOVA	21	26.835	P < 0.01
Leaves content (as a percentage of dry stem weight)	Welch	16.261	82.399	P < 0.01

Using SPSS Version 18, One Way ANOVA was used to test for significant differences in the means of various constituents between whole plants sampled in the “Early” harvest window (October, November, December) and whole plants selected in the “Late” harvest window (March, April). In all cases the means were compared for the samples of *Miscanthus x giganteus* plants with three stem sections and the whole plant composition was determined as the weighted average of the

compositions of the different fractions (leaves, nodes, internodes etc.). Table 16-2 summarises the results of these tests. It shows whether an ANOVA test or a Welch F test was used (the former being used under conditions where the variances of the groups were the same) and provides the F value and degrees of freedom for each of these tests. It also specifies the significance level of the difference with either $P < 0.01$, $P < 0.05$, or no value (indicating that the difference between the means was not significant).

Figure G-63 plots the means, of the Early and Late groups, for the glucose, xylose, Klason lignin, leaves to stems percentages, and hemicellulose to cellulose percentages whilst Figure G-64 has the corresponding plots for the extractives, ash, arabinose, galactose, rhamnase, mannose, ASL, and nitrogen contents.

Changes within Individual Plant Fractions

Many of the trends seen in Table G-19 and Table 16-2 can be explained by the changes in the relative proportions of stems and leaves over time. Whether or not the compositions of separate plant fractions change over time was also examined. The correlation statistics, using the same plants and groupings as in Table G-19, are presented for the dead leaf blades in Table G-20, for the dead leaf sheaths in Table G-22, and for the whole-stems (the weighted average of all internode and node sections of the plant) in Table G-23. Some observations about these data are summarised below:

- For all groups there is a statistically significant positive correlation between sample collection date and the galactose content of the dead leaf blades (e.g. Figure G-65 for samples from Carlow). This was confirmed in an ANOVA that compared the galactose contents for the dead leaf blades between samples collected in an “Early” harvest and those in a “Late” harvest. Levene’s test resulted in the rejection of the null hypothesis that the variances of the groups are the same so the Welch F value was used. It was found that there was a significant effect of harvest period $F(1,17.529) = 74.088$, $P < 0.01$.
- For all groups, except Adare-H, there is a statistically significant positive correlation between sample collection date and the mannose content of the dead leaf blades. This was also confirmed in an Early/Late harvest ANOVA; $F(1,40) = 71.320$, $P < 0.01$.
- These trends may not necessarily be related to chemical changes in these leaf blades over time, but may instead be related to the types of leaves that stay on the plant for longer periods and those that fall earlier.
- Most of the other relationships seen for the dead leaf blades are inconsistent between locations and groups.

- There are no consistent strong correlations between harvest date and constituent concentration for the dead leaf sheath samples.
- There are statistically significant positive correlations between harvest date and KL (see Figure G-66 for samples from Carlow) and AIR concentrations for the whole-stems of 3-stem-section and 2-stem-section plants. The Pearson correlation coefficient values for the one-stem-section plants are similar but not statistically significant (at $\alpha = 0.05$) due to the limited number of samples of these plants. These observations were explored with ANOVA tests comparing Early/Late harvests and are summarised in Table G-21 which shows that there are significant differences ($P < 0.05$) for the 2-stem section and 3-stem section plants. This relationship could exist because KL is required for structural support and stems with lower contents are likely to be more prone to lodging (only the standing stems were sampled).

16.5 Potential Yields from Biorefining

This Section links the trends seen in total dry matter yield and the chemical composition of the crop available for harvest (at the time of sample collection) to potential yields from processing the harvested Miscanthus in several biorefining technologies. Six different hydrolysis technologies are examined; these are discussed in detail in a paper by the Author (Hayes and Hayes, 2009), and will be summarised here:

(A) – Dilute acid hydrolysis of biomass in two plug-flow reactors (Nguyen, 1998). This can be considered to representative of a near-commercial dilute-acid hydrolysis facility.

(B) – Dilute acid hydrolysis of cellulose in a counter-current reactor with an uncatalysed steam hydrolysis pre-treatment (DiPardo, 2000, Sun and Cheng, 2002). This more efficient process (for cellulose hydrolysis) may be commercially viable in the future.

(C) – Concentrated acid hydrolysis of biomass (Broder and Barrier, 1988). This technology, and the predicted yield, is similar to that employed by BlueFire Ethanol (Farone and Cuzens, 1996) which, in 2011, is constructing a commercial-scale biorefinery to produce ethanol from lignocellulosic wastes.

(D) – Enzymatic hydrolysis of biomass. Involves a dilute acid pre-treatment and separate fermentation of the monosaccharides from cellulose and hemicellulose (sequential hydrolysis and fermentation – SHF) (Hamelinck et al., 2005). Cellulase enzymes are produced in a separate reactor to that for hydrolysis. This is the likely setup of the first commercial enzymatic hydrolysis facilities planned by companies such as Iogen.

(E) – Enzymatic hydrolysis and fermentation of biomass via consolidated bioprocessing (CBP) (Sun and Cheng, 2002) with a liquid hot water pre-treatment step (Wyman et al., 2005). Here hydrolysis of cellulose, fermentation of the sugars and production of cellulases all take place in one reactor and involve a single micro-organism. This process can be considered to potentially be the most efficient and economical enzymatic hydrolysis technology (Lynd, 1996); however, it is currently not sufficiently developed for commercialisation. There is substantial ongoing research, however, (Mosier et al., 2005, Lynd et al., 2002, SunEthanol, 2007) and it is expected that such a process could be viable before 2020.

(F) – The DIBANET technology (see Chapter 18). It is conceptually similar to the Biofine process which involves the dilute acid hydrolysis of lignocellulosic polysaccharides and subsequent dehydration of the products to levulinic acid and co-products (Fitzpatrick, 1997, Hayes et al., 2005). Polysaccharide-derived hexoses yield levulinic acid (LA) and formic acid (FA) (50% LA and 20% FA by mass hexoses are assumed (Hayes et al., 2005)). Hemicellulose-derived pentoses yield furfural (50% by mass (Hayes et al., 2005)), which can be converted to LA (93% conversion by mass (Timokhin et al., 1999)). Here it is assumed that all furfural is converted to LA and that ethyl levulinate (EL) is then produced via esterifying LA with fuel-grade ethanol (with a yield of EL that is 95% of the theoretical maximum). EL can be used in regular diesel engines up to 20% (Hayes et al., 2005). According to their molar masses, approximately 400 kg of ethanol is required per tonne of LA.

Technologies A to E produce ethanol, the following formulae are used to calculate the yield of ethanol according to the hexose and pentose contents of the feedstock and the efficiencies of the technology (Hayes and Hayes, 2009):

$$E_{cl} = \frac{1110^a \cdot 0.5111^b}{(0.789)^c} \frac{100 - Cl_{con}}{100} \frac{H_{cl}}{100} \frac{F_{gl}}{100} \quad (16.1)$$

$$E_{hc} = \frac{1136^d \cdot 0.5111^b}{(0.789)^c} \frac{H_{hc}}{100} \frac{F_{hc}}{100} \quad (16.2)$$

where

- E_{cl} – litres ethanol per tonne of cellulose, dm^3/t ;
- E_{hc} – litres ethanol per tonne of hemicellulose, dm^3/t ;
- Cl_{con} – cellulose consumed by cellulases, %;
- H_{cl} – hydrolysis efficiency for cellulose to glucose, %;
- H_{hc} – Hydrolysis efficiency for hemicellulose to sugars, %;
- F_{gl} – the fermentation efficiency of glucose;

- F_{hc} – the fermentation efficiency of hemicellulosic sugars.
- a – the mass yield of glucose per tonne of cellulose with 100% conversion efficiency is 1110 kg due to the conversion from the polymeric to the monomeric form;
 - b – the theoretical maximum mass yield of ethanol via conventional fermentation (100% efficiency) is 51.11%;
 - c – the density of ethanol is 0.789 kg/ dm³;
 - d – the mass yield of pentoses per tonne of pentose-sugars with 100% conversion efficiency is 1136 kg due to the conversion from the polymeric to the monomeric form;

The data for the Shanagolden site was used to examine the variations in potential ethanol/levulinic acid yields over the harvest window. As described in Section 16.4.1, the average percent weight leaves (calculated in terms of the total stem mass) at the last date for sample collection is used as a baseline for an expected yield of 12 dry tonnes per hectare, and any changes in this percentage are used to change the expected total yield at that point in the harvest window (Figure G-67). It is assumed that stem dry mass is constant over the window. These yields are then linked to the whole plant compositions determined from either wet chemical or NIR-predicted data (these are represented by the total hexose and pentose curves in Figure G-67), in order to calculate the expected biorefinery yields per hectare, Figure G-68. Only the data for 3-stem section plants (the modal class) are used.

Table 16-3 shows that Technology E has the highest yields of ethanol from both the pentose and hexose sugars. Correspondingly, this technology achieves the highest yields at any given point in the harvest window. These yields are greatest in the Early harvest scenario. Figure G-67 shows that the pentose content is stable over the course of the harvest window while the glucose content rises. This represents an increase in the cellulose content of the standing stock of biomass over the harvest window. Figure G-69 shows how these changes in the composition of the feedstock influence the yields of end-products per tonne of feedstock according to technologies A to F. So that technologies A to E can be compared with the DIBANET process (F) the end products are quantified in energy terms (GJ/tonne feedstock). For DIBANET, the mass proportion that LA contributes to EL is used to ascertain the energy value contribution provided by the LA produced from the biomass. Figure G-69 shows an increase in yields per tonne for each technology over time. The net effect of this dynamic is that the slopes of the curves representing decreased ethanol/levulinic acid yields per hectare with a delayed harvest, Figure G-68, are less than that of the curve for total biomass yield, Figure G-67.

Table 16-3: Conversion factors and yields for the ethanol-producing hydrolysis technologies A-E. A = near-term dilute acid hydrolysis process; B = advanced dilute acid hydrolysis process; C = near-term concentrated acid hydrolysis process; D = near-term enzymatic hydrolysis process; E = advanced enzymatic hydrolysis process.

	(Biorefining Technology) (Reference)				
	(A)	(B)	(C)	(D) (Hamelinck et al., 2005)	(E) (Hamelinck et al., 2005)
Cellulose consumed by cellulases (%)	0	0	0	6	4
Hydrolysis of cellulose to glucose (%)	50 (Nguyen, 1998)	84 (DiPardo, 2000)	87 (Broder and Barrier, 1988)	75	98
Hydrolysis of hemicellulose to sugars (%)	85 (Sun and Cheng, 2002)	55 (Sun and Cheng, 2002)	95 (Broder and Barrier, 1988)	82.5	93
Kg hexoses per tonne cellulose ^a	555	932	966	783	1044
Kg sugars per tonne hemicellulose ^b	966	625	1079	937.2	1056
Ethanol from glucose (% of theoretical) ^c	90 (Nguyen, 1998)	95 (DiPardo, 2000)	95 (DiPardo, 2000)	87.5	93.5
Ethanol from hemicellulose sugars (% of theoretical) ^{c, d}	59 (Sonderegger and Sauer, 2003)	86 (Hamelinck et al., 2005)	59 (Sonderegger and Sauer, 2003)	59	93.5
Litres Ethanol per tonne cellulose	324	574	594	444	633
Litres Ethanol per tonne hemicellulose	369	348	412	358	640

a = the mass yield of glucose per tonne of cellulose with 100% conversion efficiency is 1110kg; b = the mass yield of pentoses per tonne of pentose-sugars with 100% conversion efficiency is 1136kg; c = the theoretical maximum mass yield of ethanol is 51.11%; d = for the near-term commercial technologies (A, C, D) a lower efficiency (Sonderegger and Sauer, 2003) is used since high yields have not yet been demonstrated at commercially-viable production rates.

Figure 16-1 illustrates this point by using the example of the yields of LA from the DIABNET process. These are expressed in terms of a percentage increase in LA yield per hectare associated with earlier harvests, compared to that which would be experienced from harvesting the biomass at the latest point in the harvest window. It can be seen that a maximum improved yield of up to approximately 20% can be achieved with an early harvest. This is lower than the 31.9% increase in dry matter yield discussed in Section 16.4.1. Furthermore, there is a significantly lower change in the expected yield in the months of October and November and March and April than for the interim period. These two periods fit the concept of the “Early” and “Late” harvest categories, Section 14.4.2, reasonably well. Considering the “Early” window, the harvest of biomass could be taken towards the latter part of this period, perhaps at a point where more of the nitrogen has translocated to the rhizomes, with little negative effect on biomass yield. Yield considerations become important moving beyond the Early period until the “Late” period is reached whereupon yields and compositions will be stable for

several months. Hence, the exact date of harvesting samples in the “Late” window could be flexible according to other considerations (e.g. consistency of supply rates of feedstock to the biorefinery, weather conditions etc.).

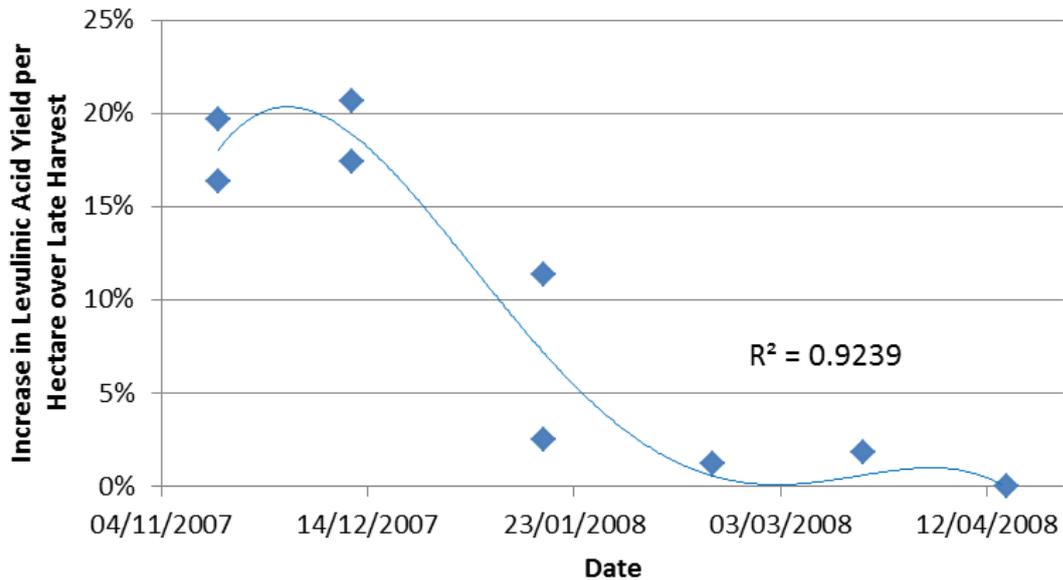


Figure 16-1: The increased yield per hectare of levulinic acid, over that experienced from the latest point in the harvest window (the datapoint from April), associated with earlier harvests.

Technologies A, C, and D are considered to be the near term options for commercial facilities and, of these, C provides by far the greatest yields. However, this process requires that the moisture content of the material is not significantly greater than 10%. The high moisture contents (over 50%) seen for Early harvest samples may therefore prohibit processing these in C. The ethanol yields from a Late harvest of process C are greater than the Early harvest yields of technologies A and D, therefore the relative advantage of an Early harvest may only be relevant for biorefining once more modern technologies (B, E, F) are commercially viable. When such a point is reached, the maximum potential ethanol yield, of approximately 6,250 litres per hectare per year associated with processing an Early harvest through technology E, is significantly greater than the expected ethanol yields from current first-generation biofuel crops such as wheat (3142 l/ha), barley (2065 l/ha), and sugar beet (4685 l/ha) in Ireland (Power et al., 2008). Indeed, the Late harvest ethanol yields from technology E would also be greater than all these first-generation alternatives.

16.6 Summary

This Chapter has explored the trends, both within and between plants, in compositional data of relevance to biorefining processes. The sampling methodology employed has allowed for estimations to be made regarding how a field of *Miscanthus* may change over the course of the harvest window. The trends that take place over this period have been examined for plants of differing heights and stand ages.

The collection of these plants and their separation into the different anatomical fractions was a time consuming process. For that reason only a handful of plants were sampled from each location at any particular date. These plants also often varied in their heights, meaning that these could all not be included in the same experiments. That means that, for any given sample collection date, there are typically one or two datapoints for samples that are representative of the modal height class of plant at each particular site. This limited number does bring some error into the projections made for total biomass yield and average field composition, as shown in the instances whereby the percentage leaves is less at an earlier date than it is at a later date (see Figure G-49, for example). If a significantly greater number of personnel were available for this project then an increased number of modal plants would have been selected per site, smoothing out any errors associated with atypical individual plants. However, even given the limited number of plants collected, in most cases the datapoints fit the polynomial regression curves reasonably well and, using these curves, clear trends regarding the changes in leaf proportions and whole plant constituent concentrations can be understood. These then allow for informed discussions regarding the potentials for *Miscanthus* as a biorefining feedstock to be made.

Some of the important points brought forward by this research are summarised below:

- Some varieties other than *giganteus* have been analysed. In most cases these varieties have increased xylose and arabinose but lower cellulose contents. Whether this is an advantage or disadvantage for potential biorefining yields will depend on the technology used for processing these samples.
- The differences in compositions between the DS and DF fractions of samples have been studied in more detail and these differences vary according to anatomical fraction. The greatest differences are seen for the stem section samples, particularly the nodes. The improved sample comminution methods developed over the course of the research have reduced the relative proportions of the DF fraction, so reducing the discrepancies between

linking the reference data for DS samples to the WU/DU/DG scans. An alternative would be to, in the future, process all samples to the more homogenous DF state enabling the analysis of the whole sample. This would require the development of new NIRS calibrations, however, and tests will need to be carried out to ensure that there are no analytical errors associated with the hydrolysis of fine samples.

- Statistically significant differences exist between many of the plant fractions for numerous constituents. There are also trends in constituent concentration with height up the stem.
- There are also differences between smaller and taller plants in the compositions of their various plant fractions.
- When these variations are coupled with the much greater proportions that leaves contribute to the total mass of small plants, it can be seen that the glucose and total sugar contents of the shorter *Miscanthus x giganteus* plants are lower than those of the taller plant. Hence, in addition to lower per hectare biomass yields, smaller plants will provide lower per tonne chemical yields in biorefining processes than taller plants would.
- While the total sugar contents of leaves are less than that of the stems, the added biomass provided by the leaves does allow for significantly (approximately 20%) improved potential biorefinery yields associated with an early harvest.
- The potential per hectare yields of second generation biofuels, or platform chemicals, from *Miscanthus* plantations can potentially be far in excess of those possible with first generation feedstocks. This, coupled with the lower energetical and monetary costs involved in their production (Hayes and Hayes, 2009, Hayes, 2008, Bullard, 2001), indicate that *Miscanthus* is an attractive feedstock for biorefining in Ireland.

17 Analysis of Waste and Other Feedstocks

In January 2008 the Author submitted to the Irish EPA a proposal for a €108,000 research project. This application was successful and the project, entitled “The laboratory analysis of Irish municipal and agricultural biomass wastes and evaluation of their utilisation in biorefining technologies”, started on December 2008. It involved the collection and subsequent wet-chemical analysis of representative samples of a wide variety of biomass wastes. The results of this comprehensive analysis would inform an evaluation of the suitability of these wastes for processing in a variety of biorefining processes.

17.1 Thermochemical Biorefining Technologies

This Chapter will compare the potential biofuel/chemical yields from processing selected feedstocks in different biorefining technologies. Hydrolysis technologies A to F, and the yields that they could provide from lignocellulosic feedstocks, have been discussed in Section 16.5. Two other technologies, based on the thermochemical platform, will be presented in this Chapter.

Unlike the hydrolysis technologies, which specifically target the structural polysaccharides of feedstocks, the thermochemical processes degrade the whole feedstock. Determinations of process yields are often based on the lower heating value (LHV) or higher heating value (HHV) of the feedstock, or on the elemental composition of the material (the relative amounts of carbon, hydrogen, oxygen and nitrogen, along with the amount of ash present, are most important). In this Chapter the dry HHVs and LHVs have been determined according to the equations in Section 2.8.

The two representative thermochemical processes used in this study are described in a paper by the Author (Hayes and Hayes, 2009), and summarised below:

Technology (G) – Synthesis of mixed alcohols via the catalytic processing of syngas derived from the gasification of biomass. The efficiency of the process is based on the LHV of the feedstock, giving a conversion efficiency of 48.8% to ethanol and 9.6% to higher alcohols (Phillips et al., 2007) although this study will only consider ethanol. Feedstocks with ash contents greater than 20% are considered unsuitable for this technology. The process is considered to be beyond the current state of the art and more likely for commercialisation closer to 2020.

Technology (H) – The Fischer-Tropsch (FT) synthesis of a mixed range of linear hydrocarbons from biomass-derived syngas. This study uses the data given by Tijmensen *et al.* (2002) for an Institute of Gas Technology, direct, oxygen blown, pressurised gasifier (Katofsky, 1993) with full gas recycle. The overall conversion efficiency, based on the LHV of the feedstock, was estimated at 47.7%, with 37.92% for FT liquids and 6.65% for net power (Tijmensen *et al.*, 2002). Hydrocracking of the waxy FT product is necessary to maximise diesel yields with these cracking conditions producing 60% (by mass) diesel, 25% kerosene and 15% naphtha (Tijmensen *et al.*, 2002). Hence, yields (according to the dry LHV) will be 24.68% for diesel, 6.33% for naphtha and 10.04% for kerosene. Despite FT-derived naphtha having a much lower octane number than normal naphtha (Tijmensen *et al.*, 2002), its production is included in energetic calculations in this study. In contrast, while the FT-kerosene product has potential for utilisation in the aviation industry (providing it receives the appropriate certification) it will not be considered as a relevant biofuel since the EU alternative-fuels mandate only considers the consumption of petrol and diesel in the transport sector (EC, 2007). Biomass with a moisture content over 60% is assumed to be unsuitable, as are feedstocks with ash contents greater than 20%.

17.2 Agricultural Wastes

There exist comprehensive surveys of the quantities of some agricultural wastes and residues in the Irish Republic (RPS MCOS, 2004, Collins *et al.*, 2005). However, chemical compositional data of relevance to Ireland are critically lacking. Where they do exist they are based on outdated analytical techniques (e.g. the detergent fibre methods (Van Soest, 1963a)). Previous Irish studies concerning these wastes and the various potential technologies for their utilisation predominately focused upon their use as combustible fuels (Williams *et al.*, 2000, van den Broek *et al.*, 2001). Similarly, most existing studies on biofuel production in Ireland focus on the first generation (Hamelinck *et al.*, 2004) or lack the necessary thoroughness to be useful reviews of biorefining technologies (Murphy and McCarthy, 2005).

The main types of agricultural wastes that the Author focused on were straws, animal manures, and spent mushroom compost (SMC).

17.2.1 Straws

17.2.1.1 Background

Straw is a general term that can cover most solid plant residues from crops. It can include oilseed rape, rye, barley, wheat, oats, beans, and peas. In Ireland, the cereal straws are the most abundant. When wheat or barley is threshed, straw is a byproduct that is typically laid back on the field during the combining process. Some may be collected and used for bedding, feed supplement, or mushroom compost production. However, much is left to rot on the field.

The yield of straw per unit mass of grain will vary according to the plant type and the local environment. However, the average yield of wheat straw is 1.3–1.4 kg per kg of grain (Montane et al., 1998). In Ireland, it has been estimated that about 1.5 million tonnes of straw are produced annually. However, there is a large variation between the years (CEN, 1999). Agricultural practices require that much of this material remains on the land for fertilisation, and so quantities available are significantly less. The mushroom industry requires 1 tonne of straw per 2.5–3 tonnes of compost (Rice, 2001), and so the current production of 290,000 t of SMC will require about 80,000 t of straw.

The chemical compositions of straws will be dependent on the relative proportions of the components of the plant (e.g. nodes, internodes etc.) and the chemical compositions of these components. The harvesting procedure is also important since it determines how well the different components are collected. For example, Ernst *et al.* (1960) found the proportions of components of baled wheat straw (by mass) to be 68.5% internodes, 20.3% leaves (Sheaths), 5.5% (leaf blades), 4.2% (nodes and fines), and 1.5% (grain and debris). In experiments where straw was harvested by hand, there was a significantly higher leaf mass of nearly 50%. The specific cultivar of the species may also be important as the relative proportions of leaves and internodes may differ between varieties and different cultivars may also lose differing quantities of leaves in the harvesting process, due to variations in the brittleness of leaves. Table 17-1 shows, with secondary data, how the type of cultivar is also important in the chemistry of plant components. For example, with “Madsen” the cellulose content in the internode is 2.1% lower than that in the node, yet with “Rod” the cellulose content is 6.6% higher in the internode than in the node.

There has been much research on characterising the wheat straw hemicelluloses, which are a significant fraction of the dry matter. Sun *et al.* (1996) found that the hemicelluloses consisted of a (1→4)-linked β -D-xylan main chain. They found a D-glucopyranosyluronic acid (or 4-O-methyl- α -D-

glucopyranosyluronic acid) group attached at position 2, and L-arabinofuranosyl and D-xylopyranosyl groups attached at position 3. For every 26 D-xylopyranosyl residues in the main chain, there was one uronic acid unit. For 13 such D-xylopyranosyl residues, there was one L-arabinofuranosyl group, and for 18 such D-xylopyranosyl residues, there was one D-xylopyranosyl group.

Table 17-1: The chemical composition (% whole mass) of the mass components of 6 American wheat varieties (Clean Washington Center, 1997).

Name	Cellulose(%)	Hemicellulose(%)	Lignin(%)	Extractives(%)	Ash(%)	Total (%)
Internodes						
Madsen	36.7	34.7	18.0	1.1	7.4	97.8
Eltan	35.7	31.2	19.5	1.1	5.7	93.2
Stephens	35.7	35.3	19.0	1.0	7.6	98.6
Lewjain	35.2	32.3	18.9	1.0	7.2	94.6
Cashup ₃	48.3	20.4	19.7	1.1	9.1	98.5
Rod	35.3	35.2	20.3	0.9	6.8	98.5
MEAN	37.8	31.5	19.2	1.0	7.3	96.9
Nodes						
Madsen	37.8	24.1	15.7	0.8	8.6	87.1
Eltan	35.6	23.1	15.8	0.7	9.8	85.0
Stephens	34.3	25.9	15.8	0.6	10.2	86.8
Lewjain	34.5	20.9	15.2	1.1	13.1	84.8
Cashup	34.8	28.7	14.4	1.0	12.7	91.5
Rod	28.7	27.8	14.8	1.0	11.1	83.4
MEAN	34.3	25.1	15.3	0.9	10.9	86.4
Leaves						
Madsen	32.6	33.4	20.8	2.4	13.	102.6
Eltan	23.1	28.8	19.7	2.8	8.7	83.0
Stephens	30.2	26.3	24.0	4.1	13.8	98.5
Lewjain	26.3	28.4	22.9	2.2	12.3	92.1
Cashup	28.5	29.0	24.2	3.1	15.1	99.9
Rod	27.0	30.0	21.1	2.8	11.1	92.0
MEAN	28.0	29.3	22.1	2.9	12.4	94.7

17.2.1.2 Sample Collection and Processing

Samples of the straws of several varieties of cereals were collected from two locations operated by the Department of Agriculture, Fisheries and Food (DAFF). These were experimental plots that were used to examine the performance of new varieties with the target that some of these could have potential for large scale commercial production. The first location was Backweston Farm in Leixlip, Co. Kildare and the second was Kildalton College in Piltown, Co. Kilkenny. These samples were collected by hand by DAFF staff and stored in sampling bags and subsequently collected by the Author and taken to the Carbolea laboratories. Table H-1 lists the samples that were collected, providing information about their species, variety type, date of collection, and location. All samples

were processed whole and no manual separation of the anatomical fractions (e.g. internodes, leaves etc.) took place, except for the rapeseed straw samples in Table H-1.

17.2.1.3 Compositional Data

Wet chemical analysis of the DS samples followed the standard method outlined in Section 11, and the results for the extractives and lignocellulosic components are presented for each sample in Table H-2 and summary statistics are provided in Table H-3 and Table H-4. These results show that, as with the sugarcane bagasse and *Miscanthus* samples, the principal constituent of the straws is glucose. This is a sugar that will primarily be present in the polysaccharide cellulose. Xylose is the second most abundant constituent, closely followed by Klason lignin. Arabinose is the next most abundant sugar but is present in concentrations approximately 10 times less than that of xylose. Galactose is present in concentrations that are typically around three times less than arabinose, whilst rhamnose and mannose are minor components, although mannose is present in slightly higher concentrations than it is in many of the *Miscanthus* samples. Among all the samples, the arabinose to xylose ratio ranged from 0.09 to 0.14 and the galactose to xylose ratio from 0.03 to 0.05. In both cases the lowest ratios were for the winter barley group and the highest ratios for the winter oats group. The inverses of these arabinose to xylose ratios (i.e. xylose to arabinose ratios) are between 7.1 and 11.2; these values are below the ratio of 13 determined in the research of Sun *et al.* (1996); however, that paper involved the analysis of Chinese straws, which may very well be expected to differ in composition from Irish samples.

Table H-4 provides data for the range in concentration values for various constituents. It can be seen that, generally, there is a low variation between different varieties of a certain species. For example, the range in glucose content for of all winter wheat samples analysed was only 0.82%. However, there is a greater variation in lignocellulosic components between different species (see Figure H-1). For example, the average glucose content of spring-oat-straw is approximately 5% higher than that of spring-wheat straw. While the variation between all samples is greater than the variation within species groups, it is still quite low. For example, the standard deviation for glucose is less than 2% and the range less than 10% for all constituents.

This limited variability should be attractive in the sense that it improves the confidence that can be associated with any predictions of potential biofuel yields. This is perhaps even true for unknown varieties of the straw types listed above since no single sample of a given species type has been

observed, in these data, to deviate greatly from the properties of the other samples of that type. This limited variation would, however, make the job of developing precise NIRS calibrations much harder. For example, using the threshold RER ratio of 15 for the development of an NIRS model suitable for the quantitative prediction of unknown samples, the RMSEP could not be greater than 0.51% for a glucose model and not greater than 0.28% for a xylose model. Achieving such low RMSEPs would be very difficult and require, as a first condition, extremely accurate reference methods. The Author has not yet attempted to develop NIRS models for straw samples, since sufficient samples have not yet been analysed to allow for a robust model; however, the limited variation in the straws does suggest that NIRS calibrations may not be warranted.

Table H-5 provides the ash and elemental data for the DS and DF fractions of selected straw samples. It also provides the HHV and LHV for the combined sample. The data show that, for every sample, the ash and nitrogen contents of the DF fraction are greater than those of the corresponding DS fraction. However, there is less relative variation in the carbon content, with the DS and DF values within 1% of each other for all but one sample. The hydrogen contents of both fractions are also similar. The differences in ash contents, however, mean that the DF fractions have slightly lower HHVs/LHVs than the DS samples. For example, the LHV of the DS fraction of STWW4 is 18.15 MJ/kg while the LHV of the DF fraction is 18.09 MJ/kg. Nevertheless, these heating values are reasonable, particularly in the context that straw samples are typically reasonably dry (less than 20% moisture) meaning that their effective heating values will be closer to those of their dry LHVs and little or no drying would be required.

The data show, however, that the ash contents of these straw samples are slightly higher than those of *Miscanthus* stems. Furthermore, according to the literature, within this ash there can be relatively high amounts of components such as chlorine and potassium that may be problematic in certain reactors and catalytic systems (Kavalov and Peteves, 2005). A potential solution to these inorganic components is to leave the straw on the field after cutting so they can be washed off by rain or to wash the straws at the biorefinery (Kavalov and Peteves, 2005), although such treatments may necessitate an additional drying step.

17.2.1.4 National Straw Quantities

Section 17.5 discusses the projected national yields from utilising a variety of lignocellulosic feedstocks in different biorefining technologies. These projections require estimates on the arisings

of each feedstock. In Ireland, the cereal straws are the most abundant (CSO, 2007). Specific data for straw production in Ireland are lacking, hence these are estimated based on data for the harvested area and yields (CSO, 2007) and on the straw/grain ratios for each feedstock (RPS MCOS, 2004, Kaltschmitt and Hartmann, 2000). A total figure of 0.96×10^6 tonnes (on a dry basis) is calculated for the maximum harvestable straw available.

A study estimated that between 80 kt and 325 kt of the total straw resource could be available for energy purposes (RPS MCOS, 2004). These figures are distributed proportionately between the various available straw types according to cereal production figures (CSO, 2007), with the total straw arisings under three scenarios presented in Table H-6. These scenarios are (i) all straw produced potentially available for biorefining; (ii) 319 kt of total straw potentially available (equivalent to 325 kt minus the estimated production of rapeseed straw) ; (iii) 78 kt of total straw potentially available (equivalent to 80 kt minus the estimated production of rapeseed straw).

17.2.2 Animal Manures

17.2.2.1 Background

Table 17-2 provides data (Crowe, 2000) for the indoor production of silages and animal manures in wet tonnes per year. The estimations of manure produced were based on animal numbers, average waste production per animal, and the length of time that animals are kept indoors. Average housing periods of 6, 20, and 26 weeks were used for sheep, cattle, and horses, respectively. For pigs and poultry it was assumed that all the slurries or litter produced are collected as wastes.

Table 17-2: Total quantities of animal and silage wastes (wet tonnes per year). Taken from Crowe (2000).

Waste Category	Quantities Arising	
	(tonnes/yr)	% of Total
Cattle manure and slurry	37,098,470	57.4
Sheep manure	338,063	0.5
Horse manure	365,310	0.6
Pig manure and slurry	2,623,350	4.1
Poultry manure	1,847,531	2.9
Silage effluent	2,684,500	4.2
Dirty water (dairy only)	19,621,500	30.4
Total	64,578,724	100

The compositions of animal wastes is a complex issue, with carbohydrate composition dependant on the class of animal, diet, digestibility of food, bedding, and stage of growth, among other factors. Table 17-3 shows the results of some of the carbohydrate analyses that were found in the literature for different livestock categories whilst Table 17-4 shows how the composition of cattle manure can vary with different diets.

Table 17-3: Major mass constituents (% DM) in some animal wastes, note that detergent methods were used and, hence, may not be totally reliable. Results from (Smith, 1973b)

Waste	Neutral Detergent Solubles	Hemicellulose	Cellulose	TOTAL SUGAR	Lignin	Ash
Broilers (caged)	69	16	11	27	4	22
Laying hens (caged)	65	17	15	32	3	28
Pigs (growing & fattening)	60	20	15	35	5	17
Beef cattle (fattening)	53	22	17	39	8	7
Dairy cattle (lactating)	41	21	25	46	13	9

Table 17-4: Relevant mass compositions (% dry matter) of cattle fed four different diets (Anthony, 1972)

Constituent	Diet			
	1	2	3	4
Dry matter	25.15	22.48	28.97	25.59
pH	4.68	4.96	4.81	5.74
Noncell wall content	47.13	45.41	51.50	60.10
Cell wall content	52.87	54.29	48.50	39.90
Ash	6.89	8.02	7.55	11.50
Monosaccharides				
Glucose	19.63	18.40	17.07	18.87
Galactose	4.31	4.20	2.75	5.48
Mannose	1.71	3.87	1.32	2.41
Arabinose	2.22	2.94	2.58	1.29
Xylose	5.41	5.27	7.41	9.18
Ribose	2.05	1.00	1.38	2.80
TOTAL SUGARS	35.33	35.86	32.51	40.03
Crude protein	13.37	16.56	16.84	20.26

Regarding the production of excreta from pigs, it was shown in a study by Henning and Poppe (1977), see Table 17-5, that the quantity of waste production is dependent on the weight of the animal - with an increase in body weight the proportional quantity of pig waste decreases significantly. That study used the same diet; lower digestibility feeds would result in more waste. Pig manure has a low dry matter content of 6-10%. Faeces represents about 46% and urine 54% of wastes on a fresh matter basis, but on a dry matter basis faeces represent 77% and urine 23% (Henning and Poppe, 1977). Pearce (1977) documented a large amount of analytical data from 24 commercial piggeries in Australia, the relevant data are presented in Table 17-6. The polysaccharide

contents were calculated using detergent fibre analysis (Van Soest and Wine, 1967). Hillard (1977) compared the chemical composition of the pig faeces to that of its feed. The results from that experiment, Table 17-7, show that lignin and cellulose were practically undigested, as demonstrated by their accumulation in faeces.

Regarding dry matter contents for Irish samples, McCutcheon (1997) analysed a range of samples of pig manure in Ireland and found the range to be 0.4% to 13.1 % with a mean of 5.1%.

Table 17-5: Quantity of pig waste according to weight of animal. Taken from Henning and Poppe (1977).

Live Weight (kg)	Quantity of pig waste per head/day	
	Kg (wet)	% of Pig Weight
41.9	3.62	8.6
59.7	4.08	6.8
89.8	4.45	5.0
128.7	4.89	3.8

Table 17-6: Mean and range values for important mass constituents (% dry matter) in the composition of pig faeces. Taken from (Pearce, 1977)

Constituent	Mean	Range
Cellulose	17	6 – 23
Hemicellulose	20	3 – 36
Lignin	5	3 – 6
Crude protein	19	11 – 31
Crude fibre	18	7 – 23
Ash	17	10 – 28

Table 17-7: Composition (% dry matter) of pig feed and pig faeces. Taken from (Hillard, 1977).

Constituents	Feeds	Faeces	Index (Feed = 100)
Hemicellulose	13.8	20.3	147
Cellulose	5.2	16.9	325
Lignin	1.1	4.9	445
Ash	6.7	17.4	259

17.2.2.2 Sample Collection and Processing

On March 5th 2009 several samples of dairying cattle slurry were collected from the Teagasc Johnstown Castle Research centre in County Wexford. Each sample was collected from a different storage house. The labels given to these samples, and the different conditions they represent, are outlined below:

- MNDY1 = 24 milking dairy cows, fed a mixture of grass silage and concentrated feed. Slurry was in storage since 14/10/08 and last agitated on 2/3/09.
- MNDY2 = 24 milking dairy cows, fed a mixture of 40% grass silage, 20% maize silage and 40% concentrated feed. Slurry was in storage since 14/10/08 and last agitated on 2/3/09.
- MNDY3 = 24 milking dairy cows, fed a mixture of 34% grass silage, 17% maize silage and 49% concentrated feed. Slurry was in storage since 14/10/08 and last agitated on 2/3/09.
- MNDY4 = Calf house. Last agitated on 20/2/09.
- MNDY6 = Fattening Dry cows. Fed grass silage and concentrated feed. Slurry was in storage from 20/10/08, last agitated on 2/3/09.

On 3/6/09 several samples of pig slurry (again each representing variations in livestock class and/or diet) were collected from the Teagasc research centre in Moorepark, Co. Cork.

Drying and sample preparation methods took place quickly after sample collection in order to prevent the degradation of the samples. Hence, the dairy slurry samples were prepared first. As shown by the results in Table H-7, the moisture contents of these samples were very high (up to 97% on a wet basis) and require a different protocol for drying than the standard air-drying method outlined in Section 11.1. Instead they were placed on trays and oven dried at 60°C. However, it was found that this took over one week in some instances and the smells caused problems for laboratory workers and adjacent facilities. A different drying approach, freeze drying, was therefore employed for the piggery slurry samples that were collected.

Regarding particle size comminution, initially the FOSS cyclotec mill with a sieve aperture of 1 mm was used for the dairy manure samples. However, this approach resulted in the DF fraction being greater in weight than the DS fraction and also resulted in the production of significant quantities of dust in the lab. As a result, for the pig manure samples, the dried excreta were comminuted by hand (initially) and by mortar and pestel (later), with numerous sieving steps in between in order to try and maximise the DS fraction.

17.2.2.3 Compositional Data

Table H-7 presents the ash, extractives, and lignocellulosic data for the manure samples that were analysed. It can be seen that there is a significant variation both within and between groups. For example, the total sugars content varies from 3.27% (MNPG5) to 46.19% (MNPG1) for the pig

slurries and from 18.53% (MNDY4) to 39.31% (MNDY1) for the dairy cattle slurries. For both groups the lowest total sugar samples came from the young animals that were being reared; principally on a milk diet. Table H-9 presents the sugar data for each sample as a percentage of the total sugars content and also provides a hemicellulose to cellulose ratio for each sample. This allows for the differences in the sugar compositions within and between the groups to be seen more clearly. Some important points are listed below.

- Hemicellulose is the major polysaccharide in the pig slurries with a content that is up to 89% greater than cellulose. In contrast, cellulose is the major polysaccharide in dairy slurries. This is in agreement with the literature (Pearce, 1977, Smith, 1973b).
- The arabinose contents are proportionately greater in the pig slurries.
- The slurries from young animals of both the dairy (MNDY4) and pig (MNPG5) classes show similar differences compared against the other samples in their groups:
 - Proportionately greater concentrations of galactose, rhamnose and mannose.
 - The highest hemicellulose to cellulose ratios of their animal type.
 - The lowest proportionate glucose contents.
 - The greatest AIA contents.
- Sample MNPG5 has a significantly lower KL content (8.92%) than all other samples.
- Ash contents are high in all cases meaning that these manures can not be processed in the thermochemical technologies.
- The (95% ethanol-soluble) extractives contents are, on average, lower in the dairy manure samples than the pig manure samples. However, this may be a result of the oven drying procedure that was used for the dairy samples resulting in the loss of some of the volatile extractives (see Section 3.4.3).
- Extractives are a major component (over one third of total mass) of MNPG5.

The pig slurry sample with the greatest total sugars content (MNPG1, 46.19%) could be considered to be a reasonably attractive feedstock for utilisation in hydrolysis technologies given that close to 50% of the sample could potentially yield chemical products and that this is a waste feedstock which could be received free or for a gate fee. The high moisture contents of these samples, however, will mean that their use in any hydrolytic biorefining process will probably be as a mixing agent for a drier feedstock (e.g. straw and waste wood). There would also be the question of the transportation costs involved for samples with such high moisture contents.

The number of pigs in Ireland has been steadily growing since the 1970s and in December 2009 there were a total of 1,602,100 pigs in Ireland (CSO, 2009). There has also been a trend towards a lower number of pig-holdings with higher average herd sizes (CSO, 2002). In 1997 there were a total of 2,000 holdings with pigs, compared with 35,700 in 1973, with an average size of herd of 858.5, compared with 29.0 in 1973 (CSO, 2002). This shift towards larger herd sizes is beneficial as far as issues concerning collection and transport are concerned.

If it is assumed that the mean weight of these pigs is 59.7 kg and that the quantity of waste produced per pig is 4.08 wet kg per day (Table 17-5), this would result in a daily production of pig manure at an average sized farm of approximately 3.5 wet tonnes per day, which is equivalent to 175 dry kg per day (assuming a 5% dry matter content). Clearly, several hundred such farms would be required to provide sufficient dry matter for a biorefinery, even if the feedstock was only used as a co-feed with other, less wet, materials. If solutions could be found to make the transportation costs manageable then perhaps pig slurry can be a viable feedstock although this will also depend on the expected composition of the excreta since, as Table H-7 shows, this varied significantly between the samples. As a first principle, the farms from which slurries were sourced would need to be in close proximity in order to reduce these logistical problems and costs. However, it should be noted that the total national resource of pig slurry would only be approximately 6500 wet tonnes per day, equivalent to only 325 dry tonnes at 95% moisture, meaning that this feedstock could only be a relatively minor component in a multi-feedstock commercially-sized biorefinery.

While pigs are housed throughout the year, cattle are not. The National Waste Database 1998 (Crowe, 2000) assumed that cattle were indoors for 20 weeks of the year in calculating the total arisings of 37 m wet tonnes of cattle slurry. Assuming a moisture content of 93% this represents approximately 7,900 dry tonnes per day of feedstock over these 20 weeks. This is a sizeable resource but the number is theoretical and not practical. As with the calculations for pig excreta this national number is much less relevant than the resource that would need to be available in a reasonably small catchment to enable the collection and transport of this waste to be practical and economical. Excluding the sample of calf excreta (MNDY4) the total sugars contents of the dairy samples that were analysed by the Author varied less than those of the (mature) pig samples (a range of 6.09% compared with 24.98%) but the maximum content was less than that for the pig samples. It is questionable whether feedstocks, even if received for a gate fee, with only approximately a third of the total mass balance being sugars could be economically viable for processing in hydrolysis technologies.

Table H-8 presents the results for the elemental analysis of the DF and DS fractions of two samples of pig excreta and one sample of dairy cattle excreta. The LHVs and HHVs are also presented for the combined (DS + DF) samples. However, the high moisture and ash contents of these samples will prohibit their use in technologies G and H.

17.2.3 Spent Mushroom Compost

17.2.3.1 Background

Spent mushroom compost (SMC) is the substrate remaining after mushroom production, with approximately 5 kg of SMC produced for each kg of mushrooms (An Bord Glas, 1996). In 2001 290 kt (wet basis, or 85 kt on a dry basis) of SMC were produced per annum (Teagasc, 2002), with 96% coming from the Border counties (Williams et al., 2001) (including 70 kt, wet basis, from Co. Monaghan (Williams et al., 2000)). Importantly, SMC is produced all year round, and hence predictability of supply should be high and seasonality issues avoided. Also, the shift that has been seen over time to larger producers concentrated in a relatively small area is an advantage when organising supply logistics for a biorefinery; transport costs should be relatively low.

Mushroom compost is a mixture of 60-70% straw, 28-34% poultry litter, and 2-4.5% gypsum (CEN, 1999). It is made in a series of stages, termed phases. In the first phase the components (e.g. straw, litter, gypsum) are mixed and then placed in long windrows for a period of up to 2 weeks with the resulting product being termed Phase I compost. The second phase takes up to 18 days and takes place indoors in plastic tunnels that allow for the environment to be controlled so that any unwanted organisms or diseases in the compost can be controlled. Once the compost is of a quality suitable for mushroom production the compost is mixed with "spawn", a monoculture of mushroom mycelium on grain (Jordan et al., 2008). This compost is termed Phase II. Phase III involves the spawning and growth of the mycelium and takes place under controlled conditions. It is considered complete when the mycelia have fully colonised the compost.

Mushroom producers either receive Phase II or Phase III composts. Once the compost is fully colonised mushroom production involves placing a casing layer of peat on top of the compost. This layer promotes the formation of promordia; mushroom pins. Approximately three weeks after this point the first crop (first flush) of mushrooms can be harvested. The compost can then be rewet allowing for the harvesting of subsequent flushes at approximately 7 day intervals. Typically, in

Ireland, up to three flushes are harvested from each compost shipment. The remaining material is known as SMC and can sometimes be sterilised (“cooked out”) by heating for 12 hours at 70°C.

There are serious issues concerning the disposal of SMC. The compost is considered an unwanted waste by most producers. SMC can be disposed by a contractor at a charge of approximately €10 per tonne; a study found that 72% of all SMC in Ireland is applied to land (Maher, 1993). There are problems associated with this use, however. A survey (Teagasc, 1994) in Co. Monaghan reported significant deterioration of the four main cross border rivers in the area. It was also found that the amount of phosphorus being generated by local agriculture was well in excess of the County's nutrient capacity.

The overall composition of SMC will vary according to the time of year, the amount of peat casing put on by the grower, the compost manufacturers, and the amount of water added to the mushroom by the grower. The chemical composition of the ultimate spent material will be significantly different from the composite of the materials that make up the mushroom compost and casing layer, however, due to the effects of the composting process and mushroom growth.

The average moisture content of SMC has been measured as 65%, and the volatiles and ash contents, on a dry basis, as 61% and 39%, respectively (Maher, 1993). There has been some research on the use of Irish SMC for energy generation (Williams et al., 2001). The high moisture content, however, indicates that combustion of this feedstock would not be an effective utilisation of the material – the HHV of SMC is around 12.2 MJ/kg on a dry basis, but the moisture content results in the effective heating value being around 2.4 MJ/kg (CEN, 1999).

While there are several sources detailing the moisture content and calorific value of SMC, there are less data on carbohydrate and polysaccharide contents (Van Lier et al., 1994, Sharma, 1996, Jordan et al., 2004). Jordan *et al.* (2008) carried out an examination of the composition of SMC in Ireland in order to determine its suitability as a fertiliser. Their results are presented in Table 17-8. Table 17-9 also presents the results of an unpublished survey of the chemical properties of SMC that was carried out on 20 samples at the Teagasc Kinsealy Research Centre in July 2003 (Michael Maher, personal communication). If the mean cellulose and hemicellulose contents determined by Jordan *et al.* (2008) are reflective of the real polysaccharide composition of SMC then this represents an attractive feedstock for biorefining given that it would be approximately 57% carbohydrate and potentially provided to the biorefinery for a gate fee. Unfortunately, both these studies used detergent fibre methods. These indirect methods of analysis will carry a high degree of uncertainty (see Section 3.1.1), particularly given the complex degraded nature of these samples.

Table 17-8: The composition of SMC taken from various producers in the Republic and Northern Ireland. Data are in % dry matter. Taken from (Jordan et al., 2004)

Parameter	Min	Max	Mean	SD
Cellulose	18	62	38	8.6
Hemicellulose	2	41	19	8.71
Lignin	11	49	25	9.5
Organic matter	41	76	64	5.9

Table 17-9: Data on the composition of SMC (% dry matter), from the analysis of 20 samples by Michael Maher (2003 – unpublished)

Parameter	Min	Max	Mean	S.D.
Dry Matter (% wet basis)	27.8	38.7	38.7	1.13
Water soluble carbohydrates	1.7	2.9	2.2	0.15
Acid Detergent Fibre	34.5	47.7	41.6	1.50
Lignin	16.2	21.5	19.1	0.80
Ash			33.4	

17.2.3.2 Sample Collection and Processing

It was planned that samples of both the mushroom compost and the SMC would be collected and analysed in order to see the changes that occur during the composting and mushroom production stages. On 2/9/09 the Author visited Monaghan Mushrooms, the largest mushroom compost supplier in Europe. The site of compost production is Tyholland in Co. Monaghan. The samples collected from this site are detailed below.

- MCPR6 = Day 1 of Phase I compost production. This sample represents the mixture of the component fractions (wheat straw, poultry litter, gypsum) prior to the stacking of this mixture in windrows. It is therefore the compost feedstock before the composting process. Correspondingly, it was of a much larger particle size than the subsequent samples.
- MCPR5 = Phase I compost, 14 days old compost. This had a lower particle size than MCPR6 but larger than MCPR4, MCPR8, MCPR3.
- MCPR4 = Phase II compost without mycelium.
- MCPR8 = Phase II with mycelium added.
- MCPR3 = Phase III compost.

After these samples were collected the Author went to a Monaghan mushrooms farm, located in Northern Ireland, and obtained a sample of SMC (MCSP3) that had two mushroom harvests (two flushes) and had been “cooked out”. The Author also visited Carbury mushrooms in Carbury, Co.

Kildare and obtained 2 samples of 3-flush SMC (MCSP6, MCSP8). Kiernan Mushrooms in Co. Cavan was also visited and a sample of 2-flush SMC (MCSP7) collected.

These samples were reduced in particle size using the Retsch SM2000 chipper in order to obtain the DS and DF fractions.

17.2.3.3 Compositional Data

Table H-10 shows the ash, extractives, and lignocellulosic compositions of mushroom composts at various stages of preparation. It can be seen that there is a significant drop in total carbohydrate content from the first day to Phase II/III compost with a reduction in total sugars content of around 30%. All of the major constituent sugars experience a drop over this period. Table H-12 presents the sugar data for each sample as a percentage of the total sugars content and also provides a hemicellulose to cellulose ratio for each sample. It can be seen that the proportions of the different sugars do not change greatly from day 1 (MCPR6) to the final phase III compost (MCPR3). It would perhaps be expected that hemicellulose would be more readily degraded than cellulose and that this would be demonstrated by a fall in the hemicellulose to cellulose ratio; however, this only drops from 0.63 for MCPR6 to 0.60 for MCPR3, and so it appears that the composting process is almost as effective on cellulose. The minor sugars mannose and rhamnose do experience greater relative changes over the period, with their proportions of total sugars increasing. This is likely to be a result of the increased microbial mass in the sample after composting. As the proportionate contributions of the sugars decrease with time in the composting process, the quantities of KL and, particularly, ash increase.

Regarding the SMC, Table H-10 shows that the total sugars content falls significantly further, down to as low as 18.1% for sample MCSP8. Sample MCSP3 is the SMC that results after two flushes of mushrooms have been harvested from a compost that is equivalent to MCPR3 (with a layer of peat added). There is a 17.8% drop in total sugars from MCPR3 to MCSP3. A drop in the concentrations of some sugars is to be expected since these are utilised by the mushrooms for their growth. It is interesting to note that the proportion that each sugar contributes to the total sugars content differs between MCSP3 and MCPR3. While the glucose content is roughly similar, the relative xylose content decreases by over 7%. In contrast, the proportions that rhamnose, mannose, and galactose contribute to the total sugars content are greater for the SMC samples than for the mushroom samples. Indeed, there are whole mass increases in the mannose, rhamnose and galactose contents.

It is possible that this is a result of the contribution of peat to the SMC sample as well as the preponderance of some of these sugars in microbial communities. The greater relative decrease in the xylose content compared with the glucose content confirms the reports from another study (Adamovic et al., 1998) which noted that the extracellular enzymes of the mycelium degrade hemicellulose at a faster rate than cellulose. MCSP6 and MCSP8 are 3-flush composts while MCSP3 and MCSP7 are 2-flush composts. It would be logical that the total sugars content of the SMC would decrease with an increasing number of flushes. However, this is not observed in these samples since, although MCSP8 has the lowest total sugars content, MCSP6 has the highest. Since these composts were obtained from different suppliers, however, making judgements about the effect of flush number is not possible since the growing conditions and composition of compost and peat casing could vary. A more detailed study could involve the sampling of SMC at several stages in one site, each stage corresponding to flush number. However, it is questionable, given the low total sugar contents, as to whether further research is warranted, at least under the concept of studying feedstocks for potential utilisation in commercial biorefining schemes.

Table H-10 shows that, while the sugars content of SMC is less than that of the mushroom compost, the KL content is greater. The change in KL, glucose, xylose, and total ash contents, is presented in Figure H-2, which represents all of the Monaghan Mushroom samples. An increased KL content may suggest an improved heating value and hence potential for utilisation in thermochemical biorefining schemes. However, in this case the ash content also increases. Table H-11 presents the C, H, N, S, and ash data for the DS and DF fractions and for the whole sample. It can be seen that the combined ash content of the SMC samples is very high - over a third for all of the samples with DS and DF data. Even without considering the technological problems that high ash content feedstocks bring to thermochemical processes, the situation regarding dealing with the post-treatment waste (approximately 1 tonne of ash for every 3 dry tonnes of sample) would be a great concern. The high moisture contents of this feedstock (over 65%) would, in any case, preclude SMC from being a practical, in energy terms, feedstock for thermochemical processing.

These data therefore suggest that SMC is not an attractive feedstock for any of the biorefining technologies under consideration. This is in contrast to what a literature review of the secondary data suggested (Hayes and Hayes, 2009). However, the data that informed that previous evaluation by the Author were critically flawed. There can be a much greater degree of confidence that the data presented in this chapter represent a much more accurate portrayal of the real composition of SMC and the chemical changes that take place in its life-cycle. The question still stands, therefore, on how SMC can be dealt with. A possible alternative to the technologies discussed in this Thesis is the slow

pyrolysis of the feedstock for the production of a biochar. This biochar can then be added to the land with the aim of improving plant growth production (Hayes, 2006). Since the ash will be incorporated into the final saleable product, the ash content will not be considered a waste. While the high moisture content of the SMC may result in a negative energy balance for the biochar production scheme, it is possible that the value of the end-product could still allow for the process to be economically viable. Whether this will be the case will depend on the quality and suitability of the SMC biochar as a plant growth medium; the evaluation of biochar properties and plant growth potential are areas being explored by colleagues in the Carbolea group.

17.3 Municipal Wastes

It has been estimated that, in 2005, a total of 3,050,052 tonnes (wet basis) of municipal solid waste (MSW) were generated in Ireland (EPA, 2006). Approximately 72% of this was biodegradable, Figure H-3 categorises the biodegradable municipal waste (BMW) fraction of this and revises the figures to dry tonnes according to the moisture contents provided by Porter (2007). BMW is composed of wood, various papers and cardboards, and organics (all other BMW material; principally food and garden wastes).

The Landfill Directive (EC, 1999) sets progressive targets to reduce the amount of BMW land-filled, when compared against the BMW produced in the baseline year of 1995. The target for 2016 for Ireland is for a maximum of 35% of the quantity of BMW generated in 1995 to be land-filled. Hence, approximately 1.8 m tonnes of BMW will need to be diverted from landfill (Roche, 2006). One of the currently favoured methods for reducing the quantity of BMW and all municipal solid waste is incineration. It is forecasted that the incinerator planned for Poolbeg (which will process wet 600,000 tonnes of MSW of which 225,300 dry tonnes are expected to be BMW) will produce only a relatively small output of electricity. It is appropriate to consider alternative waste treatment technologies that may offer higher revenue streams. Biorefineries are one such option (Hamelinck et al., 2005).

There has been highly comprehensive work already conducted in terms of the characterisation, by mass, of the various waste fractions in the Irish household and non-household (commercial/industrial) sectors (Gaillet et al., 2005) according to a modification of the standard EPA "Waste Characterisation Methodology" (EPA, 1996). Household surveys were carried out in nine local authority areas, and included a total of 37 separate survey events. These were distributed

between cities, towns, and rural areas, and between 2-bin and 3-bin collection systems proportionately according to the contributions made by these to the national household waste mix. The end results of this analysis were estimated figures for the proportion, by mass, of various waste categories to the total amount of household waste according to each bin type (e.g. the black bin, the recyclables bin, and the brown bin). These data are present for all local authority areas. There are also similar data available for non-households.

The categories provided in that report that have relevance to potential utilisation in lignocellulosic biorefining technologies are outlined below along with the example materials (also provided by the report) for each category:

- Papers:
 - Packaging – e.g. Brown or white paper bags, wrapping paper, fast food wrapping, egg cartons.
 - Newspapers/brochures – e.g. Local and national newspapers, newsprint-type advertising publications, other newsprint.
 - Magazines and glossy paper – e.g. Magazines and ads on glossy paper, shop catalogues.
 - Office papers – e.g. envelopes, letters, print-outs.
 - Tissue papers.
 - Other papers – e.g. Till receipts, books, telephone directories, Golden Pages, non-glossy junk mail, loose leaf paper, non-glossy brochures and catalogues.
- Cardboards:
 - Flat packaging – e.g. Cornflake boxes, toy boxes, washing powder, containers, food containers, cleaning product cartons.
 - Corrugated packaging board – e.g. Corrugated packaging cardboard used for household items packaging (TVs, PCs, furniture).
 - Other cardboards – e.g. Birthday cards, postcards, files and folders, tickets.
- Composites:
 - Liquids packaging – Beverage cartons (Tetrapak)
- Wood:
 - Wood packaging – e.g. Bottle corks, cork packaging, pallets, wine presentation boxes.
 - Non-packaging wood – e.g. Wood fencing, wood from DIY, kitchen units, particle wood (chipboard, plywood, MDF).

- Organic Waste:
 - Biodegradable kitchen and canteen waste – e.g. Bread, fruit, cooked or uncooked food items, meat and fish, pet foods, vegetable skins, tea bags.
 - Biodegradable waste from garden and park – e.g. Grass and bush cutting, twigs, soil, flowers, leaves, tree branches, weeds.

Each of the subcategories in the list above have the mass proportions that they contribute to each bin in the survey. The Author sampled, processed and analysed many of the example materials listed above in order to gain an understanding regarding the variability in compositions of these wastes.

17.3.1 Paper and Cardboard Wastes

The samples of paper and cardboard wastes that have been collected, processed, and analysed by the Author as part of the EPA project are detailed below:

- PCWC1 = fast food wrapping (cardboard). A combined sample of a Big Mac box, a french fries box, and a happy meal box (all from McDonalds).
- PCBB1 = Brown paper bags. These comprised shopping bags and magazine holder bags.
- PCWB1 = White paper bags. These are the kind of paper shopping bags that are provided by retailers (to avoid the plastic bag levy).
- PCEN1 = A mixture of various mail envelopes.
- PCGP1 = Glossy paper (flyers etc.). This mixed sample comprised one-sheet pieces of glossy paper and booklets (e.g. a brochure for DELL computers).
- PCPF1 = Food packaging cardboard boxes. Several items were mixed (e.g. a pizza box).
- PCCB1 = Breakfast cereal boxes. Included a cornflakes box and a Weetabix box.
- PCBC1 = Birthday cards.
- PCTR1 = Till receipts.
- PCPP1 = Printouts from an office printer.
- PCNP1 = A newspaper, the Irish Independent. This sample comprised just the main paper and none of the subsections that might have been made of more glossy materials.
- PCTP1 = Tetrapak. Juice cartons.
- PCCK1 = Wine corks (only elemental/ash data are provided).
- PCPD1 = Phone directory (only elemental/ash data are provided).

Table H-13 provides the extractives, ash, and lignocellulosic data for the DS fractions of these samples, whilst Table H-14 presents, for selected samples, the ash, elemental, and heating value (DS+DF only) data for the DS, DF, and combined (DS+DF) fractions. It can be seen that many of these samples have high total sugars contents, with values ranging from 65.3% (for the fast food wrapping) to 94.54% (for the Tetrapak cartons). That suggests that these samples could be excellent feedstocks for hydrolysis biorefining technologies, although the CaCO₃ filler in office paper may affect acid-catalysed hydrolysis. Amongst the sugars, glucose is by far the most abundant, usually followed by xylose. Mannose is a much more important sugar, in proportional mass terms, in these samples than it is in many of the samples presented so far (e.g. sugarcane bagasse, Miscanthus, straws). Indeed, it is the second most abundant sugar in the newspaper and white-bag samples. This is a reflection of the paper coming from woody materials rather than herbaceous feedstocks, a fact that is also responsible for the arabinose and xylose contents being lower than seen for Miscanthus and bagasse samples.

In softwoods galactoglucomannan is the principal hemicellulose component and constitutes around 20% of the dry weight (Lundqvist et al., 2003). The glucose to mannose ratio is about 1:3, whereas the ratio of galactose to glucose can vary from 1:1 to 1:10 (Hakkila, 1989). Arabinoxylans are also present in softwoods but at lower quantities (Casey, 1980). In hardwoods, the xylans are the principal hemicellulose, but the type of xylan found is typically a glucuronoxylan rather than an arabinoxylan. The concentration of this xylan in hardwoods varies between 15 and 30% by weight (Sjostrom, 1981). In a few species, for example in some birches, the xylan content can reach as high as 35% (Hakkila, 1989).

Paper made from hardwoods tends to be smoother (and therefore easier to write or print on) than paper from softwoods, and softwood paper also tends to be weaker. The fibres from hardwoods and softwoods can be blended into a single paper type, with the proportional quantities varying according to preferences for strength, whiteness, roughness etc.

The type and severity of pulping used will determine the amount of lignin that is retained in the final paper. Table H-13 shows that the KL content of the paper/cardboard samples varies greatly (from 1% for printouts to 26.4% for newspapers). Given that the ASL and extractives contents of the samples are relatively small, the only other major contributor to the total mass balance, apart from the sugars and KL, is the ash content. This also varies substantially, Table H-14, from 1.3% for the Tetrapak sample to 34.9% for the glossy paper sample. Under a situation where the ash content is constant, for a given particle size fraction, between two samples, e.g. PCGP1 and PCWC1, an increase in the KL content will usually be accompanied by a decrease in the total sugars content. This

would mean that the relative advantage provided by hydrolysis or thermochemical biorefining processes would shift since the KL has proportionately higher carbon, and proportionately lower oxygen, contents than the polysaccharides (see Section 2.8). This is shown in the case of the carbon contents of the DS fractions of these two samples (Table H-14). However, even in instances where the KL content is greater, some of these paper/cardboard samples are not ideal feedstocks for thermochemical processing due to their high ash contents and since the paper additives may cause problems in some gasification schemes. However, there are exceptions. For instance, PCCK1 (wine corks) has a high carbon content and a low ash content meaning that its LHV is significantly greater than the other samples.

Table H-14 also shows that there are often quite significant differences in elemental and ash compositions between the DS and DF fractions. For all samples the ash contents of the DF fractions are greater than those of the corresponding DS fraction. In the case of the glossy paper sample (PCGP1) the ash content of the DS sample was 19.5% but that of the DF sample was 51.9%. This sample also had the lowest DS proportion of all the paper samples analysed.

Processing these paper and cardboard samples was difficult. The samples were torn up by hand and then passed through the FOSS Cyclotec mill. This process was observed to significantly reduce the particle size of the material, with a lot of (uncollectable) dust also produced. Sieving the output of the mill was problematic since the particles tended to agglomerate preventing a lot of DF-sized particle from passing through the sieve apertures. Extra care had to be taken in the hydrolysis stage to ensure that sample particle agglomeration was prevented otherwise the acid would not reach all of the material and analytical precision would be poor. The differences seen in DS/DF elemental and ash compositions suggest that an extractives/lignocellulosic analysis of the DF fractions is warranted in order to produce weighted average constituent values that will be more representative of the starting material.

17.3.1.1 Arisings

Approximately 67% of the recovered BMW in 2005 was paper and cardboard; however only 2.6% of this was recycled in Ireland, a drop from 31% in 2004 due to the closure of a paper mill (EPA, 2006). Ireland exported 387 kt (on an oven dry basis) of collected paper/cardboard, with 39.3% sent to the UK and 27.1% sent to Asia (EPA, 2006). The specific make-up of this paper/cardboard stream is not

known; however it can be calculated from the data that 62.4% of it came from the commercial sector, with 37.6% coming from households.

A paper by the Author (Hayes and Hayes, 2009), using secondary chemical data, assumed that 50% of all the paper/cardboard from the commercial sector was chemically and energetically equivalent to office paper, that 40% was equivalent to cardboard and 10% to newspaper. The proportions were shared equally between these three for the household sector. In this Thesis these proportions are maintained and office paper is given the composition of sample PCPP1, newspaper the (lignocellulosic) composition of PCNP1, and cardboard the composition of PCPF1. The weighted averages of these samples result in the compositional values provided in Table H-13 and Table H-14 for the “Household”, “Commercial” and “Export” (62.4% commercial and 37.6% household) categories.

17.3.2 Garden Wastes and Composts

The Irish survey that quantified the various fractions of household and commercial sector wastes (Gaillot et al., 2005) had a category for “organics” that was broken up into “biodegradable kitchen and canteen waste” and “biodegradable waste from garden and park”. The former subcategory will be discussed in Section 17.3.3, whilst the latter will be discussed here.

On 26/8/09 the Author visited a local composting facility in Mungret, Co. Limerick. At this facility compost is produced from garden waste that has been delivered by members of the local community. The operators of the facility informed the Author that the facility receives approximately 3 wet tonnes of material (garden wastes, gardeners wastes, council garden wastes etc.) per day in summer months. This garden waste is put through a chipper and then piled in mounds where the composting process takes a total of 16 weeks. The piles are turned approximately every 2-3 days during this period. After 16 weeks the compost is put through another chipping/screening device. The screened material is considered to be suitable for addition to soil.

Compost samples were collected from several piles, representing different stages in the composting process. Sample of the garden waste that had just been delivered to the facility were also collected in several boxes and returned to the Carbolea laboratories. The contents of these boxes were then studied for their principal fractions and these were partitioned for separate analysis. The samples collected at this site are described below:

- COMX1 = Mixed garden waste that has just been chipped and not yet started to decompose.
- COMX2 = Mixed garden waste after 4 weeks of composting.
- COMX3 = Mixed garden waste after 8 weeks of composting.
- COMX4 = Mixed garden waste after 16 weeks of composting.
- COMX5 = Screened (i.e. fine particle size) 16-week compost.
- GRMX1 = Non-composted grass that had been delivered that day to the Mungret facility.
- GRMX2 = Recently-cut verge-grass (roadside) from a nearby location.
- TREG1BL = A branch of Lawson's Cypress, *Chamaecyparis lawsoniana*, an evergreen tree.
- TREG1LL = Lawson's Cypress leaves.
- BUVR1LD = Leaves of Griselinia, an evergreen bush.
- BUVR3LD = The leaves of a holly, *ilex*, variety that produced small holly leaves.
- BUVR2TD = The woody part (twigs) of the holly plant.
- BUVR5LD = The leaves of an ivy (*ivies*) plant.
- BUVR8TD = The woody part of a Pyracantha plant.

Table H-15 provides the extractives, ash, and lignocellulosic data for the DS fractions of all of these samples, whilst Table H-16 presents the ash, elemental, and heating value (DS+DF only) data for the DS, DF, and combined (DS+DF) fractions of samples COMX1, COMX3, COMX4, GRMX1, and GRMX2. Table H-15 shows that the total sugars contents are greatest for the woody fractions of the plants (samples BUVR2TD, BUVR8TD, and TREG1BL) and for the second grass sample (GRMX2). The major sugar in sample TREG1BL is glucose (32.5%) followed by xylose (6.9%), mannose (6.5%), galactose (5.0%) and arabinose (1.6%). These proportions make sense given that Lawson's cypress is a softwood so, after cellulose, the major polysaccharide will be a galactoglucomannan with lower quantities of an arabinoxylan, as discussed in Section 17.3.1. This sample also has a large KL content of 31.7%.

Sample GRMX2 differs significantly in composition (lower extractives, higher KL, higher sugars contents) from sample GMRX1. This is because GRMX2 was a much taller and more mature roadside grass than the standard garden-lawn cuttings of GRMX1.

The leaves samples, particularly TREG1LL, BUVR1LD, and BUVR5LD, have low total sugars contents that suggest that these are not attractive feedstocks for hydrolysis biorefining technologies. The 95% ethanol-soluble extractives contents of these samples also tend to be high (e.g. 21.9% for BUVR1LD, Griselinia leaves) and the total mass balance of all the analysed components low (e.g. 71.5% for sample BUVR1LD). It is likely that the remainder of the mass balance is provided by extractives

components that are not soluble in 95% ethanol (e.g. hot water extractives). These other extractives could possibly be collected in future experiments and analysed to determine if these can offer an additional supply of carbohydrates.

Leaf sample BUVR3LD (holly leaves) has higher KL and total sugars contents than the other leaf samples; however these contents are still probably too low to encourage the use of such a material in a hydrolysis biorefining technology.

Sample COMX1 was taken from a pile of recently chopped garden waste. It can be considered to be a representative composite of the different garden wastes that had been supplied to the facility and therefore, perhaps, most indicative of what the composition of biorefinery-gate, non-composted, garden waste would be. The total sugars content of this sample is disappointingly low and suggests that such a resource is not an attractive feedstock for hydrolysis biorefineries.

Consideration needs to be given to the month of collection of these samples, August. In the summer the majority of garden wastes are likely to be of grasses (lawn cuttings) or the cuttings of bushes (small twigs and leaves) rather than bulky woody materials. Given the observations seen in Table H-15 regarding the compositions of the individual leaf/grass samples it is perhaps unsurprising that the total sugars content of COMX1 is low. In the winter months the relative proportion of woody materials might be expected to increase. However, in that period the total amount of garden waste material received at the facility would be expected to be significantly lower than in the summer months. Hence, while using samples collected in August to assess the suitability of garden wastes for biorefining schemes is unlikely to be fully representative of the year-round situation, it should be more reflective of the approximate compositions of the modal (in terms of waste arisings) months. Of course, monthly sample collection and analysis would be preferable, but there was not the available time for this given all the other work that was being done.

Regarding the changes seen over the 16-week composting period, there are no clear trends with regard to the total sugars content; with the exception of sample COMX3 it decreases with time but sample COMX3 (8 weeks of composting) has the highest total sugars content of all the compost samples. A possible explanation for the outlying values seen for COMX3 is that there was a significant change, compared with the other periods, in the type of material that was placed in this compost pile.

The trend is somewhat clearer with the KL content; it consistently increases with time, rising from 27.4% in COMX1 to 33.6% in COMX4 (16 weeks); a relative increase of 22.6%. The relative increase in the ash content seen over this period is much greater. This increase in ash is not attractive for

thermochemical biorefining technologies, nor is the decrease in carbon and hydrogen contents seen from COMX1 to COMX4, Table H-16. COMX5 (the screened, fine particle size, 16-week compost) was not analysed for its elements but was characterised for ash, extractives and lignocellulosic components, Table H-15. It can be seen that the ash content is significantly greater and the total sugars content significantly less than sample COMX4, indicating that this fine fraction of the compost, while it may be suitable for soil amendment, is a very poor feedstock for biorefining. According to these data there appears to be no benefit, in terms of the properties relevant for biorefining technologies, in the composting of these garden wastes.

17.3.3 Brown Bin Wastes

As part of a research project, several samples of brown bin waste (i.e. food and garden waste) were collected, processed, and analysed in order to provide information, to a large company working in this area, as to whether this resource had potential as a biorefining feedstock. These samples were taken in two phases. In the first period four samples obtained from the waste processing facility in February 2009 were taken. Two of these, labelled MSW2 and MSW3, were of uncomposted organic waste and two, labelled MSW1 and MSW4, were of composted organic waste. These samples were delivered directly to the Carbolea laboratories from the waste processing facility. The Author had no involvement in the sampling methodology for these materials.

Upon observation of these samples it was clear that these had been taken directly from the facility with no means of homogenisation employed. This was considered to be an issue that would limit how representative the results of their analysis would be for that waste type as a whole, particularly given that the quantities of each sample (approximately 2 kg wet) were relatively small. For instance, one of the samples had a fully intact orange and the other a large carrot - these contributed to the total mass balance of the samples to a much greater degree than they would be expected to for the total organic-waste stream.

In May 2009 the Author visited the waste processing facility in order to collect more samples. At this location there is a vertical composting unit (VCU) that takes “brown-bin” organic waste that has been mixed with chipped wood waste (predominately composed of waste pallets from the construction and commercial/industrial sectors) as a bulking agent and the mixture is composted over a period of approximately five days. Five different samples were collected and these are labelled and described below:

- **MSW8 - Untreated brown bin waste** - This material was standing in a pile at the facility and consisted of waste in a similar state to the uncomposted waste that was sampled previously. This time, however, it was planned that a greater amount of material would be collected and that this collection would be more representative than before. This target was achieved by firstly using a mechanical digger to take sections from the pile and place these on some plastic sheeting. The material on this sheeting was then mechanically broken apart and mixed by the digger over a period of five minutes. Following this, the material was spread by hand evenly across the sheeting and the sheeting was divided up into nine equally-sized sections. Equal amounts of material were hand sampled indiscriminately from each of these sections and placed in a large plastic container until the container was filled. It was noted that there was much more garden-waste (e.g. wood, grass etc.) in the brown bin waste than was observed in the previous four MSW samples (which seemed to be primarily composed of waste food). This is likely to be a reflection of the different season in which these samples were collected.
- **Wood - Waste wood** - A large container was filled with the chipped wood that is used as a bulking agent for MSW10. This wood is predominately from waste wood-pallets.
- **MSW10 - Pre-VCU treatment mixture of brown bin waste and wood waste** – Prior to processing in the VCU, brown bin waste (such as that which made-up MSW8) is mixed with waste wood in a homogenising device. The mixed material then travels up a conveyor belt towards the top of the VCU. Hand samples of this mixture were randomly taken from the conveyor belt and added to a large plastic container until it was filled.
- **MSW7 - Post-VCU sample** - Material was taken as it left the VCU (i.e. after the composting process had completed).
- **MSW9 - Screened composted waste** – This is a sample that consists of material that, similar to MSW7, was taken after 5 days treatment in the VCU. However, following this treatment the sample was then sieved down to <15 mm (which may have resulted in the removal of some of the wood bulking agent) and then fully composted in an aerated room where it was turned periodically over a 58-day period.

Sample Preparation of These Wastes

Two different processing methodologies were employed on the organic waste samples. The first method was carried out on the initial set of four samples received in February. Each sample was placed on a baking tray and put in an oven set at 60°C in order to dry the sample. This took several days as each day the waste needed to be broken apart (by using a knife) so that the drying process

could be thorough. The material was then put through a Retsch SM2000 chipper with a 20 mm mesh size for initial particle size reduction. The output was then sieved with the DS and DF fractions retained. Further particle size reduction of the “Large” fraction resulted from putting it through the FOSS Cyclotech 1093 mill (1 mm mesh size).

However, it was found that the use of the Cyclotech mill resulted in a minor amount of the original sample ending up as the DS fraction with a predominance of DF material resulting, see Table H-18. It should also be noted that there was probably also a significant amount of extra fines material produced that was not measured in the collected DF fraction but instead leaked, as airborne dust, from the mill itself.

For the second set of organic waste samples (MSW7-10 and the Waste Wood sample) a revised methodology was employed in order to increase the amount of material that resulted in the DS fraction and also minimise the carbohydrate degradation and loss of volatile components that may have occurred with the original drying procedure.

Instead of oven drying, the samples were taken to a farm in Labasheeda, Co. Clare where they were spread out on plastic sheeting in a barn on the assumption that they would dry over the course of several days. However, after a week the samples were still wet so, given that the weather was hot and sunny and was expected to be for several days, the samples were taken outside and put on plastic sheeting that was spread out on a field at the farm. The sheeting was held down with weights and measures were taken to ensure that material would not blow away or transfer between samples. The samples were left to air-dry during the hours of sunlight and the sheets were covered overnight to prevent rewetting by rainfall or the absorption of morning dew. Periodically the samples were broken apart by hand in order to facilitate a more thorough drying of the samples. It took three days before the samples had reached an acceptable moisture content at which point they were returned to the laboratory. As before, the samples were comminuted by processing through the Retsch SM-2000 chipper with the 20 mm mesh attached. The output was sieved with the DS and DF fractions retained. The “Large” fraction was reprocessed in the Retsch SM-2000 using the 1 mm mesh and the resulting output was sieved. All of the resulting “Large” fractions were reprocessed in the chipper and sieved until only DS and DF fractions resulted. Table H-18 shows that this procedure resulted in an increase of the percentage of DS material, compared with samples MSW1-4.

The results of the lignocellulosic analysis of the DS fractions of these brown-bin waste samples are provided in Table H-17. There are AIR but no KL or AIA data for samples MSW1-4 since the correct filter crucibles for the subsequent ashing of the AIR were not present in the lab at the time of

analysis of these samples. There are also no ASL data since the UV-Visible spectrophotometer had not been purchased at this point. It can be seen that the total sugars contents of the brown-bin waste samples varies greatly (from 12.4% for MSW9 to 51.7% for MSW4). Given the imperfect sampling and processing methodologies employed for samples MSW1-4, it would perhaps be unwise to interpret the data as being representative of the wider organic waste stream. However, samples MSW7-10 came from a much improved analytical procedure and provide interesting insights into the value of these materials for processing in hydrolysis technologies. The brown bin waste (MSW8) has approximately 30% total sugars, probably an insufficient content to allow this feedstock to be used in hydrolysis biorefining technologies. The wood that is added to it as a bulking agent, for composting, has a total sugars content of 53.9% and this results in the pre-VCU sample (MSW10) having a total sugars content greater than that of the brown bin waste (MSW8). The same is true for the KL content, whilst MSW10 also has a lower ash content than MSW8 due to the relatively low ash content (6.2%) of the wood sample.

Comparing samples MSW10 (pre-VCU) and MSW7 (post-VCU), the total sugars contents are very similar but the KL content has increased significantly (from 21.0 for MSW10 to 30.3% for MSW7) and the ash content has increased by 2.5%. Also, the total mass balance provided by the constituents analysed-for rises substantially, from 84.1% for MSW10 to 99.4% for MSW7. It appears that many of the non-ethanol-soluble extractives components that may have been present in MSW10 have been lost in the composting process. MSW9 is a sieved, fine particle size, post-VCU fraction and shows a similar relationship to MSW7 (a lower total sugars content and a substantially greater ash content) as COMX5 does to COMX4 (see Section 17.3.2).

Table H-18 presents the C, H, N, S, and ash data for the DS and DF fractions for some of the samples and the combined data where both the DS and DF fractions were analysed as well as heating value data for the combined samples (MSW7-10). In the case of MSW1-4 the DS samples were not analysed since these only contributed a small proportion of the total mass balance. The data provided in this Table are somewhat academic; the ash contents of these samples are too great to allow for their utilisation in thermochemical biorefining processes.

17.3.4 Autoclaved Municipal Wastes

Two samples of the organic fraction of black-bin-bag (i.e. heterogeneous non-sorted) domestic wastes were also analysed. These samples came from two different companies each of which had

developed similar processes. These involved the black-bin-bags and their contents being chopped up and then put through a rotary autoclave which helped to sterilise and homogenise the material. The output of this autoclave then went through a series of steps in order to remove the metallic and plastic components leaving a material that was thought to contain a significant quantity of cellulosic material. The first sample, MSW6, was sent by a company to the Carbolea laboratories and was a quite fine white power whilst the second, MSW11, was collected by the Author at the facility. This was a brown/grey fluffy material that had numerous small fragments of glass inside. These were removed prior to the wet-chemical analysis of the sample.

The ash, extractives, and lignocellulosic data for the DS fractions of these samples are provided in Table H-17. It can be seen that these samples have superior total sugars contents to the brown bin waste samples MSW7-10. The total sugars content of MSW6 is particularly attractive given that it is over 50%. Glucose was the most abundant constituent, it comprised 78.2% of the total sugars content of this sample, with the next largest constituents being KL and ash.

The probable reason for the higher sugars contents of these samples is the presence of paper in the black-bins; as Section 17.3.1 shows, paper/cardboard samples have large concentrations of polysaccharide sugars, glucose in particular. The autoclaving/sorting process has removed the non-organic components from this highly heterogeneous waste stream and produced a reasonably attractive feedstock for biorefining. In the case of sample MSW11, however, automatic means of removing the glass from the resulting sample will need to be sought and the ash content may be problematic. In some technologies, there could be approximately 1 tonne of waste ash for every 5 tonnes of this autoclaved material. However, if acid treatment technologies (e.g. DIBANET (F), and technology C) were to be used then some of this ash would be soluble; in which case the AIA figure is more important. There is a significant absolute difference (15.14%) between the ash and AIA contents of MSW11. If it is assumed for the DIBANET process that the AIR figure represents the solid post-hydrolysis residue from KL and AIA, and that 50% of the total sugars content also ends up in this residue (see Section 16.5), then for every tonne of feedstock there would be 520 kg of a residue that would have an 8.4% ash content. This appears to be a reasonable material for combustion/gasification given that the ash content is not too high. Hence, the ultimate solid waste from this process may only be approximately 4.4% (the AIA content).

17.4 Potential Yields from Biorefineries

Selected representative samples were taken from each waste feedstock category and their lignocellulosic and elemental data were put into a spreadsheet designed to determine the potential biofuel yields from processing these in up to eight different biorefining technologies (A to H). Some combined feedstocks were also analysed; these are summarised below:

- Household paper/cardboard (see Section 17.3.1).
- Commercial paper/cardboard (see Section 17.3.1).
- Exported paper/cardboard (see Section 17.3.1).
- For the straws, the average compositional values of each species (e.g. winter straw, winter wheat etc., see Table H-3) were used to represent the lignocellulosic data for that species and the LHVs provided in Table H-5 were used.

Not all technologies were considered suitable for these feedstocks. The following conditions were used:

- Moisture contents over 60% - Technologies C and H were considered unsuitable.
- Moisture contents over 80% - Technology G was considered unsuitable.
- Ash contents over 20% - Technologies G and H were considered unsuitable.
- Total sugars contents less than 40% - Technologies A-F were considered unsuitable.

These conditions resulted in:

- The only manure sample that meets the conditions for inclusion is MNPG1 which is considered a potential feedstock for processes A,B,D,E, and F.
- The only suitable garden waste samples, for which elemental and lignocellulosic data are available, are COMX1 (for process G only) and GRMX2 (for processes A,B,D,E, and F).
- SMC samples were considered not suitable for any of technologies A-H.
- PCBC1, PCGP1, and PCWC1 considered unsuitable for the thermochemical processes.
- MSW6 and MSW11 are considered suitable for technologies A,B,C,D,E,F. All of the brown bin waste samples are considered unsuitable.
- The waste wood sample is considered suitable for all processes, an LHV of 18.25 GJ/tonne has been used for this feedstock (FEC Consultants, 1990).

- The “organic” fraction of MSW/BMW that is often used in reports detailing the quantities of wastes in Ireland (Gaillot et al., 2005) mostly comprises food wastes and garden wastes (paper and wood wastes are categorised separately). A study (Williams, 2007) of the composition of this fraction, using Canadian wastes, found that it comprised 73.8% food waste, 16.3% leaves/grasses, and 9.9% prunings. Given the compositional data for brown-bin (i.e. mostly food) wastes (Section 17.3.3) and garden wastes (Section 17.3.2) it appears that such a feedstock would be unsuitable for all technologies, because it has an ash content that is too high for the thermochemical processes and a total sugars content that is too low for the hydrolysis processes.

Table H-19 provides the yields possible per tonne of feedstock for the suitable samples. These yields are expressed in litres (of ethanol, FT-diesel, or FT-naptha), in kilogrammes of levulinic acid for process F, and in energy terms (gigajoules (GJ) per dry tonne) for all technologies so that the end products can be compared on equal terms. The cells with the energy yields are formatted so that the technology that provides the greatest yield for a given feedstock has the darkest background whilst the technology that provides the lowest yields has the lightest background.

Table H-19 shows that Technology E has the highest energy yields of all the hydrolysis technologies whilst G provides superior yields compared with H. Technologies E and G achieve very similar yields for the straw feedstocks; however, the paper/cardboard samples, due to their high carbohydrate contents (see Section 17.3.1), offer more attractive yields from E than G. In contrast, the waste wood sample will provide higher yields from process G compared with process E. This is because the sample contains more lignin and less carbohydrate than the paper samples; the lignin providing a superior heating value compared with the polysaccharides as discussed in Section 2.8.

Of all the feedstocks, the paper samples clearly offer the greatest potential yields, through their processing in hydrolysis technologies. Indeed, the potential yield from sample PCTP1 (Tetrapak) in process E is over double the yield that could be achieved from MSW11 (autoclaved black-bin waste) in this process. Considerations other than total yield should be made when comparing feedstocks, however. For instance, the potential volumes of supply and the costs associated with these are highly important. If a material that offers lower yields is available at a significantly lower cost, or can be processed for a gate-fee, then it may still be commercially viable. Commercial considerations also eliminate processes B, E, F, and G from implementation in the near term. In their absence technology C offers the highest yields for the feedstocks which are considered suitable for it. This is

a concentrated acid hydrolysis process similar to that of BlueFire Ethanol, a company that is constructing a commercial biorefinery in California USA. Concentrated acid hydrolysis schemes have traditionally had high hydrolysis efficiencies and their commercial development has been hindered due to the high costs (both capital and operational) associated with the utilisation of strong acids (Graf and Koehler, 2000). BlueFire Ethanol claim that their technology has improved the acid recovery rate and can produce ethanol from the non-food fractions of MSW at a cost of less than \$1 per US gallon; however, it is not clear whether this low price is due to a significant contribution from gate-fees. If this is the case then its utility may be limited when the feedstock will need to be paid for (e.g. some papers, straws).

In the absence of technology C, processes A and D are left for potential near-term deployment. Process D is based on a much less refined version of enzymatic hydrolysis than that used by technology E, and is similar to the type of processes planned for commercialisation by companies such as Iogen and Verenium Biofuels. Given their heterogeneity, MSW6 and MSW11 may be problematic feedstocks for enzymatic hydrolysis since the enzymes typically perform best on a feedstock that is known and to which they can be tailored. The straw and waste paper samples should be much more suitable in this regard, and these could offer real near-term potential for such a process. Technology A uses dilute acids to hydrolyse biomass and may be more suitable for the MSW samples than process D; however, the fermentation of the complex hydrolysate may be problematic. Technology F, DIBANET, produces its end products through a mechanism that is entirely chemical and so is likely to be less sensitive to the heterogeneity of the biomass it processes. This process is still being developed at the University of Limerick, however.

Table H-19 presents the yields of potential biofuels possible from these feedstocks and technologies. However, considerations can also be made regarding the value of the byproducts from these biorefining processes. These include: kerosene in H; the formic acid by-product from F; and propanol, butanol and methanol in G. Butanol is considered to be an attractive transport fuel for current vehicles (Qureshi and Blaschek, 1999), however it is likely that its concentration in the final product stream of G will be too low for its economical separation. If value is given to the higher alcohols (propanol and butanol) produced in G, the energy yield for this process, according to the LHV, will rise from 48.8% to 58.4%. Also, no value is given to the surplus electricity that may be produced in the thermochemical schemes or to the value of the lignous residue that remains from hydrolysis technologies. It is anticipated that this residual material will be used to provide process electricity and heat for the hydrolysis processes in order that they may be energetically self-sufficient. This

contrasts with first-generation schemes which traditionally require the combustion of fossil fuels. However, it is possible that there will be a surplus amount of residue remaining after these demands have been met (Hayes et al., 2005). Surplus hydrolysis residues could offer the potential for increased biofuel yields if used in thermochemical technologies (G-H).

17.5 National Projections

Given the range of feedstocks considered suitable, by the Author, for processing in biorefining technologies (presented in Table H-19), it is clear that there are two main types of biomass that offer attractive yields: waste paper/cardboard and straws. A scenario is presented whereby all of the paper that is exported from Ireland is instead biorefined (the quantities provided by the EPA (2006) are used, see Section 17.3.1.1) . This resource offers the yields per tonne listed for the “Exported Paper” feedstock in Table H-19. In addition, the upper estimates for the practical and sustainable quantities of barley, wheat, and oat straws that can be used, see Table H-6, are also biorefined. The calculated total biofuel yields, in million litres (million kg for process F) and terajoules (TJ), from processing these feedstocks in Technologies A to H are presented in Table H-20. For each process the yields are summed across all feedstocks to provide a total biofuel energy yield which is then expressed as a percentage of the total energy demand, estimated for the year 2010 (in a previous paper by the Author (Hayes and Hayes, 2009)), for petrol and diesel transport fuels in the Irish Republic.

Table H-20 shows that exported paper/cardboard contributes the greatest quantity of biomass (54.9% of the total), followed by spring barley straw (22.9%), whilst the other feedstocks contribute much less to the total biomass resource. These total quantities of biomass allow for between 1.76% and 3.34% of the estimated demand for petrol and diesel transport fuels in Ireland to be met. This is a significant amount and is possible from using sustainable quantities of residues and wastes.

The removal of straws from the field, rather than allowing these to contribute to the soil organic matter, can be a contentious issue. However, the scenario put forward still allows for the retention of a portion of these straws on the land (RPS MCOS, 2004). The alternative is that only the waste paper/cardboard resource is biorefined. This scenario results in between 1.03% and 1.97% of the estimated demand for petrol and diesel transport fuels in Ireland to be met. Taking the yield from technology E as an example, this level of biofuel supply is 58.5% of the level of biofuel supply in the straw and paper scenario. Therefore, the drop in output is less than the loss in total biomass. This is

a result of the exported paper resource offering superior yields per tonne to the straws in the biorefining technologies (see Table H-19). However, if instead technology G was used to process these national resources of biomass, then the biofuel yield in an exported-paper only scenario would be 51.4% of the biofuel yield in a straw and exported paper scenario. This is because paper provides lower yields per tonne, compared with straws, in this process.

17.6 Other Energy Crops

In addition to *Miscanthus* (Sections 14 to 16), two other energy crops were analysed. These were sampled by the Author from the Teagasc Oak Park Research centre in Carlow. These were subsequently processed to an appropriate particle size and then analysed with reference methods. The feedstocks are briefly summarised below and their lignocellulosic data are presented in Table H-21. The whole plant compositions of two (over 2 m tall) *Miscanthus* plants, one collected in September and the other in February, are provided for comparison.

Switchgrass (*Panicum virgatum*): This is an American perennial C₄ grass that is established from seeds and has received much attention in the US as a potential lignocellulosic energy crop (Wolf and Fiske, 1995). Christian and Riche (2000) carried out trials, in the UK, of 8 varieties over six years and found that the average yield of the four top-yielding varieties in the final year was 12 t ha⁻¹ y⁻¹ (on an oven dry basis). Its yields are said to be sustainable over reasonably long periods (approximately 15 years) and, while these yields are generally lower than *Miscanthus*, the ability to be established via seed will result in significantly lower establishment costs (FAIR, 2000).

However, the productivity of the switchgrass *Shawnee* variety that was sampled from Oak Park was very poor, as shown in Figure H-4 (a). The low yields of the crop are reflected in its analytical data; the total sugars and KL contents are very low. In contrast, the ASL, ash, and extractives contents are high. The total mass balance for this sample was only 81.0%. It is likely that much of the remainder of the mass balance consisted of extractives components not soluble in 95% ethanol.

Short rotation willow coppice: Short rotation coppices (SRC) can comprise a variety of species including poplars and robinia but willow (*Salix*) is the SRC-type grown commercially in Ireland. The plantations are established from stem cuttings and the stems that grow in the first year are coppiced (cut back) to encourage multiple sprout formation. Subsequently, the plantation is harvested every 2-5 years. The harvest window can span the whole year but harvesting traditionally occurs in the

winter. The harvested material can be used in power generation schemes, and for the production of wood chips for domestic/commercial boilers. Section 10.7 details the total land in Ireland which was planted with SRC between 2007 and 2010. Where the land is most suitable for *Miscanthus*, it has been favoured over coppices in recent years (Styles and Jones, 2007) due to higher yields and less disease risks (Rice, 2001, Mitchell, 1995, Royle et al., 1992). The sample analysed by the Author was collected in February 2010. It consisted of the chips of material that had been blown from the forage harvester used to harvest the crop on the same day. Figure H-4 (b) provides a photograph of a willow SRC plantation at this site.

The lignocellulosic data for this sample show that the total sugar content is significantly greater than that of the switchgrass sample, however it is less than both of the *Miscanthus* plants, particularly the February plant. Compared with the *Miscanthus* samples there are different proportions of each of the sugars; glucose, rhamnose, galactose, and mannose contribute more to the total sugars content, whilst arabinose and xylose contribute less. This is as would be expected for a hardwood species, as discussed in Section 17.3.1. Its KL content is higher and ash content lower than the *Miscanthus* plants, also as would be expected for a woody species.

17.7 Summary

An earlier paper written by the Author (Hayes and Hayes, 2009) details the waste feedstocks that the Author considered, at the time, to have potential for commercial utilisation in a range of biorefining processes in Ireland. Data were collected regarding the national quantities of these wastes that were produced in Ireland and these linked with secondary data concerning their lignocellulosic compositions and heating values in order that estimated yields could be determined for eight different biorefining technologies. This chapter has detailed the work conducted by the Author since that time in order to determine the yields, for many of those feedstocks and some additional materials, based on primary analytical data obtained in the Carbolea laboratories. These primary data have been obtained from Irish samples and using advanced analytical methodologies that allow for the determination of the constituent sugars of the polysaccharides (rather than using detergent methods). The following points list the most important differences, as a result of this analytical work, that have been noted compared with the data and conclusions presented in the earlier paper (Hayes and Hayes, 2009):

- Spent mushroom compost samples have much lower sugar contents than suggested by articles in the literature (Jordan et al., 2008). These articles used detergent fibre methods which can not provide certainty that the true polysaccharide contents are the same as the weight losses determined in these gravimetric methods. The complexity of SMC is likely to allow for many non lignocellulosic components ending up in the ADF, NDF etc. fractions. A much higher degree of confidence can be placed in the results of this Chapter.
- As a result of these updated data, SMC is not considered suitable for any of the eight biorefining processes.
- Similarly, the composition of the “organic” fraction of MSW/BMW (i.e the green waste and brown bin waste) is poor for ultimate utilisation in biorefining processes with ash contents that are too high and sugar contents that are too low.
- Autoclaved (and then sorted) black bin wastes appear to be more attractive feedstocks, with the presence of paper in the waste stream improving the carbohydrate content.
- Some pig manures have reasonably attractive sugar contents but the issue of high moisture contents remains a problem.
- Waste papers offer the highest potential yields in hydrolysis processes, in agreement with the previous observations (Hayes and Hayes, 2009), however the carbohydrate composition can vary greatly with the type of material.

It has been concluded that the most practically and commercially feasible feedstocks in the near term are waste straws and waste paper. In particular, the resource of waste paper that is already collected (for subsequent distribution overseas) represents a feedstock of great potential since the logistical issues regarding collection and distribution have already been dealt with. The use of this resource in national biorefineries in order to address European biofuels regulations is surely a strategically superior means of utilising this feedstock than the current status-quo.

18 The DIBANET Project

In June 2008 the Author submitted a proposal to the EU 7th Framework Programme for a €3.7 m (with € 1.42 m going to UL) research project. This proposal, codenamed “DIBANET” and entitled “The production of sustainable diesel-miscible biofuels from the residues and wastes of Europe and Latin America”, was submitted to the Energy.2008.3.2.1 call which was termed “Enhancing international agreement between the EU and Latin America in the field of biofuels”. The proposal was successful and the 42-month project commenced in July 2009. There are a total of 13 partners in the project, 6 from Europe and 7 from Latin America.

18.1 Concept

The central concept of DIBANET is the production of levulinic acid (and co-products) from lignocellulosic biomass and the conversion of this to ethyl-levulinate, a novel diesel miscible biofuel (DMB) produced by esterifying ethanol with levulinic acid (LVA).

There are five key scientific objectives to the project:

1. Optimise the yields of levulinic acid (and co-products), from the conversion of biomass, while minimising chemical/energy requirements.
2. Improve the energy balance of the production of levulinic acid and the total biofuel yields possible from a feedstock by sustainably utilising the residues in pyrolysis processes to produce a bio-oil that will be upgraded to a DMB.
3. Reduce the energy and chemical costs involved in producing ethyl-levulinate from levulinic acid and ethanol.
4. Select key biomass feedstocks for conversion to levulinic acid, analyse these, and develop rapid analytical methods (NIRS) that can be used in an online process.
5. Analyse the DMBs produced for their compliance to fuel requirements and, if non-compliant, suggest means to achieve compliance.

It was decided at the launch of the project that Miscanthus would be the primary European feedstock for the project and sugarcane bagasse and sugarcane “trash” (field residue after mechanical harvesting) would be the primary Latin American feedstocks.

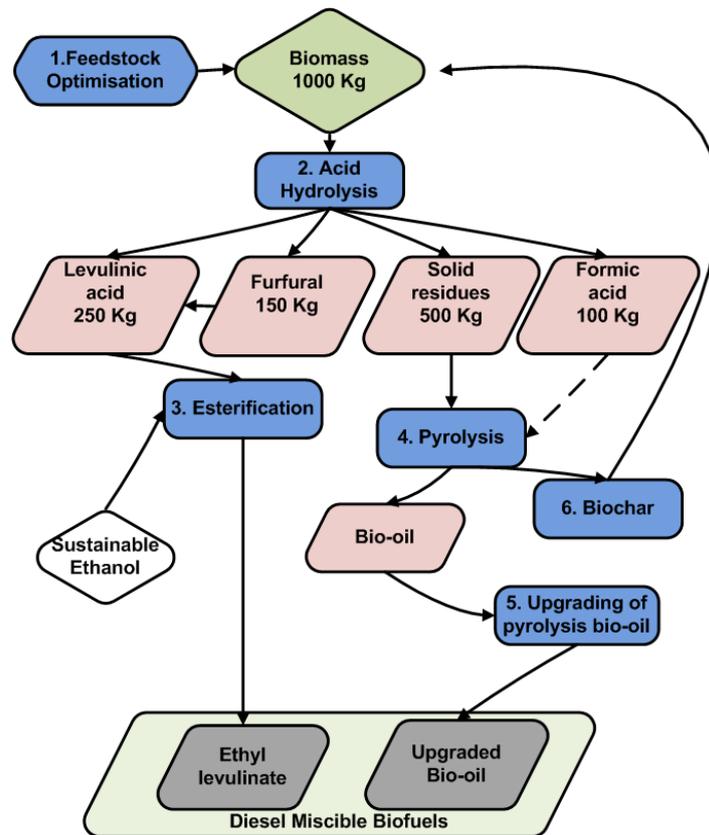


Figure 18-1: The DIBANET process flow. A representation of the processes (in blue); their primary products (in pink); the final downstream biofuels (in grey); and their linkages.

The DIBANET process flow is presented in Figure 18-1, it involves the following steps:

1. Optimisation of the sourcing, selection, analysis, and preparation of the feedstock.
2. The hydrolysis and subsequent degradation of biomass. This can produce (i) levulinic acid, (ii) furfural (which can be converted to levulinic acid via hydrogenation), (iii) formic acid, and, (iii) solid residues (SR).
3. The esterification of levulinic acid with (sustainable) ethanol to produce the DMB ethyl-levulinate.
4. Pyrolysis of some or all of the SR to produce a bio-oil and a biochar. Pyrolysis can be enhanced by using the formic acid produced in (2) as a co-feed.
5. Upgrading (catalytic) of the bio-oil to produce an upgraded bio-oil that is miscible with diesel.
6. Utilisation of the biochar as a soil-amender for plant-growth promotion or to fuel the processes

The experience of the Author in NIRS has led to his involvement in parts of Step 1 and Step 2, and these will be detailed in the rest of this chapter.

18.2 NIRS Work on Latin American Feedstocks

Step 1 in the DIBANET process flow, Figure 18-1, involves the evaluation of potential feedstocks for the DIBANET hydrolysis process and the wet-chemical analysis of those that are considered to have potential. Quantitative NIRS calibrations will be developed for selected feedstocks.

The previous chapters in this Thesis have outlined the data collected for Irish feedstocks that informed the work in DIBANET. It was decided that Miscanthus would be focussed on for NIR analysis and the results of this are provided in Chapters 14 to 16. The analysis of a wide variety of waste materials, Chapter 17, has also informed which of these could give attractive yields in a DIBANET-type process.

In addition to UL there are four other project partners involved in this feedstock-analysis Work Package. These partners, and their roles, are detailed below:

- CTC (Centro de Tecnologia Canavieira) – This is a research organisation, with a headquarters in Piracicaba, Sao Paulo, Brazil, that is focussed on sugarcane. In DIBANET it is involved in the collection and analysis, via NIR and wet-chemical reference methods, of sugarcane bagasse and sugarcane “trash” (field harvesting residues, mainly leaves). This partner also sends these residues to UL and other partners so that they may be processed in the DIBANET technologies (hydrolysis and pyrolysis). CTC also has an online NIR system at a sugar mill in Sao Paulo state. This system will be used in the latter part of the project.
- UNICAMP (University of Campinas) – This is a University located in Campinas, Sao Paulo, Brazil. Its involvement is in the wet-chemical and NIRS analysis of Latin American feedstocks other than sugarcane residues. At the time of writing the Author understands that residues from the banana and coffee industries will be the materials that will be focused on for the development of quantitative NIR models.
- FOSS – This company, based in Denmark, is a manufacturer and supplier of NIRS equipment. Its role will be to aid in the transfer of lab-based NIRS calibrations to the online system.

- FUN – Fundacion Chile – This non-governmental-organisation, located in Santiago, Chile, is involved in the evaluation of the socioeconomic factors regarding the sourcing of lignocellulosic feedstocks and in the integration and systems analysis of the DIBANET process flow, Figure 18-1.

At the time of writing, June 2011, neither CTC nor UNICAMP have any full wet-chemical datasets for the analysis of any samples of Latin American feedstocks. Following completion of his PhD viva the Author will need to travel to Brazil to resolve this problem. These partners have, however, collected samples and obtained WU (wet unground), DU (dry unground), DG (dry ground), and DS (dry sieved) spectra of some of these. The spectra collected by CTC will be examined in this Chapter. The number of samples analysed by WU, DU, DG, and DS methods are outlined in Table 18-1.

Table 18-1: The number of samples of sugarcane bagasse and trash scanned by CTC in each dataset.

Scan Dataset	Number of Samples Scanned	
	Sugarcane Bagasse	Sugarcane "Trash"
WU	201	37
DU	159	28
DG	94	36
DS	60	36

18.2.1 Application of Miscanthus and Australian Sugarcane Bagasse NIRS Models to CTC Spectra

In December 2010 the Author visited the laboratories of CTC. This partner has an identical FOSS XDS device (see Section 5.3.2.3) to that used in UL. The XDS is specifically designed to facilitate the cross-device transfer of calibrations. At that point WU spectra existed for 157 bagasse samples. Attempts to apply early Miscanthus WU quantitative NIRS calibrations to these samples appeared to be unsuccessful. The values obtained differed substantially from the DS-model predictions of the corresponding DG-samples (there were no DS scans at this point).

In December 2010 CTC did have data for the moisture content of 92 of the 157 WU samples. The Author undertook a PCA of the WU spectra (transformed by the first derivative) and a PCR for moisture content. It was found that the majority of the y-variance (moisture content) was explained by the first PC, a fact illustrated by the loading vector of PC1 which bore many similarities to that of water. A similar PCA was conducted on the WU Miscanthus spectra obtained at UL followed by a PCR for the glucose content. It was found that the first PC explained no significant amount of the Y-

variance; instead its loading vector looked very similar to that of water (as with the WU bagasse PCA). The largest individual increments in explained-glucose-variance for the Miscanthus WU-PCR were seen for PC's 2-4. Given that glucose is the most abundant constituent in Miscanthus this is logical. It was therefore considered that, even though the loading vectors for PCs 2-4 of the bagasse-PCA looked different to those of the Miscanthus model, there could be a good possibility that these could explain significant amounts of the variation in glucose content among these samples.

The Unscrambler X software has a feature that allows for the selection of equally spaced samples along a PC axis. This feature was used to select, using PCs 2 to 4, 15 samples each along PC. This resulted in a total of less than 60 samples since some were represented in the selected samples of more than one PC. The Author suggested to the researcher at CTC that these samples should be processed and analysed by reference methods first.

All of the work conducted in December 2010 was based on early, unrefined, Miscanthus NIR models. The final models presented in Chapters 15 and 12 allow for the prediction of the most important constituents of the CTC samples, using the Miscanthus and Australian sugarcane bagasse models, and using the expanded CTC NIR datasets provided to the Author in June 2011 (detailed in Table 18-1).

18.2.1.1 PCA Projections

Firstly, tests were undertaken to see how well the spectral variability of the WU and DS bagasse and trash datasets (the "projected" datasets) could be modelled by PCA models based on the WU or DS spectra of Miscanthus or BSES bagasse samples (the "modelled" datasets). Hence, for example, the DS BSES bagasse dataset was transformed by SG-1,1,10,10 and a PCA conducted over the spectral range 1100-2500 (these PCA model conditions were consistent for all the models developed). This produced a model onto which the (similarly-transformed) CTC bagasse DS dataset was projected. This projection involves the application of the model loading vectors to the spectra and allows PC scores to be computed for each sample along with statistics for their explained X-variance and leverage (using models of varying PC numbers). The important point is that these projected samples do not influence the model, meaning that the results of their projection will demonstrate how the loading vectors of the model can represent the spectral variability of these "unknown" samples.

Figure I-1 presents plots for this projection (CTC DS bagasse samples projected onto a BSES DS model). Figure I-1 (a) shows the explained X-variance with an increasing number of PCs in the model.

The blue line represents the model samples in calibration, the red line the model samples in cross validation, and the green line represents the projected samples. It can be seen that, as would be expected, the projected samples are less well explained by the model than the BSES samples. Figure I-1 (b), (c), and (d) present scores plots for PC1 vs PC2, PC3 vs. PC4, and PC5 vs. PC6, respectively. In these plots the model samples are coloured blue and the projected samples are coloured green. A model that represents the spectral variability of the projected samples well should not see these samples outside the PC score space of the model samples. However, many of the projected samples in Figure I-1 lie outside the score-regions of the model samples, particularly in the plots of later PCs (e.g. PC5 vs. PC6, Figure I-1 (d)).

Figure I-1 (e) presents an influence plot for a 7 PC model. It can be seen that, compared to the projected samples, the model samples occupy a very small region of this plot, with leverage and residual X-variance values far below those of the projected samples. Since the projected samples are not involved in the determination of the model loading vectors it is possible for these samples to have leverage values above 1 (see Equation (6.56) in Section 6.4.3.5), and this is the case for all of the projected samples in Figure I-1 (e). It can therefore be said that this model is not a good representation of the spectral variability of the projected samples and quantitative predictions of the unknown samples, based on models (e.g. PLSR) developed on the spectra that formed the basis of this PCA, cannot be trusted.

Figure I-2 present plots for the projection of the CTC trash DS samples onto the BSES DS model. The fit of the model here is even poorer than in Figure I-1, with the projected samples having higher values for their leverage and residual X-variance.

Figure I-3 presents plots for the projection of the CTC bagasse DS samples onto the Miscanthus DS model. The fit is much better here, with the projected samples typically lying within the score-space of the model samples and having leverage and residual X-variance values that do not exceed those of the model samples.

Figure I-4 presents plots for the projection of the CTC trash DS samples onto the Miscanthus DS model. These projected samples occupy a quite limited region in the scores plots and most of them do not have leverage or residual X-variance values in excess of the Miscanthus samples. However, there is a group of 9 trash samples that have leverage values greater than those of any Miscanthus samples.

Figure I-5 presents plots for the projection of the CTC bagasse WU samples onto the BSES WU model. The explained X-variance of the projected samples, Figure I-5 (a), is much higher than in the

previous projections; however, this is primarily a result of the spectral effects of moisture being represented by the model loading vectors (mainly PC1). This PC will be of minimal or no relevance in predicting the chemical compositions of these unknown samples. Thus the scores and influence plots need to be studied before assessing the suitability of the model samples. These plots show that the projected samples have leverage and residual X-variance values far in excess of the model samples and so are not well represented by the model.

Figure I-6 presents plots for the projection of the CTC trash WU samples onto the BSES WU model. As with the DS projections, this model is an even poorer fit to these projected samples than it is to the CTC bagasse WU samples.

Figure I-7 presents plots for the projection of the CTC bagasse WU samples onto the Miscanthus WU model. The fit here is better than for the BSES WU model. However, some of the projected samples do take up, for some PCs (e.g. 5 and 6), regions of PC score space that are unoccupied by the model samples. Some projected samples also have leverage and residual X-variance values in excess of the model samples.

Figure I-8 presents plots for the projection of the CTC trash WU samples onto the Miscanthus WU model. These samples occupy much smaller regions of PC scores and influence-plot spaces than the projected bagasse samples and have reasonable leverage values, however the residual X-variance values for these samples can be high.

These Figures show that, despite being based on the same species, the BSES models do not represent the CTC samples well. The Miscanthus models perform significantly better, with the Miscanthus DS model offering advantages over the WU model.

18.2.1.2 Predictions of Compositions of the CTC Samples

The best models for glucose content, Klason lignin content, and xylose content for the WU, DU, DG, and DS datasets of the Miscanthus samples (see Section 15.2.2) and for the WU, DU, and DS datasets of the BSES bagasse samples (see Section 12.3.3) were used to predict the composition of the CTC bagasse and trash samples based on their WU, DU, DG, or DS spectra. No BSES DU xylose model was used since the best model developed for this constituent using the BSES bagasse samples gave very poor regression statistics (see Table C-34). Each of the models, in prediction, used the number of PLS factors determined via Haaland's criterion and listed in the appropriate Tables for these models in

previous Chapters. As well as the predicted compositional values, the deviation in prediction (see Section 6.6) was calculated for the given PLS factor. Figure I-9 provides quantile plots for the predictions of the glucose content of the CTC bagasse, Figure I-9 (a), and CTC trash samples, Figure I-9 (c), as well as quantile plots for the deviation in glucose content prediction for the bagasse, Figure I-9 (b), and trash, Figure I-9 (d), samples. It can be seen that the structure of the glucose content predictions varies according to the model used and that the deviations in prediction are much more varied for the BSES models (labelled AUS in Figure I-9).

Interestingly, in Figure I-9 the average value and range in values for the deviation in prediction are greater for the DU models than for the WU models. It is possible that the researcher in CTC is using a different sample presentation method for the collection of the DU spectra to that used in UL and that this may be the reason for this. The Author did notice on his visit to the CTC laboratories in December 2010 that the DU bagasse contained a significant amount of very fine material that could, if the sample presentation method was not adjusted, lead to a bias towards this particle size fraction being presented to the NIR cell window at unproportionate quantities (see Section 11.1).

For both the trash and bagasse groups the range in predicted glucose values was greater when using the BSES models for prediction. A large concentration range would be a help in developing quantitative NIRS models for these samples. However, the poor fit of the CTC samples to models based on the BSES spectra suggests that the predictions based on BSES models cannot be trusted and that the Miscanthus models should be preferred. Table I-1 presents histograms and associated statistics for the predicted glucose, KL, and xylose contents of the 60 CTC bagasse and 37 CTC trash DS samples using the Miscanthus DS models for these constituents. It can be seen that the ranges in these concentrations are low; in the case of glucose it is only 4.18%. Such a small range will make the development of accurate and precise NIRS models difficult – an RMSEP of 0.28% would be the maximum permissible to allow an RER of 15 and provide a model that is suitable for the quantitative prediction of unknown samples.

Until reference analytical data are available for these samples it will be impossible to say if these predicted values and ranges are accurate, or even if the application of Miscanthus models has potential in discriminating between samples of high/low concentration values (and so allowing for targeted CTC sample-selection for future reference analysis). If the predicted glucose values using different Miscanthus models (WU, DU, DG, DS) are compared between these models, Figure I-10, then the R^2 values are disappointing. In these plots a sample with a relatively high glucose content using one model may not have a relatively high glucose content when using another model. There are also biases apparent in these comparisons, for example the DS model tends to predict higher glucose

contents than the WU model, Figure I-10 (a), whilst the WU model tends to predict higher glucose contents than the DG model, Figure I-10 (d). However, given that the deviations in prediction tended to be least for the Miscanthus DS model, it is probably most prudent to use this model as a basis for sample selection rather than the WU, DU, or DG models.

Figure I-11 presents the quantile plots for the KL predictions and Figure I-12 presents these for the xylose predictions. Similar trends to the glucose quantiles plots are seen for these constituents: there are greater deviations in prediction for the BSES models, and there are lower ranges in predicted concentration values for the Miscanthus models. The BSES models tend to perform most poorly on the trash samples, particularly in the case of the DU and DS models where the mean KL prediction is around 4% less than for the other models.

Table I-1 shows that the ranges in the predicted KL and xylose compositions for the DS bagasse/trash samples are even lower than the range for the predicted glucose content. If these ranges are accurate then it is questionable as to whether the development of quantitative NIRS calibrations is warranted given that the variability in composition is so low. The range in compositional values will be dependent on the methodology employed in obtaining samples, however. Using too narrow a field of sample selection will limit the variations between the samples collected. The Author suggested to CTC that samples be taken from as many mills as possible; however, it is not clear if this suggestion was followed. There are clearly many issues that the Author will need to address on his visit to the CTC laboratories.

18.3 NIRS Spectra and Calibrations for Pretreated Miscanthus Samples

Studies (Aden et al., 2004) by researchers from the US National Renewable Energy Laboratory have identified valuable sugar-based building blocks from lignocellulosic biomass. Of more than 300 initially selected candidates 12 were selected based on the size of the potential markets for these and their derivatives and on the complexity of the synthetic pathways involved in their production. One of the 12 is levulinic acid (LvA), referred to as 4-oxopentanoic acid. It is a C5-molecule with both ketone and carboxylic acid functionalities, see Figure 18-2, which provide interesting synthetic pathways for use as a chemical building block for a wide variety of derivative compounds.

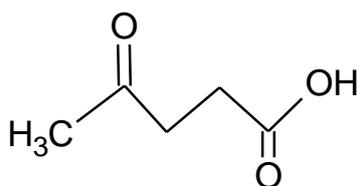


Figure 18-2: Levulinic acid

Controlled degradation of hexose sugars by acids is the most widely used approach to prepare LvA from lignocellulosic biomass (Hayes et al., 2005). The theoretical yield of LvA from C6-sugars is 64.5 wt % due to the co-production of formic acid; however, the recovered yields are usually significantly lower due to the formation of undesired insoluble-materials called humins (Hayes et al., 2005). Other possible by-product of biomass hydrolysis include furfural, formed by the decomposition reactions of C5-sugars.

The Biofine technology (Hayes et al., 2005) uses a two-stage process for the production of LvA. Carbohydrate feedstock and a dilute sulphuric acid solution are mixed, and the slurry is supplied continuously to a tubular reactor. This reactor is operated at a temperature of 210–220 °C and a residence time of 12 s in order to initially hydrolyse the carbohydrate polysaccharides into their soluble monomers (hexoses and pentoses). The product of the first reactor is fed to a continuously-stirred tank reactor operated at a lower temperature (190–200 °C) but with a longer residence time of 20 min. LvA is removed by drawing-off liquid from the second reactor and recovered by solvent extraction. Yields of LvA of up to 50% by cellulose weight have been quoted as obtainable from a pilot-plant processing one tonne of feedstock per day (Hayes et al., 2005); however, these yields have typically been obtained from processing paper sludge which has a low lignin content. It is considered that lignin could interfere with the polysaccharide to LvA/furfural conversion pathways, resulting in the formation of humin products instead of these platform chemicals.

The DIBANET proposal put forward that a continuous reactor system would be constructed and operated at UL and process selected samples of lignocellulosic biomass from Europe and Latin America. This technology was initially conceptually similar to the Biofine process, in that it would be a multi-reactor system that targeted the production of LvA and fufural in high yields. It was considered that the state of the art would be improved via the following strategies:

1. Design and engineer the conditions (reactor configurations, time, temperature, pressure, acid concentration and type, residence times) in the reactor(s) for maximal yields at reduced costs;

2. develop solid acid catalysts to convert carbohydrates to levulinic acid and evaluate their performance in this regard; and
3. use ionic liquids, in a pretreatment step, to swell the biomass matrix so that the polysaccharides are more amenable to hydrolysis.

Initially a batch reactor was operated at UL, using dilute sulphuric acid for the hydrolysis of biomass and the production of LvA and co-products. As of June 2011 a continuous system has been operational. Its configuration and planned use are now significantly different from that of the Biofine process. In particular, there has been a focus on the pretreatment of the biomass prior to hydrolysis. It is considered that increasing LvA yields beyond the current art is feasible if the feedstock is pre-treated in such a way that heterogeneities and complexities in the lignocellulosic matrix are decreased (Girisuta, 2007). Initially ionic liquids were evaluated for this purpose, but the high cost of these makes their use uneconomical in any commercial process. Instead, the use of a mixture of formic acid (a co-product with LvA) and hydrogen peroxide has been studied in, to date, batch pretreatment experiments. This technique has been observed, in experiments in the Carbolea laboratories, to solubilise the lignin with the majority of the polysaccharide sugars remaining in the solid residue. This residue is then a potential feedstock for hydrolysis to LvA, and it is considered that improved yields, over the virgin biomass, could result due to the reduced KL content.

Colleagues at Carbolea have been carrying out pre-treatment experiments whereby they vary the conditions (time, concentrations of formic acid and hydrogen peroxide etc.). To date all these experiments have been conducted on the same batch of, late harvest, Miscanthus stems. These researchers have analysed the solid residues that remain after the pre-treatment using the methods outlined in Section 11. They have found that the reaction conditions do have a significant influence on the ultimate lignocellulosic composition of the residual material.

Prior to the ethanol-extraction stage of the analysis of these samples their spectra were taken using the small NIRS cell. The combinations of these spectra with the reference-analytical data have allowed the Author to develop PLSR models for the important constituents. Since, to date, only 24 unique samples have been analysed (including one sample with no pretreatment), these models have been tested only via (full) cross validation.

Table I-2 presents regression statistics for the models that were developed. For every model the spectral dataset was transformed by SG-2,2,25,25 and the PLSR regression took place over the 1100-2500 nm spectral region.

Figure I-13 presents the predicted y vs. reference y plots for the samples, in calibration and cross-validation, for the glucose, xylose, KL, and AIR models, whilst Figure I-14 presents these for the ASL, galactose, arabinose, and ash models. Figure 18-3 presents the plots for the total sugars and ethanol-soluble-extractives models. In each of these Figures the sample that was not pretreated is labelled “Untreated”. Observing these Figures it can be seen that there is a significant variation in the concentration values. For example, the range in KL content is 20.5% and the range in glucose content is 39.6%. In many cases the range in values for these 24 samples is greater than for the numerous *Miscanthus* samples used to develop the models in Chapter 15. These extended ranges mean that, while the RMSECVs may often be larger than those for the corresponding DS/DT models in Chapter 15, the RER_{CV} and R_{CV}^2 values are greater.

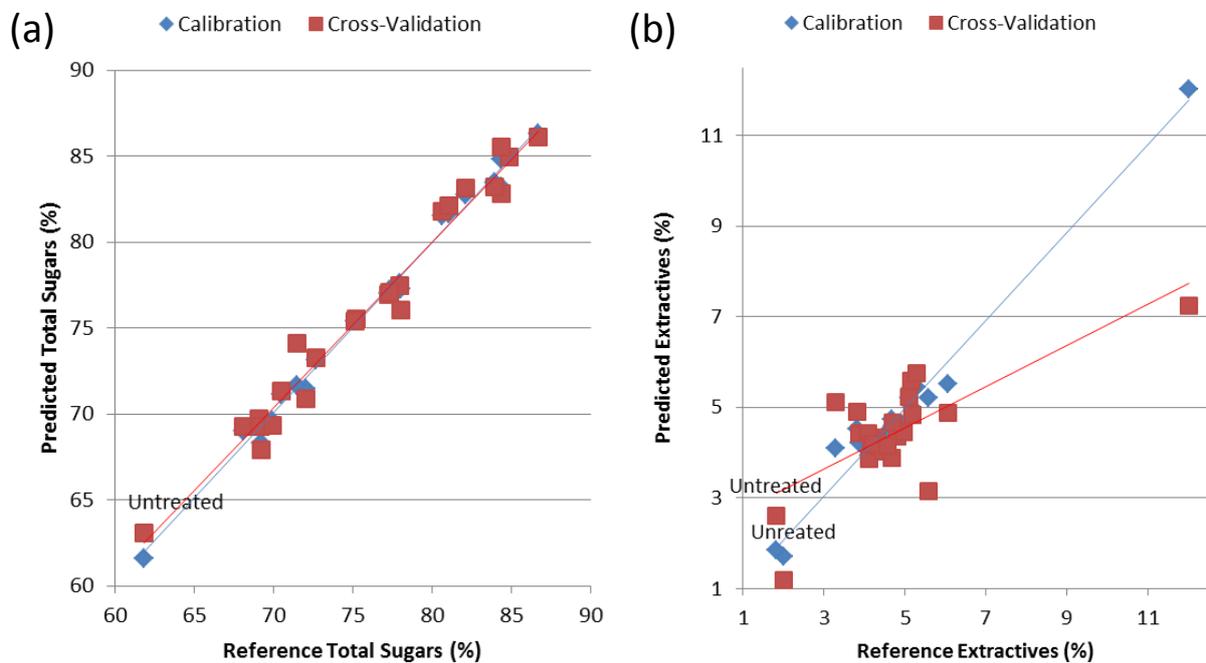


Figure 18-3: Predicted y vs. reference y for the 23 pretreated *Miscanthus* samples and the one sample that was not pretreated. (a) Total sugars content; (b) extractives content.

While the models for the most important constituents (glucose, xylose, KL) have good regression statistics, particularly given the limited number of samples used to develop the model, there are some poorly performing models. R_{CV}^2 values are less than 0.7 for the rhamnose, mannose, ethanol-insoluble-ash (EIA), and extractives models. With the exception of extractives, these models are for relatively minor constituents and of little relevance to biorefining yields. The extractives content model performed poorly partly because the majority of the samples had very similar extractives contents, except for one outlying sample with a significantly higher extractives content and another outlier (the untreated *Miscanthus* sample) with a lower extractives content than the modal group, as

illustrated in Figure 18-3 (b). It is also important to note that the ethanol soluble extractives content increased significantly after the pretreatment methods were applied. Indeed, this increase was greater than the proportional loss of dry mass in the whole sample. This implies that these extractives were being “produced” in the pretreatment step and subsequently incorporated into the solid residual material that was used for reference extractives analysis. It is unlikely that more of the types of conventional “extractives” that were present in the “untreated” sample (e.g. waxes, chlorophyll, etc.) were being made in this way. What is more likely is that the pretreatment process may degrade polysaccharides or lignin-components to such a degree that they do not become solubilised in the pretreatment process but that they are then amenable to extraction with ethanol. Chemical characterisation of the liquid extract would be a good start in trying to figure out what lignocellulosic components are giving rise to the increased extractives contents of the pretreated samples. Developing an NIRS model for this constituent will therefore be complex, given that a wide variety of constituents may be responsible, and will certainly require a larger calibration set within which the concentration values are more evenly spread across the range.

Indeed, results of all these models on pretreated samples, while promising, are only indicative at this stage, and many more samples should be supplied to increase the size of the calibration set and also to allow for an independent validation set. At the present time the use of these models to predict the compositions of unknown samples is not advised; when they were applied to predict the glucose contents of samples pretreated with a different method (using solid acid catalysts) the predicted glucose content was over 100%.

18.3.1 Comparison of Spectra

Figure I-15 presents a plot of the, SG-2,2,25,25 transformed, spectra of two samples, 112000, the untreated sample, and 112016, a sample that underwent relatively severe pretreatment conditions. From sample 112000 to sample 112016 the total content of hemicellulosic sugars fell from 22.2% to 12.9%, whilst “extractives” increased from 1.8% to 6.1%, and ash fell from 3.9% to 2.0%. The most important compositional changes, however, were a fall in the KL content (from 21.4% to 0.9%) and a rise in the glucose content (from 39.6% to 73.8%). It should be expected, therefore, that the absorbance values should be less in the regions associated with the vibrational frequencies of molecular bonds in lignin and greater in the regions associated with the vibrational frequencies of the molecular bonds of cellulose.

Figure I-15 shows that there are wavelength regions where the two spectra differ quite significantly. One of these is labelled “A” and is associated with an absorbance at 1672 nm that is a characteristic region for lignin, representing the first overtone of a C-H stretching vibration in aromatics (Krongtaew et al., 2010a). The untreated sample, having a higher KL content, has a stronger absorption band in this region (represented as a valley in 2nd derivative spectra) whilst the pretreated sample absorbs much less here.

An absorbance at 1724 nm has been attributed to the 1st overtone of a C-H stretch in the aliphatic lignin groups (Shenk et al., 2008), and this wavelength, labelled as “B” in Figure I-15, once again shows a stronger absorption for sample 112000 than sample 112016.

Label “C” in Figure I-15 highlights the stronger absorbance seen for the pretreated sample around 2100 nm, a region associated with combination bands involving the bonds of polysaccharides (Shenk et al., 2008) and a characteristic absorption region for cellulose (Üner et al., 2011, Martin and Aber, 1994). Also, label “D” highlights the stronger absorbance seen for the pretreated sample at the 2270 nm wavelength, a region of the spectrum associated with an O-H stretch/C-O stretch combination band in polysaccharides (Shenk et al., 2008)

18.4 Summary

This chapter has summarised early work that has taken place regarding the assessment of Brazilian sugarcane bagasse/trash samples for their suitability for the development of quantitative NIRS models, and also regarding the development of quantitative models for pretreated *Miscanthus* samples. Regarding the bagasse samples, it is unfortunate that at this stage (June 2011) there still have been no analytical data from CTC. That makes it very difficult to test the utility and accuracy of employing models developed on different species/varieties to predict these samples. Until such data are available the spectra can only be used to assess how well the unknown samples fit to a model of known samples. The PCA projections that took the form of this evaluation have suggested that the *Miscanthus* datasets provide a greater potential for explaining spectral variability in the CTC samples than the Australian sugarcane bagasse datasets do. The application of the *Miscanthus* quantitative models to the CTC spectra have resulted in predicted compositional values that are reasonably in line with what would be expected for sugarcane bagasse samples (see Section 12.1.3); however, the predicted values do vary according to the *Miscanthus* model used (DS, DG, DU, WU).

The most logical dataset for accurate prediction would be DS since particle size and moisture-content differences between the sample types will be minimised. Indeed, DS models do tend to give the lowest deviations in prediction for the Miscanthus models. The first task that the Author will undertake when he travels to the CTC laboratories will be to quickly put in place an accurate reference-analytical protocol and use this to find the compositions of a limited number of samples chosen along the predicted concentration ranges using the Miscanthus DS models. If these fit reasonably well, even if a bias correction is needed, then this strategy can be continued to strategically select samples for reference analysis and calibration development.

However, if the DS Miscanthus models are useful for prediction, then these predictions suggest that the compositional variability of sugarcane bagasse and trash samples is low. This may be because the CTC samples that have, to date, been processed to a DS state are poor representations of the true variability of Brazilian bagasse/trash samples. There are a total of 141 samples that have WU, but no DS, spectra. These will need to be evaluated for the added compositional variability that these could potentially bring to a model.

The ultimate target of the development of laboratory-based NIRS calibrations for sugarcane bagasse samples is that these can be transferred to an online device at a Brazilian sugarmill. The Author plans that quantitative laboratory calibrations can be transferred and tested before the end of the Brazilian 2011 harvesting season (November).

The early work on developing quantitative NIRS calibrations on pretreated Miscanthus samples has been highly promising. It is planned that this work will be extended with the inclusion of more samples representing different pretreatment conditions and different samples (sugarcane bagasse and trash). NIRS may also be evaluated for potential application in other areas of the DIBANET project, for instance in the determination, using the spectra of the process hydrolysates, of the concentrations of end-products (levulinic acid, furfural etc.). The NIRS spectra of biomass samples (pretreated or not) could potentially be used to predict the ultimate yields of levulinic acid and co-products. Calibrations could be developed based on the predicted compositions of the samples, or by providing data to the calibration set regarding the actual LVA yields achieved. This would allow the PLSR algorithm to determine important wavelengths for this regression. For such predictions the reactor conditions would need to be consistent so that these would not be a factor that could influence end-product yields. The regressions developed could, coupled with known information about which molecular bonds have relevance to the important wavelengths, help to suggest which biomass constituents have an influence on yields in this process.

19 Conclusion

The Summary sections of Chapters 12 to 18 summarise the important observations made in those Chapters. This Chapter will provide overall conclusions regarding the work conducted in this Thesis.

19.1 Quality of the Analytical Data

Chapter 3 details the wide variety of potential techniques available for the wet-chemical characterisation of lignocellulosic materials. The methods that were chosen to be used for this Thesis were selected on the basis that these would provide highly accurate and precise data for the important lignocellulosic constituents and that, in the most important cases, these constituents would be analysed directly rather than indirectly. Using indirect methods, for example gravimetric techniques such as the detergent fibre methods, could potentially result in components other than those of interest being categorised as part of the desired analyte. This is not only a poor result for accurate characterisation of the biomass but also for the development of precise near infrared spectroscopy (NIRS) models since these would need to account for the wide variety of potential constituents that may be assigned to a given gravimetric fraction.

A relevant example is the determination of the Klason lignin content of peat samples. For this characterisation a gravimetric method was used and the resulting NIRS models were less precise than those developed on the direct analyses of many of the constituent sugars. These sugars are the most important constituents for hydrolysis biorefining technologies and have therefore been a major focus in the development of analytical methods for biomass samples in this research. It is considered that the ultimate methods developed, as discussed in Chapters 3 and 4, have allowed for these sugars, particularly the most abundant sugars (glucose and xylose) to be characterised with a very high level of precision in procedures that are reasonably quick compared to other techniques. For instance, process hydrolysates only need to be diluted prior to introduction to the ion chromatography (IC) system where all sugars can be resolved in approximately half an hour. In contrast, gas chromatography requires a lengthy derivatisation procedure to be undertaken to transform the sugars to their alditol acetate derivatives in order that they will be volatile for introduction to the gas chromatography (GC) column.

If the precision of wet-chemical analysis is determined by the standard error of laboratory (SEL) of an analyte, then superior precision, for most constituents, has been obtained for the Miscanthus samples. This is primarily a result of the number of samples available for analysis and the much larger quantities of each sample when compared with many of the peat and, particularly, bagasse samples. This allows for repeat analysis of a sample if imprecise analytical data are obtained from its analysis.

The lack of sufficient quantities of sample was a hindrance in the analysis of the bagasse samples and is the main reason why the SELs for this feedstock are often greater than those for Miscanthus. There is therefore likely to be a knock-on effect regarding the levels of precision possible in NIRS models. Regarding the peat samples, the larger SELs often seen for these are a result of the increased heterogeneity of the samples, even when processed to the dry and sieved (DS) state. This was also observed for the various compost, brown bin waste, and animal manure samples, as is demonstrated by the standard deviation of duplicates (SDD) values for many constituents in Chapter 17. Since the density of these samples can be greater than that of the bagasse/Miscanthus samples, but the weight used in analytical hydrolysis is maintained, the variable effects of individual particles carry greater weight in these samples.

Solutions to this would involve a greater quantity of material being hydrolysed allowing for these variable effects to be decreased. That was not undertaken in this research since it would have required the purchase of a new set of laboratory equipment (pressure tubes, autoclave, filter crucibles etc.) to accommodate the increased volumes of hydrolysate that would result. Such an investment was not considered warranted given that the main feedstock of interest for quantitative NIRS calibration was Miscanthus. An alternative to using more sample would be to process all of the sample down to a much finer and more homogeneous particle size (e.g. something similar to DF (dry fines)) and then hydrolyse that. Sample presentation and handling issues do become more problematic with reductions to a very small particle size, however. Also, tests would need to be conducted to ensure that these new conditions do not lead to differential rates of sugar degradation.

If there is an interest in developing quantitative NIRS models for some of the less precisely characterised samples (e.g. manures, composts, municipal wastes) then employing either/both of these two strategies (increased sample size, increased sample homogeneity) could aid in improving the resulting regression statistics. However, Chapter 13 demonstrates that, even with increased SELs, good quantitative models could still be developed for peat samples using the standard methods.

As a result of this research a number of standard lab practice methodologies with step by step procedures have been written by the Author. These have been used by assistants/interns and other researchers in the Carbolea laboratories and have also been distributed to DIBANET partners CTC and UNICAMP. These procedures cover the preparation, NIRS analysis, ethanol-extraction, and analytical hydrolysis of samples. There are also data recording sheets for writing the results of the experiments and Microsoft Excel spreadsheets for inputting the data. These spreadsheets are all interlinked and enable a user to obtain the full chemical composition and all relevant details of a sample just through inputting its sample number. The standard methodologies and data recording sheets are available to download from the DIBANET website: <http://www.dibanet.org/links.php>.

19.2 Differences Between the Spectral Datasets

As discussed in Chapter 9, the vast majority of the research in the literature regarding the development of quantitative NIRS models for lignocellulosic constituents has focussed on the spectra of dry sample of a homogeneous particle size (e.g. DS). A prime target of the Author's research has been to, not only consider samples in this state, but also samples with less intensive methods of sample preparation (or no sample preparation at all, if possible). Chapter 11 details the reasoning behind the selection of the WU (wet and unground), DU (dry and unground), DG (dry and ground), and DS datasets for model development and the hypothesis that the accuracy and precision of NIRS analysis should increase from the first (WU) to the last (DS) of these. It was an aspiration that accurate WU models could be developed but the hindrances in achieving these were well understood.

The results of this research have not only demonstrated that accurate WU models can be developed but also have shown that, in contrast to the original hypothesis, these can sometimes be superior to models developed on dry feedstocks, particularly the DU models. This Thesis has shown, through the use of NIRS models on the replicate scans of samples, and through the differences in precision observed between models developed on the various datasets, that sample presentation to the NIR cell is of key importance. The WU samples reduce the particle size biases that can occur when placing dry samples into the cell. These biases can be significant for the DU samples of some feedstocks (e.g. Miscanthus and bagasse) and can result in a disproportionate amount of fine material being scanned whilst the larger particle size fractions are not allowed to interact with the NIR radiation.

The DT sample presentation method (a new method for DS samples), which has only been used to date on *Miscanthus* samples, improved the reproducibility when predicting replicate scans of the same material. For some constituents it also improved the root mean square error of prediction (RMSEP) statistic and other statistics of the NIRS models. However, the samples in this state ($180 \mu\text{m} < X < 850 \mu\text{m}$) are reasonably homogenous, at least when compared to the DG and, particularly, DU sample states. An insufficient number of samples have been scanned using the DH (a new method for DG samples) and DV (a new method for DU samples) protocols to allow for conclusions to be drawn concerning whether these offer improvements for the development of NIRS models compared with the standard DG and DU methods; however, it seems logical that they should. With regard to the bagasse samples, a lengthier alternative to the DV spectral collection method was used whereby the DU sample was sieved and a weighted average spectrum of the spectra of the various particle size fractions was determined. This method (DW) allowed for more precise NIRS models for some constituents, although in some cases the differences were minor.

Of course, sample presentation methods would be of much less importance if all particle size fractions of the biomass had similar chemistry, but the analyses of the DF fractions of *Miscanthus* samples show that this is not the case. These DF samples were the most homogenous and provided the most consistent predictions when NIRS models were applied to replicate scans of the same sample. The combination of the DF and DS data in a weighted average for NIRS model development (i.e. the WU_{DG} and WU_{DGP} models, see Section 15.1.1) also resulted in more precise models than for using the DS data only.

19.3 Wavelength Regions for Regression

The FOSS XDS NIR unit that was used for the collection of spectra in this Thesis allows for absorbances over the wavelengths 400-2500 nm to be recorded. Chapter 9 and the data provided in the Tables in Appendix B show that accurate NIRS models do not necessarily need this whole range and, in some cases, superior models can result from using limited wavelength regions. As part of this research numerous spectral ranges were examined for partial least squares regression (PLSR) model development, particularly for the WU datasets. Many of the results for the wavelength regions tested are provided in Chapter 12, which discusses the development of models for bagasse samples. Similar tests were also carried out for the peat and *Miscanthus* samples although, for the most part, the results of these tests are not provided in order to reduce the final size of the Thesis. The final

results presented in the various chapters represent the best wavelength regions found for the respective feedstock, dataset, and constituent. In most cases the visible and short-wavelength-NIR region (400-1100 nm) did not aid in the development of accurate quantitative NIRS models and the 1100-2500 nm region was chosen instead. It may initially seem strange that a model developed on a limited wavelength region can be more precise than one that includes this, and other, regions (e.g. a 400-2500 nm model). However, it is logical when considering that, while PLSR includes the variance of Y and the covariance of Y and X in the development of its loading weight vectors, it also needs to model the variance of X (i.e. the spectra). The PCAs that were carried out on full wavelength (i.e. 400-2500 nm) spectra for bagasse, peat and Miscanthus all demonstrated high loadings in the visible region due to the greater degree of relative variation in this region. PLSR models will, therefore, still need to account for this and so excluding these regions in cases where they do not provide large discriminatory power for the constituent of interest will result in a simpler model and loading vectors more focused towards the relevant spectral variability.

There have been some instances where the inclusion of the 400-1100 nm region has improved the model. These include the prediction of the ash content of bagasse samples and in the discrimination between various categories of Miscanthus samples (Chapter 14). Indeed, in the level of sample discrimination good models have been shown to be possible using only this region, or only the visible region.

Regarding the WU models, in some instances slightly improved RMSEPs were possible using only the 1100-1800 nm region. This region avoids the large absorbances attributed to the molecular bonds of water at around 1930 nm, although the moisture absorbances around 1390 nm will still need to be modelled. The regression coefficients plots for the dry datasets (e.g. DS) often show that there are important wavelengths for the model in the wavelengths beyond 1800 nm, for instance at 2100 and 2270 nm for cellulose and at around 1900 nm for lignin. However, the WU models show that these regions can be avoided and accurate NIRS models can still be developed. In most cases similar wavelengths in the 1100-1800 nm region demonstrate important regression coefficients in both the full region and limited region models; however, their importance is magnified in the limited region models. For example, the lignin absorbance at 1662 nm is much more important in the 1100-1800 nm models than it is in the 1100-2500 nm models.

All of the tests carried out in this Thesis regarding regions of the spectra (windows) to use for PLSR model development have been done manually. There are automated methods for examining the effects of varying these windows. One of the techniques used for picking the most useful spectral interval for calibration is known moving windows PLSR (Du et al., 2004), MWPLSR. In this technique a

series of PLS models are constructed for every window that moves across the whole spectral region. If h is the window size there will therefore be $(k-h+1)$ windows over the whole spectrum, where k is the total number of wavelength variables in the full spectrum. Changeable size moving window PLS (CSMWPLS) investigates how also varying the value for h can potentially further improve the calibration (Du et al., 2004).

There is also a method known as Iterative Predictor Weighting PLS (IPW-PLS) (Forina et al., 1999). This involves a cyclic repetition of the PLS algorithm with model-wise elimination of useless predictors (wavelengths) after each step. Forina *et al.* (1999) tested the IPW algorithm on a data-set consisting of 176 predictors (variables) from the NIR spectra of 60 samples of soy flour. Models were developed for moisture, protein and oil content. It was found that IPW retained the least number of predictor variables (6 for moisture) and had a lower SEP than standard PLS. There is a derivation of this method, developed by Chen *et al.* (2004), known as Modified Iterative-Predictor Weighting PLS (mIPW-PLS). In testing this method the authors took a dataset of the NIR spectra of 38 samples of four-sugar aqueous solutions and attempted to calibrate for the glucose content and fructose content using, among other methods, PLS, IPW-PLS, and mIPW-PLS. It was found that mIPW-PLS reduced the number of predictor variables compared with IPW-PLS and also improved the predictive accuracy.

Neither MWPLSR or the IPW-PLS methods are provided in the Unscrambler X software; however, the Author understands that there are tools for their utilisation in MATLAB. It could be interesting to see if these could improve the regression statistics obtained in this study, although care would need to be taken in the IPW-PLS method to ensure that spectra covering a wide variety of moisture contents are presented to the model in order to ensure that the variable bathochromic and hypsochromic effects of moisture and hydrogen bonding are represented in the spectra so that wavelengths relevant to these are not excluded from the model.

19.4 Spectral Pretreatments

How the spectra are transformed is also an important consideration in NIRS model development. The literature review (Chapters 8 and 9) indicated that the transformation of spectra often helps but that there is no clear situation regarding what are the best treatments to use. Numerous transformations were attempted in this research for all feedstocks, constituents, and datasets and there was no definite winner that provided superior models in all instances. There were some

general observations that were made, however. Firstly, with the exception of the WU bagasse models, the WU models developed on spectra transformed by scatter correction methods (standard normal variate (SNV), standard normal variate and detrend (SNVDT), multiplicate scatter correction (MSC), extended multiplicate scatter correction (EMSC)) tended to have poorer predictive abilities than the models developed on spectra transformed by Savitzky-Golay derivatives. Indeed, sometimes these models were less accurate than those developed on the raw spectra (see Table F-14). The relative poor performances of the models based on scatter-corrected spectra are in agreement with observations made by other researchers regarding the development of models for WU samples (Cozzolino et al., 2006a, Dardenne et al., 2000).

Secondly, these scatter correction methods also perform less well when they are applied to the whole spectrum rather than just the NIR region. This is because the noise in the visible region brings forward added spectral variability that hinders, for example, the ability of the MSC coefficients to remove particle size effects from a dataset.

Regarding the derivatives, the relative advantages provided by different derivative orders, polynomial fitting orders, and smoothing points, varied over the numerous models developed. In some cases (e.g. the xylose models developed for the bagasse samples) the best derivative treatment offered a significant improvement over many of the others but in many other cases (e.g. the large *Miscanthus* datasets) the models based on different derivative treatments were all reasonably similar (see Table F-14). It was noticed, however, that there was a limit (around 40 nm) for the region used for smoothing beyond which the loss in spectral detail resulted in less precise models. In general a second derivative is preferred for interpretation since it maintains peak location and does not result in the production of too many artefact peaks (as can be the case with the fourth derivative).

In terms of the qualitative analysis and discrimination of samples (Chapter 14) the use of spectral pretreatments often led to less accurate models. This was particularly the case for the automated clustering methods used for partitioning *Miscanthus* samples to either the leaves or stems groups. Particle size effects can result in significant differences between spectra and it may have been the case that the inherent differences in particle sizes between different plant fractions aided in their discrimination; so explaining the less accurate performance of models where these effects are reduced.

19.5 Quality of the Quantitative NIRS Models

The Reader is encouraged to consult with the Tables in Appendix B, based on the literature review of previous NIRS studies on lignocellulosics, when assessing the quality of the models developed in this Thesis. The most important regression statistics in these Tables are those based on the validation set, i.e. R^2 , SEP (standard error of prediction), RPD (ratio of standard error of performance to standard deviation), and RER (range error ratio). If the RMSEP was reported then it is highlighted in italics in the SEP column. The SEP measures the precision of the prediction but the RMSEP measures the accuracy, and there can be a significant difference between the two (with the SEP being lower) if there is a large bias. It is considered by the Author that the RMSEP is a much better statistic for assessing the performance of the model than the SEP; however this statistic was not provided in many of the articles read by the Author.

In any case, it can be seen that the results obtained in this Thesis, particularly those for WU samples, are superior to the majority of those in the literature. For Miscanthus, RER ratios over 15 are possible for the glucose, xylose, total sugars, and Klason lignin (KL) contents of WU samples. These are the most important constituents with regard to predicting the yields when biomass is processed in hydrolysis biorefining technologies. In demonstrating that these can be predicted with a high degree of accuracy this research has shown that there is potential for the utilisation of online NIRS systems in biorefineries to characterise in real time the important lignocellulosic properties of biomass. These systems could be used to potentially modify process conditions according to the characteristics of the feedstock, segregate different types of samples, and pay the biomass suppliers according to the quality of the material provided.

The transfer of laboratory NIRS models to an online system is one of the objectives of the DIBANET project. This will take place at a later date using the calibrations developed for sugarcane bagasse samples. DIBANET (see Chapter 18) partner CTC is responsible for the development of these models but to date no complete set of analytical data exists for any of the samples that CTC have collected. The Author has tried to evaluate the suitability of the bagasse samples that have been scanned, using the FOSS XDS at CTC, for NIRS model development by projecting these onto the Miscanthus and Australian bagasse models. It appears that the loading vectors of the Miscanthus models offer a reasonable approximation in explaining some of the spectral variability in the CTC datasets and the Miscanthus DS PLSR model has been used to predict the glucose, xylose, and KL compositions of the CTC samples.

If the ranges of the predicted concentrations for CTC bagasse and trash are similar to those of the real samples then these suggest problems in the development of accurate NIRS calibrations for bagasse. These problems were also noted in the work on the BSES bagasse samples, where the chemical variability among the samples was limited. In the case of these BSES samples, this resulted in some of the PLSR models that were developed being poor predictors of composition. While good calibrations were possible for glucose and total sugars, the models for xylose content and other constituents were much less precise. This is a concern regarding the feasibility of using online NIRS for the real-time characterisation of bagasse samples as part of the DIBANET project. In September the Author will travel to CTC and attempt to strategically select samples for analysis so that the concentration range for all important constituents is as wide as possible in the hope that this can lead to improved models. It is considered that the depth of knowledge that has been learned in the development of the Miscanthus and BSES models will allow this task to be carried out before the end of the sugarcane harvesting season in Brazil so that the models can be tested online and a decision reached as to whether the development of an online bagasse characterisation system is continued. Even if it is not warranted for this feedstock, however, there does appear to be a great potential for online NIRS to be employed in biorefineries, particularly those that may receive a diverse mix of feedstocks.

The accuracies of the peat models tended to be in-between those of the other two feedstocks; better than bagasse but less precise than Miscanthus in many cases. There are few articles in the literature concerning the use of NIRS for the characterisation of peat samples, particularly for the range of constituents analysed for in this Thesis, however, and it is considered by the Author that the results presented in Chapter 13 are superior, in many cases, to what has been published to date. The data presented in that Chapter have also shown that peat samples do not have sufficient quantities of polysaccharide sugars to allow for their processing in hydrolysis biorefining technologies and, hence, the WU models developed are unlikely to be integrated in biorefineries. However, models developed on moisture content and ash may have value for the current uses for peat (combustion in power stations) whilst the lignocellulosic models could provide value for researchers looking for rapid peat characterisation methods to inform other studies (e.g. soil amendment).

It is noteworthy that the research presented has demonstrated that good NIRS models can be developed for minor constituents (e.g. rhamnose), and for constituents which, according to the literature review, have only had imprecise models developed for to date (e.g. acid soluble lignin (ASL), see Table B-8). Furthermore, such accurate models can be developed even on wet spectra. For

example, the WU model for the rhamnose content of *Miscanthus x giganteus* samples (Table F-20) has an RMSEP of 0.03% and an R_{pred}^2 of 0.845 whilst the ASL model for the same dataset (Table F-23) has an RMSEP of 0.29% and an R_{pred}^2 of 0.929.

There are some examples of constituents for which the models developed have been poor, however. For example, the calibrations for the carbon contents of *Miscanthus* samples were much less accurate than many of the other models whilst the hydrogen and sulphur contents of this feedstock varied so little that R_{pred}^2 values were less than 0.5 for all datasets. The models for uronic acids were also less accurate than those for other carbohydrates. There are relatively few publications regarding the development of quantitative NIRS models for UA. However, those that were found (Table B-7) also performed relatively poorly. The lack of precision of the UA models is not so important, however, given the relatively small proportions that these contribute to the total mass balance of samples and that the value/hindrance of UA to biorefining processes is still unclear.

19.6 Suitable Feedstocks for Biorefining in Ireland

An earlier paper by the Author (Hayes and Hayes, 2009) collated secondary compositional data regarding some feedstocks that were considered to have potential for utilisation in biorefining processes. These data were linked with estimated efficiencies for these processes in order to determine the yields that could be achieved and these were then projected to a national scenario based on the known quantities of these resources. Chapter 17 details the primary analytical data collected for many of these feedstocks and the conclusions that were reached regarding their suitability for these biorefining processes. Those conclusions will not be repeated here. However, the general trend was that many of the feedstocks that were considered to have potential based on the secondary data appeared less attractive once the primary data were considered. Such examples include spent mushroom compost and non-paper organic wastes. Autoclave-treated black bin waste was a new type of feedstock, not considered in the previous article, that appeared to have potential for utilisation in hydrolysis biorefining technologies, however.

It was concluded that a short term scenario for the deployment of a commercial-scale biorefinery in Ireland would be to utilise all the waste paper material that is currently exported from the country. Such a scenario was attractive given the mechanisms for collecting and handling this feedstock are already in place and papers/cardboards have large total sugars contents which allow for high yields in hydrolysis biorefining technologies.

Such a facility could potentially function as the catalyst for the development of supply cycle logistics and processes to allow for other feedstocks to be sourced for utilisation in biorefineries; given that a known purchaser for these feedstocks would be operational. This scenario makes more sense in the Irish context than developing, in the first instance, a biorefinery dedicated to the processing of less well established feedstocks, such as Miscanthus.

Based on the analytical data presented in this Thesis, however, Miscanthus does have great potential as a feedstock. Furthermore, providing the nutrient loss from the stand can be corrected for, the harvesting of this resource at a much earlier point in the harvest window than under current regimes (where a low moisture content is of paramount importance) would allow for significantly improved yields and revenues for the farmer. The yields per tonne of feedstock would be less at the biorefinery, however, but a potential harvest window of between October and April (i.e. 7 months) would be attractive for biorefinery operators in that the seasonality of supply issues would be less severe than under a situation where all Miscanthus is sampled at a “Late” harvest (March, April). Sustainable practices could therefore allow for the strategic harvesting of Miscanthus whereby in one year a site may experience an early harvest but in others the harvest is later in order to prevent excessive loss of soil nutrients. There would still be the question of what feedstock is supplied to the biorefinery in the seasons during which Miscanthus is growing and can not be harvested. However, there are options that could be examined here; such as the production of short rotation coppice willows (which can be harvested all year round).

19.7 Possible Future Developments of the Research

In developing accurate calibrations for the most important constituents of wet samples, this research has achieved its goals and provided data of value to the scientific community and biorefining industry. There are also several relatively easy ways upon which the work in this Thesis can be built upon in a reasonably short timeframe. For instance, there is a very large spectral database covering several hundred samples in their various states of preparation (WU, DS, etc.). Unless all of the sample has been consumed in analysis, the DS and DF fractions are retained in the Carbolea laboratories. Therefore, the potential exists for expanding the number of models for these samples to include new biomass properties. For example, cellulose crystallinity could be analysed for via X-ray diffraction (XRD) and the variation in the crystallinity index linked to the spectra. The crystallinity of cellulose is considered by some to be an important determinant of the rates and

yields of hydrolysis processes (Yoshida et al., 2008, Kerley et al., 1988). There has been some work to date relating to the prediction of cellulose crystallinity using NIRS (Basch et al., 1974, Jiang et al., 2007), but not for Miscanthus. The Author undertook one experiment whereby samples of various Miscanthus fractions were analysed using an XRD device and it was found that there were differences in the crystallinity indices computed for these. Hence, there may be sufficient variability in the crystallinity of cellulose in Miscanthus to allow for NIRS models to be developed, if these are so desired. Models predicting the cellulose crystallinity of pretreated samples could also be of interest.

Other properties of potential relevance to biorefining that could be developed for existing samples include the degree of polymerisation of cellulose, the acetyl content of biomass, and the actual quantities of the polysaccharides (rather than their constituent sugars) present. Quantifying the polysaccharides would be quite a lengthy and complicated procedure, however, as mentioned in Chapter 3.

The key point is that much of the hard work relating to laboratory analysis, the collection and processing of samples, has already taken place meaning that only the reference analysis would need to be carried out. The knowledge developed to date regarding these samples could also inform a much more strategic means for sample selection, meaning that much fewer samples would need to be analysed in order to develop these NIRS models than was the case previously. For example, samples for cellulose crystallinity analysis could be selected on the basis of their (known) glucose content.

In addition to the DS and DF samples, all of the WC samples of Miscanthus (i.e. those that had their spectra collected but were not processed further) are retained in freezers in the Carbolea laboratories. These could also potentially be utilised for future research, for example if their spectra are selected as important by current/future models. Also, in some cases, there was such a large amount of WU material that not all of it was needed for the preparation of DS/DF samples and the remainder was retained in the freezer. The DIBANET project proposal included a Task for low field NMR (LF-NMR) equipment to be examined for the potential to develop quantitative models for the lignocellulosic parameters of interest (in a similar concept to NIRS but, clearly, using different methods of interaction with the sample). Originally, this work was to be conducted by another partner in the DIBANET project but that partner now refuses to do this work. It has been suggested that a LF-NMR system at Dublin City University (DCU) could be used by the Author for the analysis. Such a body of work could be completed reasonably quickly using wet Miscanthus samples of known compositions. To date, there has been limited use of LF-NMR in characterising lignocellulosic biomass and there are only a few publications for extractives, cellulose crystallinity and free/bound

water in relevant feedstocks (Labbé et al., 2002, Casieri et al., 2004). An exploratory study with a limited number of WU samples should indicate if further examination of this technique is warranted.

As described in Section 3.2.2, a diode array UV-Vis (ultraviolet-visible) spectrophotometer was used for the analysis of the biomass hydrolysates in order to determine the ASL content of the sample. This device records spectra over the 190 to 510 nm wavelength region. The spectra for each sample have been exported to a Microsoft Excel file where modifications to the chosen wavelength and extinction coefficient can be made and the effects on the predicted ASL content noted. The Author considers that there may be potential in resolving other information from these spectra using the chemometric techniques developed in this Thesis. For example, it might be possible to develop models for the hexose and pentose contents of the hydrolysate, and ultimately, of the biomass sample itself. This is because it is known that furfural and hydroxymethylfurfural absorb in the UV-region (see Section 3.2.3) and that these are degradation products of the secondary hydrolysis stage. A higher relative proportion of sugars in the hydrolysate at the primary hydrolysis stage would therefore result in greater quantities of these degradation products in the final hydrolysate, due to the degradation of these sugars in the autoclave treatment, and these could therefore be analysed in the UV-Vis device. Since the carbohydrate contents of these hydrolysates are already known from their IC analysis, it would be relatively easy to see if PLS models could be developed based on the UV-Vis spectra.

It is also planned to see if NIRS models can be developed for various constituents using the spectra of the biomass hydrolysates. Most of the hydrolysates obtained from the analytical hydrolysis batches have been stored frozen in the Carbolea laboratories. These could be removed from the freezer, given time to defrost and for their temperature to stabilise, and then analysed using the liquid analyser apparatus (as was used for the animal manure samples) in the XDS unit. These spectra could then be linked with the known compositions of these hydrolysates and quantitative models developed. The concentrations of these sugars would be quite low in the hydrolysate, however. For instance, a biomass sample with a 40% glucan content would provide a hydrolysate with a glucose concentration of 1.32 mg/ml (using the standard methodology outlined in Section 11.5), equivalent to 0.132% of the hydrolysate. In order to achieve an RMSEP of 1% glucan content for the biomass sample, the model would need to have an RMSEP for the glucose content of the hydrolysate of 0.0033% or 33 ppm. Blanco *et al.* (2007) found that 100 ppm was the detection limit for determining the concentration of 2-ethylhexanol in an industrial ester. If a similar detection limit could be achieved for glucose it would equate to an RMSEP of 3% for the glucan content of the original biomass sample. Hence, the NIRS analysis of the hydrolysate may not be able to provide the

level of precision achieved in the NIRS analysis of the biomass, however models developed for hydrolysates may be applicable to a wider variety of feedstocks given that many of the interfering constituents of the biomass have been removed or deconstructed. NIRS may therefore have value in the rapid determination of the analysis of the hydrolysates of unknown samples prior to their analysis on an IC system; however experimental work will be necessary to see if this is the case.

The concentrations of the important products of the DIBANET process (levulinic acid, furfural etc.) are likely to be present in higher concentrations in the hydrolysate from this process than are the sugars that result from the analytical-hydrolysis procedure. As discussed in Section 18.4, NIRS may have potential in determining their levels and so be of use in rapid screening tests that can be carried out in advance of the more accurate IC methods.

Prior to the analysis of the straw samples it was considered that these could be another potential feedstock suitable for the development of quantitative NIRS models and enough samples (62) were collected by the Author to allow for this (not all of these samples have been characterised to date). However, the results discussed in Section 17.2.1 indicate that the variation in lignocellulosic components among the samples analysed has been relatively low indicating that the development of accurate NIRS models may be difficult. Early models based on the data that do exist could be developed and applied to the spectra of the as-yet not analysed samples and the predicted values (and deviation in prediction) observed to see if these could potentially bring more variation to the model. However, with the exception of the rapeseed straw samples, all of the samples that have not yet been analysed are from species types (e.g. winter wheat) that have had at least two samples analysed to date. As discussed in Section 17.2.1.3, the relative variation within species type was much smaller than the variation between species so it is possible that the as-yet unanalysed samples will not bring great variety to the dataset. In such a case the development of NIRS straw models is probably not warranted and these would also be of less novel scientific interest (compared to the bagasse, Miscanthus, and peat models) since no WU spectra exist for these samples (they were collected dry).

It was also initially considered that spent mushroom compost could be an attractive feedstock for biorefining and that NIRS models could be developed for its composition. However, the low total sugar contents obtained for these samples indicate that this is a feedstock of no known value, to date, for biorefining. Hence, the work involved in developing NIRS models for it would not be justified.

In summary, there are many possible avenues in which the research presented in this Thesis can be expanded upon as well as other lines of research that once appeared promising but now, as a result of the observations presented, may not warrant further investigation. However, the main activities that are planned by the Author following completion of his PhD viva relate to the development of quantitative NIRS models for the Brazilian sugarcane bagasse and sugarcane trash samples collected by CTC and the deployment of these in an online system. The Author also expects to write at least 8 scientific papers based on the results presented in this Thesis.



UNIVERSITY *of* LIMERICK

OLLSCOIL LUIMNIGH

**Analysis of Lignocellulosic Feedstocks
for Biorefineries with a Focus on
The Development of Near Infrared
Spectroscopy as a Primary Analytical Tool**

Volume 2 of 2 (Appendices and References)

Thesis Presented for the award of Doctor of Philosophy (Ph.D.)

By

Daniel J. Hayes

University of Limerick

Supervisor: Dr J. J. Leahy

Submitted to the University of Limerick, July 2011

Table of Contents

<u>APPENDIX A</u>	<u>LIST OF COMMONLY USED ABBREVIATIONS</u>	<u>A-1</u>
<u>APPENDIX B</u>	<u>TABLES FOR CHAPTER 9: LITERATURE REVIEW OF QUANTITATIVE NIRS CALIBRATIONS</u>	<u>B-1</u>
<u>APPENDIX C</u>	<u>FIGURES AND TABLES FOR CHAPTER 12: ANALYSIS OF SUGARCANE BAGASSE</u>	<u>C-1</u>
<u>APPENDIX D</u>	<u>FIGURES AND TABLES FOR CHAPTER 13: PEAT</u>	<u>D-1</u>
<u>APPENDIX E</u>	<u>FIGURES AND TABLES FOR CHAPTER 14: QUALITATIVE ANALYSIS OF MISCANTHUS</u>	<u>E-1</u>
<u>APPENDIX F</u>	<u>FIGURES AND TABLES FOR CHAPTER 15: DEVELOPMENT OF NIRS QUANTITATIVE CALIBRATIONS FOR MISCANTHUS SAMPLES</u>	<u>F-1</u>
<u>APPENDIX G</u>	<u>FIGURES AND TABLES FOR CHAPTER 16: LIGNOCELLULOSIC PROPERTIES</u>	<u>G-1</u>
<u>APPENDIX H</u>	<u>FIGURES AND TABLES FOR CHAPTER 17: ANALYSIS OF WASTE AND OTHER FEEDSTOCKS</u>	<u>H-1</u>
<u>APPENDIX I</u>	<u>FIGURES AND TABLES FOR CHAPTER 18: THE DIBANET PROJECT</u>	<u>I-1</u>
<u>APPENDIX J</u>	<u>REFERENCES</u>	<u>J-1</u>

Appendix A List of Commonly Used Abbreviations

Table A-1: Summary of the abbreviations used in this chapter.

ADF	Acid detergent fibre
ADL	Acid detergent lignin
AF	Ash free basis
AIA	Acid Insoluble Ash
AIA_EF	Acid Insoluble Ash (extractives free)
AIR	Acid Insoluble Residue
AIR_EF	Acid Insoluble Residue (extractives free)
ARA	Arabinose content
ARA_EF	Arabinose content (extractives-free)
ARA_EF_SRS	Arabinose content (extractives-free) using the individual sugar recoveries from the batch
ARA_SRS	Arabinose content using the individual sugar recoveries from the batch
ASA	Acid soluble Ash
ASE	Accelerated solvent extraction
ASL	Acid soluble lignin
ASL_EF	Acid soluble lignin (extractives free)
AV.	Average
Av abs. diff	Average absolute difference (see Section 13.3.4)
BC	Bray Curtis distance
BMW	Biodegradable municipal waste
BSES	Bureau of Sugar Experimental Stations
C	Carbon
Cal:Val	Number of samples in the calibration and validation sets
CB	City block distance
CTC	Centro de Tecnologia Canavieira, a partner in the DIBANET project
CV	Cross-validation.
DB	Dry (hand) sieved fraction at BSES
DCM	Data collection method (for the FOSS XDS NIR system).
DF	A dry fine sample with a particle size less than 180 microns.
DF-E	Moisture content of DF samples prior to the removal of extractives
DF-E Dishes	Moisture content of DF samples after the removal of extractives
DG	Dry and ground sample (all particles less than 850 microns in diameter).
DH	A DG fraction that has been scanned in a different method (see Section 11.1)
DM	Dry matter
DMB	Diesel miscible biofuel
DS	Dry sieved fraction of a sample with a particle size between 180 and 850 microns.
DS-E	Moisture content of DS samples prior to the removal of extractives
DS-E Dishes	Moisture content of DS samples after the removal of extractives
DT	A DS fraction that has been scanned in a different method (see Section 11.1)
DU	Dry unground fraction scanned at BSES
DW	Weighted average scan of several DU fractions, scanned at BSES
E-H	Moisture content of DS samples prior to the analytical hydrolysis procedure
EF	Extractives-free basis
EIA	Ethanol Insoluble Ash
EIA_EF	Ethanol Insoluble Ash (extractives-free)
EMSC	Extended multiplicative scatter correction.
ESA	Ethanol Soluble Ash
ESTD	External standard
EU	Euclidean distance

EXTR_CV	Extractives content (% whole mass basis) as measured directly in collection vials
EXTR_PD	Extractives content (% whole mass basis) as measured by the mass loss in ethanol extraction
“F”	Dead leaf blade sample of Miscanthus
F	Factor (in PLSR)
F-F Test	Number of PLS factors chosen using the F-test criterion (Osten, 1988)
F-Haaland’s	Number of PLS factors chosen using the Haaland and Thomas (1988) criterion
F-Min Press	Number of PLS factors associated with the minimum PRESS value
F-UNSCR	Number of PLS factors chosen by the Unscrambler X software
F-Wold’s	Number of PLS factors chosen using Wold’s criterion
F-Wold’s 0.95	Number of PLS factors chosen using Wold’s criterion and a threshold value of 0.95
F-Wold’s 0.9	Number of PLS factors chosen using Wold’s criterion and a threshold value of 0.90
FL	Miscanthus flower sample
GAL	Galactose content
GAL_EF	Galactose content (extractives-free)
GAL_EF_SRS	Galactose content (extractives-free) using the individual sugar recoveries from the batch
GAL_SRS	Galactose content using the individual sugar recoveries from the batch
GLU	Glucose content
GLU_EF	Glucose content (extractives-free)
GLU_EF_SRS	Glucose content (extractives-free) using the individual sugar recoveries from the batch
GLU_SRS	Glucose content using the individual sugar recoveries from the batch
“H”	Dead leaf sheath sample of Miscanthus
H	Hydrogen
ha	Hectare
HAL	Hierarchical average linkage clustering
HCL	Hierarchical complete linkage clustering
HML	Hierarchical median linkage clustering
HP	Harvested plant sample of Miscanthus
HPAEC-PAD	High performance anion exchange chromatography with pulsed amperometric detection.
HSL	Hierarchical single linkage clustering
IC	Ion chromatography
ISTD	Internal standard
K	Live leaf blade sample of Miscanthus
KDC	K-Medians clustering
KL	Klason lignin
KL_EF	Klason lignin (extractives free)
KMC	K-Means clustering
KURT/KRT	Kurtosis statistic
λ	Wavelength
LA	Levulinic acid
LDA	Linear discriminant analysis.
LvA	Levulinic acid
M	Live leaf sheath sample of Miscanthus
MAN	Mannose content
MAN_EF	Mannose content (extractives-free)
MAN_EF_SRS	Mannose content (extractives-free) using the individual sugar recoveries from the batch
MAN_SRS	Mannose content using the individual sugar recoveries from the batch
MC	Moisture content (wet basis)
MIR	Mid infrared
MLR	Multiple linear regression
MSC	Multiplicative scatter correction.
N	Nitrogen
NDF	Neutral detergent fibre
NIR	Near infrared
NIRS	Near infrared spectroscopy

odt	Oven dried tonnes
PC	Principal Component
PCA	Principal Component Analysis
PCR	Principal Component Regression
PLS	Partial Least Squares
PLSR	Partial Least Squares Regression
PLS-DA	Partial least squares discriminant analysis
PLS- λ	Wavelength region used for PLSR
Pre.	Pretreatment
PRESS	Prediction residual error sum of squares
QDA	Quadratic discriminant analysis
R^2	Multiple correlation coefficient
R^2_{aal}	Multiple correlation coefficient for calibration
R^2_{valid}, R^2_{pred}	Multiple correlation coefficient for (independent) test-set validation
R^2_{CV}	Multiple correlation coefficient for cross validation
RCA	Rapid content analyser (part of the FOSS XDS NIR unit).
RER	Range error ratio
RER_{CV}	RER in cross-validation
RER_{pred}	RER using the independent validation set
RF	Reponse factor
RRF	Relative response factor
RHA	Rhamnose content
RHA_EF	Rhamnose content (extractives-free)
RHA_EF_SRS	Rhamnose content (extractives-free) using the individual sugar recoveries from the batch
RHA_SRS	Rhamnose content using the individual sugar recoveries from the batch
RMSEC	Root mean square error of calibration
RMSECV	Root mean square error of cross validation
$RMSECV_{MP}$	Root mean square error of cross validation using the number of PLS factors associated with minimum PRESS
RMSEP	Root mean square error of prediction
RPD	Ratio of standard error of performance to standard deviation (using the validation set)
RPD_{CV}	RPD in cross-validation
RSD	Relative standard deviation = (standard deviation)/average
RT	Retention time
S	Sulphur
Sam. Excl.	Samples excluded from the PLS model
SB	Sugarcane bagasse
SD	Standard deviation
SD_{EF}	Standard deviation of the extractives-free values for a given constituent and dataset
SD_{WM}	Standard deviation of the values (whole dry mass basis) for a given constituent and dataset
SDD	Standard deviation of duplicates
SEC	Standard error of calibration
SECV	Standard error of cross validation
SEL	Standard Error of Laboratory
SEP	Standard error of prediction
SG	Savitzky Golay
SIMCA	Soft independent modelling of class analogy
SKEW/SKW	Skew statistic
SNV	Standard normal variate
SNVDT	Standard normal variate followed by detrend
SPE	Solid phase extraction.
SR	Solid residues
SRS	Sugar recovery solution.
_SRS	e.g. GLU_SRS: Sugar data corrected according to the sugar recovery of the batch

St:Lf	(Number of stem samples):(number of leaf samples)
TOT	Sum of ARA, GAL, RHA, GLU, XYL, and MAN contents.
TOT_EF	Total sugars content (extractives-free)
TOT_EF_SRS	Total sugars content (extractives-free) using the individual sugar recoveries from the batch
TOT_SRS	Total sugars content using the individual sugar recoveries from the batch
UA	Uronic acids
UL	University of Limerick
wb	Wet basis
WC	Wet and chipped – A wet unground sample that has been scanned but not further processed
WM	Whole mass basis
WP	Whole plant sample
WU	Scan of wet unground sample.
X1	1 st metre of a stem
X1N	Nodes from the 1 st metre of a stem
X1T	Internodes from the 1 st metre of a stem
X2	2 nd metre of a stem
X2N	Nodes from the 2 nd metre of a stem
X2T	Internodes from the 2 nd metre of a stem
X3	3 rd metre of a stem
X3N	Nodes from the 3 rd metre of a stem
X3T	Internodes from the 3 rd metre of a stem
XYL	Xylan content
XYL_EF	Xylose content (extractives-free)
XYL_EF_SRS	Xylose content (extractives-free) using the individual sugar recoveries from the batch
XYL_SRS	Xylose content using the individual sugar recoveries from the batch

Appendix B Tables for Chapter 9: Literature Review of Quantitative NIRS Calibrations

The following Tables present statistics relating to the performance of the calibrations discussed in Chapter 9. These are grouped according to the constituent and dataset type (DG, DU, or WU). In cases where there are no values provided this is a result of the relevant data not being presented in the respective article. Summaries of the abbreviations used in the Tables are listed below and are also provided in Appendix A.

Pretreatments:

MSC = multiplicative scatter correction; WMSC = weighted MSC; EMSC = extended multiplicative scatter correction; SNV = standard normal variate; SNVD = standard normal variate and detrend SG = Savitzky Golay; 1D = first derivative; 2D = second derivative; 1,5,5,1 = derivative number, gap size, segment size, second segment size.

Wavelength Range: Units of 1×10^3 nm (i.e. μm).

Numbers in Italics: Root mean square error statistics (e.g. RMSEC rather than SEC etc.).

F: Number of latent variables (e.g. PLS factors) used.

Const. Data: Constituent Data; R = range (%), M = Mean (%), SD = Standard deviation (%).

SEC/SECV/SEP: All in % unless otherwise stated.

RER/RPD: Calculated from SECV if there is no SEP statistic

C + V: Calibration and validation

Table B-1: Statistics for NIRS calibration equations for the cellulose contents of lignocellulosic feedstocks.

Reference	Feedstock(s)	Wav. Range (10 ³ nm)	Pretreatment	F	Calibration Set					Validation Set					Const. Data			
					n	r ²	r ² _{cv}	SEC	SECV	n	r ²	SEP	RER	RPD	R	M	SD	
Cellulose																		
DG																		
(Downes et al., 2010a)	Eucalyptus wood (ground, 1 mm)	1-2.5								32	0.97	1.26						
(Schimleck et al., 1997)*d	Eucalyptus globulus wood (milled)	1.1-2.5	2D	2	11	0.95		1.04										
(Schimleck et al., 1997)*d	Eucalyptus nitens wood (milled)	1.1-2.5	2D	3	16	0.940		0.81		5		1.20						
(Raymond and Schimleck, 2002)*e	Eucalyptus globulus (1 mm)	1.1-2.5	2,20,10	4	90	0.88		0.95		30		0.88	11.36			10		
(Sinnaeve et al., 1994)	Grass-hay (1 mm)	1.3-2.4	SNV D-1551	10	838		0.95		1.30			1.22 *f	23.1			30	26.6	5.53
(Sinnaeve et al., 1994)	Tropical forages (1 mm)	1.1-2.9	MSC D-2551	13	760		0.92		2.11			2.00 *f	23.7			50	20.7	7.45
(Sinnaeve et al., 1994)	Maize (whole plants) (1 mm)	1.1-2.5	SNVDT D-1551	10	902		0.91		1.10			1.04 *f	17.6			19.4	21.5	3.7
(Liu and Chen, 2007)	Rice straw (0.425-0.25 mm)	1.3-2.44	D-1,5,10,1	13	34	0.93			1.24	9	0.934	1.1	7.20	2.55		7.9	33.9	2.8
(Beining et al., 2000)	Freeze dried peat tables (FTIR)	1.25-2.5	None			0.575										8.6		
(Reeves III and Van Kessel, 2002)	Dried dairy manure (0.85 mm)	0.4-2.5			93	0.794												
(Reeves III and Van Kessel, 2002)	Dried dairy manure (0.85 mm) – ash free DM	0.4-2.5			96	0.830												
(Downes et al., 2010b)*g	Eucalyptus woodmeal	1-2.5		11	1089	0.89		0.65	0.70	123	0.72	0.92	9.70			8.9		
(Downes et al., 2010b)*g	Eucalyptus woodmeal	1-2.5		9	1212	0.85	0.86	0.78	0.80	-		-	18.6			16		
(Hodge and Woodbridge, 2010)*h	Pine (various sp.) woodmeal (1 mm)	1.1-2.5	SNV 2D (SG)	10	305	0.84		0.83	1.08	152	0.72	1.10	12.70	1.80		14	43.0	
(Hodge and Woodbridge, 2010)	P. taeda woodmeal (1 mm)	1.1-2.5	SNV 2D (SG)	5	99	0.82		0.86		50	0.75	1.02		2.01			42.4	
(Hodge and Woodbridge, 2010)	P. tecunumanii woodmeal (1 mm)	1.1-2.5	SNV 2D (SG)	6	89	0.85		1.10		44	0.59	1.30		1.54			42.8	
DU																		
(Poke, 2006)	Eucalyptus strips			5	40	0.88		1.14		9	0.69	-						
(Jones et al., 2006) (Alves et al., 2006)(Alves et al., 2006)*a	Pine strips	1.1-2.5	2D	4	28	0.80		1.03	1.86	12	0.57	1.73			1.32			
(Downes et al., 2010a) *b	Eucalyptus disc	1-2.5		4	91	0.91	0.84	0.74	0.97									
(Downes et al., 2010a) *c	Eucalyp. disc segments (hand held unit)	0.94-1.8		7	91	0.93	0.87	0.75	0.94									
WU																		
(Vergnoux et al., 2009)	Sewage sludge compost	1-2.222	None	13	9	0.82		5.9		4	0.92	5.23	4.11			21.5		

*a = cellulose calculated by a formula based on the monosaccharide composition = glucan – (1/3 x mannan);*b = 5 outliers removed;*c = 1 outlier removed;*d = cellulose polysaccharide isolated;*e= crude cellulose content determined according to the diglyme method of Wallis et al. (1997);*f = statistics for a local regression;*g = measured using the diglyme method (Wright and Wallis, 1998);*h = TAPPI procedure T429.

Table B-2: Statistics for NIRS calibration equations for hemicellulose, holocellulose, total pentoses, and total hexoses contents of lignocellulosic feedstocks.

Reference	Feedstock(s)	Wav. Range (10 ³ nm)	Pretreatment	F	Calibration Set					Validation Set					Const. Data		
					n	r ²	r ² _{cv}	SEC	SECV	n	r ²	SEP	RER	RPD	R	M	SD
Hemicellulose																	
DG																	
(Schimleck et al., 1997)*b	Eucalyptus globulus wood (milled)	1.1-2.5	2D	4	11	0.99		0.18									
(Schimleck et al., 1997)*b	Eucalyptus nitens wood (milled)	1.1-2.5	2D	3	16	0.946		0.52		5		0.78					
(Liu and Chen, 2007)	Rice straw (0.425-0.25 mm)	1.3-2.44	D-1,5,10,1	11	34	0.91			1.11	9	0.9065	1.7	6.94	1.94	11.8	25.6	3.3
(Reeves III and Van Kessel, 2002)	Dried dairy manure (0.85 mm)	0.4-2.5			97	0.582											
(Reeves III and Van Kessel, 2002)	Dried dairy manure (0.85 mm) – ash free DM	0.4-2.5			95	0.397											
DU																	
(Jones et al., 2006) *a	Pine strips	1.1-2.5	2D	3	28	0.59		0.92	1.24	12	0.30	1.04		1.14			
WU																	
(Vergnoux et al., 2009)	Sewage sludge compost	1-2.222	None	13	9	0.85		2.13		4	0.99	1.35	6.96		9.4		
Holocellulose																	
DG																	
(Vavrova et al., 2008)*c	Plant litter (1 mm)	0.78-2.5	SNVD	3	56	0.95	0.94	3.07	3.45				14.60	4.12	50.2	53.66	14.21
(Ono et al., 2003)*d	Leaf litter (74 µm)	400-2500	2D (MLR)	3	88	0.88		3.7		44	0.83	3.5	13.57	2.30	47.5	48.6	8.4
DU																	
(Nkansah et al., 2010)	Yellow poplar wood blocks	0.8-2.5	None	7 P	37	0.77		2.93		17	0.68	4.89					
(Nkansah et al., 2010)	Yellow poplar wood blocks	1.3-1.8	None	6 P	37	0.78		2.88		17	0.63	5.18					
(Nkansah et al., 2010)	Yellow poplar wood blocks	0.8-2.5	1D (SG)	4 P	37	0.79		3.75		17	0.81	4.29					
(Nkansah et al., 2010)	Yellow poplar wood blocks	1.3-1.8	1D (SG)	4 P	37	0.64		4.51		17	0.77	4.09					
Total Pentoses																	
WU																	
(Axrup et al., 2000)	Wood chips	0.8-1.1	None	6	83			3.9		35		4.6	3.12		14.34		
(Axrup et al., 2000)	Wood chips	0.8-1.1	MSC	5	83			1.0		35		1.3	11.03		14.34		
Total Hexoses																	
WU																	
(Axrup et al., 2000)	Wood chips	0.8-1.1	None	6	83			6.7		35		8.1	1.89		15.3		
(Axrup et al., 2000)	Wood chips	0.8-1.1	MSC	5	83			2.5		35		2.9	5.28		15.3		

*a = hemicellulose calculated by a formula based on the monosaccharide composition = (arabinan + galactan + glucan + mannan + xylan) - cellulose ; *b = hemicellulose polysaccharide isolated; *c – analysed via the sodium chlorite method (Quaramby and Allen, 1989); *d – holocellulose is isolated according to (TAPPI, 1997, TAPPI, 1998).

Table B-3: Statistics for NIRS calibration equations for ADF and NDF contents of lignocellulosic feedstocks.

Reference	Feedstock(s)	Wav. Range (10 ³ nm)	Pretreatment	F	Calibration Set					Validation Set					Const. Data		
					n	r ²	r _{CP} ²	SEC	SECV	n	r ²	SEP	RER	RPD	R	M	SD
ADF																	
DG																	
(Kong et al., 2005)	Ground rice straws (1 mm)	1.1-2.5	2861 SNVD	7	136	0.96		0.71	0.82	71	0.959	0.933	20.86	4.95	19.46	36.76	4.62
(McTiernan et al., 2003)	Decomposing Pinus sylvestris needles (1 mm)				110	0.95		1.00	1.59			0.93	21.08	4.73	19.6	53.9	4.4
(Hodgson et al., 2010)	Miscanthus varieties (1 mm)	0.4-2.5	SNVD 1441	8	76	0.95	0.90	1.21	1.74								
(Reeves III and Van Kessel, 2002)	Dried dairy manure (0.85 mm)	0.4-2.5			96	0.793											
(Reeves III and Van Kessel, 2002)	Dried dairy manure (0.85 mm) – ash free DM basis	0.4-2.5			98	0.882											
WU																	
(Cozzolino et al., 2006a)	Whole maize silage	0.4-2.5	None	10	90	0.86			2.21					2.10	23.7		
(Cozzolino et al., 2006a)	Whole maize silage	0.5-1.1	2D, SNVD	6	90	0.81			2.71					2.00	23.7		
NDF																	
DG																	
(Montes et al., 2009)	Maize stover (1 mm particle size)	9.6-1.69		10	242	0.96		1.50	1.68		0.95						
(Kong et al., 2005)	Ground rice straws (1 mm)	1.1-2.5	D-2441 SNVDT	4	136	0.846		1.93	2.13	71	0.775	2.228	9.31	2.02	20.75	58.07	4.50
(McTiernan et al., 2003)	Decomposing Pinus sylvestris needles (1 mm)				117	0.91		1.11	1.44			1.07	17.10	3.55	18.3	78.9	3.8
(Hodgson et al., 2010)	Miscanthus varieties (1 mm)	0.4-2.5	SNVDT D-2641	7	76	0.96		1.05	1.36		0.93						
(Reeves III and Van Kessel, 2002)	Dried dairy manure (0.85 mm)	0.4-2.5			93	0.853											
(Reeves III and Van Kessel, 2002)	Dried dairy manure (0.85 mm) – ash free DM	0.4-2.5			93	0.735											
DU																	
(Montes et al., 2009)	Maize stover, 8 cm av. size (custom NIR unit)	9.6-1.69			80	0.70	0.56	2.64	3.21								
(Montes et al., 2009)	Maize stover, 6 cm av. size (custom NIR unit)	9.6-1.69			80	0.74	0.66	2.45	2.82								
(Montes et al., 2009)	Maize stover, 4 cm av. size (custom NIR unit)	9.6-1.69			80	0.85	0.69	1.88	2.71								
(Montes et al., 2009)	Maize stover, 0.5 cm av. size (custom NIR unit)	9.6-1.69			80	0.89	0.77	1.62	2.34								
(Montes et al., 2009)	Maize stover, 0.5 cm av. size (custom NIR unit)	9.6-1.69		8	242	0.83	0.79	3.01	3.32								
WU																	
(Cozzolino et al., 2006a)	Whole maize silage	0.4-2.5	MSC	10	90	0.60			6.71					1.20	32.17		
(Cozzolino et al., 2006a)	Whole maize silage	0.5-1.1	2D, SNVDT	6	90	0.84			5.41					1.60	32.17		

Table B-4: Statistics for NIRS calibration equations for the glucose (cellulosic) contents of lignocellulosic feedstocks.

Reference	Feedstock(s)	Wav. Range (10 ³ nm)	Pretreatment	F	Calibration Set					Validation Set					Const. Data		
					n	r ²	r _{cv} ²	SEC	SECV	n	r ²	SEP	RER	RPD	R	M	SD
Glucose																	
DG																	
(Sanderson et al., 1996)	Various crops and residues (1 mm)	1.1-2.5	SNVD		99	0.92		1.05	1.34	20	0.93	1.45		3.35		37.44	4.86
(Sanderson et al., 1996)	Various crops and residues (1 mm)	1.4-2.5	SNVD		97	0.94		0.89	1.13	20	0.94	1.30		3.74		37.44	4.86
(Schimleck et al., 1997)	Eucalyptus globulus wood (milled)	1.1-2.5	2D	2	11	0.96		0.90									
(Schimleck et al., 1997)	Eucalyptus nitens wood (milled)	1.1-2.5	2D	4	16	0.961		0.69		5		1.19					
(Kelley et al., 2004b)	Various samples (incl. pretreated), 1 mm	0.5-2.4		6	23	0.94	0.87	3.60	5.20								
(Ye et al., 2008)	Corn stover fractions (0.42 mm)	1-2.5	EMSC		35		0.97		0.93				14.31			35.31	
(Liu et al., 2010b)	Switchgrass fractions (0.42 mm)	1-2.5	EMSC		36		0.97		0.66				12.75			37.23	
(Liu et al., 2010b)	Switchgrass and corn stover (0.42 mm)	1-2.5	EMSC		71		0.97		0.85	15		0.77	12.58			38.89	
(Hames et al., 2003)	Corn stover (1 mm)	0.4-2.5			47				1.445				14.60		21.1		
(Hames et al., 2003)	Pretreated corn stover (1 mm)	0.4-2.5			96				1.549				18.90		29.3		
(Liu et al., 2010a)	Switchgrass (0.42 mm)	1-2.5	SNV	9	36		0.92		0.675	5		0.78	10.79	3.31	8.42		2.58
(Liu et al., 2010a)	Switchgrass (0.42 mm)	1-2.5	MSC	9	36		0.89		0.674	5		0.78	10.79	3.31	8.42		2.58
(Liu et al., 2010a)	Switchgrass (0.42 mm)	1-2.5	1D (SG)	9	36		0.89		0.773	5		0.86	9.79	3.00	8.42		2.58
(Liu et al., 2010a)	Switchgrass (0.42 mm)	1-2.5	2D (SG)	10	36		0.85		0.795	5		0.91	9.25	2.84	8.42		2.58
(Liu et al., 2010a)	Switchgrass (0.42 mm)	1-2.5	EMSC	9	36		0.94		0.65	5		0.75	11.23	3.44	8.42		2.58
(Liu et al., 2010a)	Switchgrass (0.42 mm)	1-2.5	1D (SG) + SNV	9	36		0.91		0.668	5		0.77	10.94	3.35	8.42		2.58
(Liu et al., 2010a)	Switchgrass (0.42 mm)	1-2.5	SNV + 2D (SG)	10	36		0.88		0.724	5		0.93	9.05	2.77	8.42		2.58
(Liu et al., 2010a)	Switchgrass (0.42 mm)	1-2.5	2D (SG) + SNV	10	36		0.90		0.718	5		0.85	9.91	3.04	8.42		2.58
(Liu et al., 2010a)	Switchgrass (0.42 mm)	1-2.5	None	10	36		0.69		1.02	5		1.36	6.19	1.90	8.42		2.58
(Liu et al., 2010a)	Corn-stover (0.42 mm)	1-2.5	SNV	9	35		0.85		1.65	5		1.36	9.79	3.10	13.32		4.21
(Liu et al., 2010a)	Corn-stover (0.42 mm)	1-2.5	MSC	9	35		0.81		1.91	5		1.37	9.72	3.07	13.32		4.21
(Liu et al., 2010a)	Corn-stover (0.42 mm)	1-2.5	1D (SG)	9	35		0.92		1.15	5		1.06	12.57	3.97	13.32		4.21
(Liu et al., 2010a)	Corn-stover (0.42 mm)	1-2.5	2D (SG)	10	35		0.92		1.12	5		1.17	11.38	3.60	13.32		4.21
(Liu et al., 2010a)	Corn-stover (0.42 mm)	1-2.5	EMSC	9	35		0.94		0.97	5		0.89	14.97	4.73	13.32		4.21
(Liu et al., 2010a)	Corn-stover (0.42 mm)	1-2.5	1D (SG) + SNV	9	35		0.92		1.08	5		1.02	13.06	4.13	13.32		4.21
(Liu et al., 2010a)	Corn-stover (0.42 mm)	1-2.5	SNV + 2D (SG)	10	35		0.85		1.19	5		1.02	13.06	4.13	13.32		4.21
(Liu et al., 2010a)	Corn-stover (0.42 mm)	1-2.5	2D (SG) + SNV	9	35		0.92		1.16	5		0.96	13.88	4.39	13.32		4.21
(Liu et al., 2010a)	Corn-stover (0.42 mm)	1-2.5	None	11	35		0.79		1.87	5		1.74	7.66	2.42	13.32		4.21
DU																	
(Alves et al., 2006)	Pine strips	1.1-2.5	2D	4	28	0.82		1.09	1.96	12	0.57	1.88		1.39			
(Kelley et al., 2004a)	Pine strips	0.5-2.4	None	4	45	0.90		2.4		27	0.78	2.7					
(Kelley et al., 2004a)	Pine strips	0.65-1.15	None	5	45	0.88		1.5		27	0.84	2.3					
WU																	
(Axrup et al., 2000)	Wood chips	0.8-1.1	None	6	83			11.6		35		12.4	0.93		11.6		
(Axrup et al., 2000)	Wood chips	0.8-1.1	MSC	5	83			5.1		35		6.8	1.71		11.6		

Table B-5: Statistics for NIRS calibration equations for the xylose contents of lignocellulosic feedstocks.

Reference	Feedstock(s)	Wav. Range (10 ³ nm)	Pretreatment	F	Calibration Set					Validation Set					Const. Data		
					n	r ²	r _{CP} ²	SEC	SECV	n	r ²	SEP	RER	RPD	R	M	SD
Xylose																	
DG																	
(Sanderson et al., 1996)	Various crops and residues (1 mm)	1.1-2.5	SNVD		95	0.96		0.65	0.77	20	0.92	1.17		3.71		18.88	4.34
(Sanderson et al., 1996)	Various crops and residues (1 mm)	1.4-2.5	SNVD		95	0.98		0.49	0.57	20	0.94	1.06		4.09		18.88	4.34
(Schimleck et al., 1997)	Eucalyptus globulus wood (milled)	1.1-2.5	2D	1	11	0.95		0.52									
(Schimleck et al., 1997)	Eucalyptus nitens wood (milled)	1.1-2.5	2D	5	16	0.988		0.19		5		0.95					
(Kelley et al., 2004b)	Various samples (incl. pretreated), 1 mm	0.5-2.4		6	23	0.87	0.71	3.50	5.80								
(Ye et al., 2008)	Corn stover fractions (0.42 mm)	1-2.5	EMSC		35		0.93		1.04				12.97			17.75	
(Liu et al., 2010b)	Switchgrass fractions (0.42 mm)	1-2.5	EMSC		36		0.96		0.58				10.41			20.28	
(Liu et al., 2010b)	Switchgrass and corn stover (0.42 mm)	1-2.5	EMSC		71		0.94		0.88	15		0.48	12.87			19.44	
(Hames et al., 2003)	Corn stover (1 mm)	0.4-2.5			47				0.949				11.40			10.8	
(Hames et al., 2003)	Pretreated corn stover (1 mm)	0.4-2.5			96				1.458				15.20			22.2	
(Liu et al., 2010a)	Switchgrass (0.42 mm)	1-2.5	SNV	9	36		0.89		0.782	5		0.70	8.61	2.76	6.03		1.93
(Liu et al., 2010a)	Switchgrass (0.42 mm)	1-2.5	MSC	9	36		0.89		0.581	5		0.61	9.89	3.16	6.03		1.93
(Liu et al., 2010a)	Switchgrass (0.42 mm)	1-2.5	1D (SG)	9	36		0.89		0.739	5		0.79	7.63	2.44	6.03		1.93
(Liu et al., 2010a)	Switchgrass (0.42 mm)	1-2.5	2D (SG)	10	36		0.77		1.018	5		0.73	8.26	2.64	6.03		1.93
(Liu et al., 2010a)	Switchgrass (0.42 mm)	1-2.5	EMSC	9	36		0.92		0.57	5		0.53	11.38	3.64	6.03		1.93
(Liu et al., 2010a)	Switchgrass (0.42 mm)	1-2.5	1D (SG) + SNV	9	36		0.87		0.805	5		0.92	6.55	2.10	6.03		1.93
(Liu et al., 2010a)	Switchgrass (0.42 mm)	1-2.5	SNV + 2D (SG)	10	36		0.87		0.589	5		0.69	8.74	2.80	6.03		1.93
(Liu et al., 2010a)	Switchgrass (0.42 mm)	1-2.5	2D (SG) + SNV	10	36		0.87		0.785	5		0.91	6.63	2.12	6.03		1.93
(Liu et al., 2010a)	Switchgrass (0.42 mm)	1-2.5	None	10	36		0.66		0.86	5		0.97	6.22	1.99	6.03		1.93
(Liu et al., 2010a)	Corn-stover (0.42 mm)	1-2.5	SNV	9	35		0.66		1.68	5		1.11	12.14	2.45	13.48		2.72
(Liu et al., 2010a)	Corn-stover (0.42 mm)	1-2.5	MSC	9	35		0.67		1.65	5		1.03	13.09	2.64	13.48		2.72
(Liu et al., 2010a)	Corn-stover (0.42 mm)	1-2.5	1D (SG)	9	35		0.79		1.21	5		1.21	11.14	2.25	13.48		2.72
(Liu et al., 2010a)	Corn-stover (0.42 mm)	1-2.5	2D (SG)	10	35		0.72		1.03	5		0.95	14.19	2.86	13.48		2.72
(Liu et al., 2010a)	Corn-stover (0.42 mm)	1-2.5	EMSC	9	35		0.86		1.03	5		0.88	15.32	3.09	13.48		2.72
(Liu et al., 2010a)	Corn-stover (0.42 mm)	1-2.5	1D (SG) + SNV	9	35		0.79		1.22	5		0.98	13.76	2.78	13.48		2.72
(Liu et al., 2010a)	Corn-stover (0.42 mm)	1-2.5	SNV + 2D (SG)	10	35		0.77		1.35	5		1.15	11.72	2.37	13.48		2.72
(Liu et al., 2010a)	Corn-stover (0.42 mm)	1-2.5	2D (SG) + SNV	9	35		0.83		1.07	5		0.88	15.32	3.09	13.48		2.72
(Liu et al., 2010a)	Corn-stover (0.42 mm)	1-2.5	None	11	35		0.52		1.92	5		1.63	8.27	1.67	13.48		2.72
DU																	
(Jones et al., 2006)	Pine strips	1.1-2.5	2D	6	28	0.98		0.11	0.44	12	0.61	0.46		1.52			
(Kelley et al., 2004a)	Pine strips	0.5-2.4	None	4	45	0.80		0.6		27	0.56	0.6					
(Kelley et al., 2004a)	Pine strips	0.65-1.15	None	5	45	0.80		0.8		27	0.54	0.6					
WU																	
(Axrup et al., 2000)	Wood chips	0.8-1.1	None	6	83			7.9		35		9.1	1.33		12.1		
(Axrup et al., 2000)	Wood chips	0.8-1.1	MSC	5	83			4.1		35		4.6	2.63		12.1		

Table B-6: Statistics for NIRS calibration equations for the arabinose and galactose contents of lignocellulosic feedstocks.

Reference	Feedstock(s)	Wav. Range (10 ³ nm)	Pretreatment	F	Calibration Set					Validation Set					Const. Data			
					n	r ²	r _{CV} ²	SEC	SECV	n	r ²	SEP	RER	RPD	R	M	SD	
Arabinose																		
DG																		
(Sanderson et al., 1996)	Various crops and residues (1 mm)	1.1-2.5	SNVD		99	0.98		0.11	0.13	20	0.99	0.12			8.75		2.05	1.05
(Sanderson et al., 1996)	Various crops and residues (1 mm)	1.4-2.5	SNVD		98	0.98		0.12	0.14	20	0.99	0.12			8.75		2.05	1.05
(Kelley et al., 2004b)	Various samples (incl. pretreated), 1 mm	0.5-2.4		6	23	0.87	0.57	<i>0.30</i>	<i>0.50</i>									
(Ye et al., 2008)	Corn stover fractions (0.42 mm)	1-2.5	EMSC		35		0.96		0.28								3.43	
(Liu et al., 2010b)	Switchgrass fractions (0.42 mm)	1-2.5	EMSC		36		0.95		0.20								3.56	
(Liu et al., 2010b)	Switchgrass and corn stover (0.42 mm)	1-2.5	EMSC		71		0.91		0.34	15		0.28	11.25				3.45	
(Hames et al., 2003)	Corn stover (1 mm)	0.4-2.5			47				0.199								3.9	
DU																		
(Jones et al., 2006)	Pine strips	1.1-2.5	2D	3	28	0.75		0.09	0.12	12	0.68	0.11			1.73			
WU																		
(Axrup et al., 2000)	Wood chips	0.8-1.1	None	6	83			2.3		35		2.5	0.78				1.95	
(Axrup et al., 2000)	Wood chips	0.8-1.1	MSC	5	83			1.4		35		1.6	1.22				1.95	
Galactose																		
DG																		
(Sanderson et al., 1996)	Various crops and residues (1 mm)	1.1-2.5	SNVD		95	0.91		0.08	0.10	20	0.87	0.12			2.75		1.06	0.33
(Sanderson et al., 1996)	Various crops and residues (1 mm)	1.4-2.5	SNVD		95	0.90		0.09	0.11	20	0.86	0.12			2.75		1.06	0.33
(Kelley et al., 2004b)	Various samples (incl. pretreated), 1 mm	0.5-2.4		6	23	0.84	0.52	<i>0.40</i>	<i>0.70</i>									
(Ye et al., 2008)	Corn stover fractions (0.42 mm)	1-2.5	EMSC		35		0.92		0.19								2.04	
(Liu et al., 2010b)	Switchgrass fractions (0.42 mm)	1-2.5	EMSC		36		0.97		0.13								2.39	
(Liu et al., 2010b)	Switchgrass and corn stover (0.42 mm)	1-2.5	EMSC		71		0.88		0.26	15		0.21	7.24				1.90	
(Hames et al., 2003)	Corn stover (1 mm)	0.4-2.5			47				0.327								2.7	
DU																		
(Jones et al., 2006)	Pine strips	1.1-2.5	2D	1	28	0.25		1.11	1.23	12	0.11	0.97			0.83			
(Kelley et al., 2004a)	Pine strips	0.5-2.4	None	4	45	0.82		1.0		27	0.80	1.0						
(Kelley et al., 2004a)	Pine strips	0.65-1.15	None	5	45	0.81		1.0		27	0.83	0.8						
WU																		
(Axrup et al., 2000)	Wood chips	0.8-1.1	None	6	83			2.3		35		2.9	1.53				1.9	
(Axrup et al., 2000)	Wood chips	0.8-1.1	MSC	5	83			1.3		35		1.8	1.06				1.9	

Table B-7: Statistics for NIRS calibration equations for the mannose, rhamnose, and uronic acids contents of lignocellulosic feedstocks.

Reference	Feedstock(s)	Wav. Range (10 ³ nm)	Pretreatment	F	Calibration Set					Validation Set					Const. Data		
					n	r ²	r _{CP} ²	SEC	SECV	n	r ²	SEP	RER	RPD	R	M	SD
Mannose																	
DG																	
(Sanderson et al., 1996)	Various crops and residues (1 mm)	1.1-2.5	SNVD		96	0.97		0.12	0.16	20	0.88	0.32		2.78		0.99	0.89
(Sanderson et al., 1996)	Various crops and residues (1 mm)	1.4-2.5	SNVD		97	0.97		0.15	0.18	20	0.87	0.34		2.62		0.99	0.89
(Kelley et al., 2004b)	Various samples (incl. pretreated), 1 mm	0.5-2.4		6	23	0.80	0.44	2.60	4.50								
(Ye et al., 2008)	Corn stover fractions (0.42 mm)	1-2.5	EMSC		35		0.91		0.23				9.74			1.16	
(Liu et al., 2010b)	Switchgrass fractions (0.42 mm)	1-2.5	EMSC		36		0.87		0.12				7.39			0.85	
(Liu et al., 2010b)	Switchgrass and corn stover (0.42 mm)	1-2.5	EMSC		71		0.89		0.21	15		0.07	8.21			0.90	
(Hames et al., 2003)	Corn stover (1 mm)	0.4-2.5			47				0.136				11.00		1.5		
DU																	
(Jones et al., 2006)	Pine strips	1.1-2.5	2D	5	28	0.96		0.23	0.64	12	0.66	0.71		1.71			
(Kelley et al., 2004a)	Pine strips	0.5-2.4	None	4	45	0.86		0.8		27	0.58	1.3					
(Kelley et al., 2004a)	Pine strips	0.65-1.15	None	5	45	0.86		0.8		27	0.69	1.0					
WU																	
(Axrup et al., 2000)	Wood chips	0.8-1.1	None	6	83			10.3		35		11.9	1.04		12.4		
(Axrup et al., 2000)	Wood chips	0.8-1.1	MSC	5	83			4.7		35		5.1	2.43				
Rhamnose																	
DG																	
(Kelley et al., 2004b)	Various samples (incl. pretreated), 1 mm	0.5-2.4		6	23	0.72	0.36	0.10	0.20								
Uronic Acids																	
DG																	
(Sanderson et al., 1996)	Various crops and residues (1 mm)	1.1-2.5	SNVD		71	0.85		0.49	0.57	16	0.76	0.73		2.07		3.21	1.51
(Sanderson et al., 1996)	Various crops and residues (1 mm)	1.4-2.5	SNVD		72	0.80		0.59	0.68	16	0.75	0.76		1.99		3.21	1.51
(Hames et al., 2003)	Corn stover (1 mm)	0.4-2.5			47				0.189				11.70		2.2		
Acetyl Content																	
DG																	
(Schimleck et al., 1997)	Eucalyptus globulus wood (milled)	1.1-2.5	2D	4	11	0.99		0.04									
(Hames et al., 2003)	Corn stover (1 mm)	0.4-2.5			47				0.2				22.50		4.5		

Table B-8: Statistics for NIRS calibration equations for the Klason lignin and acid soluble lignin contents of lignocellulosic feedstocks.

Reference	Feedstock(s)	Wav. Range (10 ³ nm)	Pretreatment	F	Calibration Set					Validation Set					Const. Data			
					n	r ²	r _{CP} ²	SEC	SECV	n	r ²	SEP	RER	RPD	R	M	SD	
Klason Lignin																		
DG																		
(Krongtaew et al., 2010b)	Pretreated wheat straw (0.08 mm particle size)*a	1.449-1.815	2D (SG)	3	80	0.95		0.9		40	0.94	0.9			12.8			
(Krongtaew et al., 2010b)	Pretreated oat straw (0.08 mm particle size)*a	1.449-1.815	2D (SG)	1	53	0.99		0.5		27	0.96	0.8			10.2			
(Kelley et al., 2004b)	Various biomass samples (incl. pretreated), 1mm	0.5-2.4		6	23	0.88	0.71	4.00	6.10									
(Kong et al., 2005)*b	Ground rice straws (1 mm)	1.1-2.5	D-1,4,4,1 + SNVDT	8	136	0.876		0.40	0.49	71	0.847	0.616	12.09	2.55	7.45	6.95	1.57	
(McTiernan et al., 2003)*b	Decomposing Pinus sylvestris needles (1 mm)			?	112 (C+V)	0.97		1.01	1.63			0.96			25.1	29.0	5.5	
(Hodgson et al., 2010)*b	Miscanthus varieties (1 mm)	0.4-2.5	SNVDT + D-2641	5	76	0.82	0.75	0.58	0.69									
(Liu and Chen, 2007)*b	Rice straw (0.425-0.25 mm)	1.33-2.44	D-1,5,10,1	11	34	0.89			0.81	9	0.8601	2.1	2.70	0.76	5.6	10.2	1.6	
(Reeves III and Van Kessel, 2002)*b	Dried dairy manure (0.85 mm)	0.4-2.5			95	0.864												
(Reeves III and Van Kessel, 2002)*b	Dried dairy manure (0.85 mm) – ash free DM	0.4-2.5			94	0.821												
(Vavrova et al., 2008)	Plant litter (1 mm)	0.78-2.5	D-2,4,4,1	6	56	0.99	0.98	0.85	1.29					19.60	6.47	25.29	34.75	8.35
(Hodge and Woodbridge, 2010)	Pine (various sp.) woodmeal (1 mm)	1.1-2.5	MSC 2D (SG)	11	345	0.97		0.34	0.45	172	0.95	0.44	25.00	4.63	11	26.4		
(Hodge and Woodbridge, 2010)	P. taeda woodmeal (1 mm)	1.1-2.5	MSC 2D (SG)	8	61	0.97		0.23		31	0.91	0.45		3.23		29		
(Hodge and Woodbridge, 2010)	P. tecumanii woodmeal (1 mm)	1.1-2.5	MSC 2D (SG)	4	104	0.92		0.38		52	0.92	0.47		3.52		25.6		
DU																		
(Poke, 2006)	Eucalyptus strips			4	40	0.78		1.02		9	0.78							
(Jones et al., 2006)	Pine strips	1.1-2.5	2D	4	28	0.85		0.47	0.91	12	0.51	1.26		1.33				
(Kelley et al., 2004a)	Pine strips	0.5-2.4	None	4	45	0.81		1.1		27	0.76	1.0						
(Kelley et al., 2004a)	Pine strips	0.65-1.15	None	5	45	0.71		1.5		27	0.67	1.4						
(Nkansah et al., 2010)	Yellow poplar wood blocks	0.8-2.5	None	8	37	0.83		6.06		17	0.64	9.22						
(Nkansah et al., 2010)	Yellow poplar wood blocks	1.3-1.8	None	9	37	0.86		5.50		17	0.39	1.10						
(Nkansah et al., 2010)	Yellow poplar wood blocks	0.8-2.5	1D (SG)	5	37	0.89		4.95		17	0.79	7.17						
(Nkansah et al., 2010)	Yellow poplar wood blocks	1.3-1.8	1D (SG)	3	37	0.7		8.28		17	0.75	7.83						
WU																		
(Axrup et al., 2000)	Wood chips	0.8-1.1	None	6	83			4.3		35		5.3	2.17		11.5			
(Axrup et al., 2000)	Wood chips	0.8-1.1	MSC	5	83			1.8		35		1.8	6.40		11.5			
(Vergnoux et al., 2009)	Sewage sludge compost	1-2.222	MSC	10	9	0.99		0.71		4	0.80	3.16	3.01		9.5			
Acid Soluble Lignin																		
(Poke, 2006)	Eucalyptus strips			6	40	0.72		0.41		9	0.12							
(Jones et al., 2006)	Pine strips	1.1-2.5	2D	2	28	0.66		0.02	0.02	12	0.36	0.02		1.50				

*a – this particle size is so small that the fraction could be considered more equivalent to a DF (see Section 11.1) than a DG; *b – ADL (acid detergent lignin) was measured here.

Table B-9: Statistics for NIRS calibration equations for the total lignin content of lignocellulosic feedstocks.

Reference	Feedstock(s)	Wav. Range (10 ³ nm)	Pretreatment	F	Calibration Set					Validation Set					Const. Data												
					n	r ²	r _{cv} ²	SEC	SECV	n	r ²	SEP	RER	RPD	R	M	SD										
Total Lignin (KL + ASL)																											
DG																											
(Sanderson et al., 1996)	Various crops and residues (1 mm)	1.1-2.5	SNVD		97	0.96		0.73	0.81	20	0.98	0.70		7.03		22.22	4.92										
(Sanderson et al., 1996)	Various crops and residues (1 mm)	1.4-2.5	SNVD		98	0.97		0.65	0.78	20	0.99	0.53		9.28		22.22	4.92										
(Schimleck et al., 1997)	Eucalyptus globulus wood (milled)	1.1-2.5	2D	1	11	0.88		0.85																			
(Ye et al., 2008)	Corn stover fractions (0.42 mm)	1-2.5	EMSC		35		0.94		1.04				13.71			20.67											
(Liu et al., 2010b)	Switchgrass fractions (0.42 mm)	1-2.5	EMSC		36		0.93		0.49				9.77			22.28											
(Liu et al., 2010b)	Switchgrass and corn stover (0.42 mm)	1-2.5	EMSC		71		0.90		1.32	15		0.82	11.23			22.49											
(Hames et al., 2003)	Corn stover (1 mm)	0.4-2.5			47				1.120				14.80			16.6											
(Hames et al., 2003)	Pretreated corn stover (1 mm)	0.4-2.5			96				1.506				7.60			11.4											
(Liu et al., 2010a)	Switchgrass (0.42 mm)	1-2.5	SNV	9	36		0.74		0.67	5		0.60	7.97	2.17	4.78		1.30										
(Liu et al., 2010a)	Switchgrass (0.42 mm)	1-2.5	MSC	9	36		0.72		0.598	5		0.59	8.10	2.20	4.78		1.30										
(Liu et al., 2010a)	Switchgrass (0.42 mm)	1-2.5	1D (SG)	9	36		0.76		0.65	5		0.76	6.29	1.71	4.78		1.30										
(Liu et al., 2010a)	Switchgrass (0.42 mm)	1-2.5	2D (SG)	10	36		0.71		0.746	5		0.79	6.05	1.65	4.78		1.30										
(Liu et al., 2010a)	Switchgrass (0.42 mm)	1-2.5	EMSC	9	36		0.82		0.48	5		0.50	9.56	2.60	4.78		1.30										
(Liu et al., 2010a)	Switchgrass (0.42 mm)	1-2.5	1D (SG) + SNV	9	36		0.86		0.517	5		0.53	9.02	2.45	4.78		1.30										
(Liu et al., 2010a)	Switchgrass (0.42 mm)	1-2.5	SNV + 2D (SG)	10	36		0.68		0.746	5		0.80	5.98	1.63	4.78		1.30										
(Liu et al., 2010a)	Switchgrass (0.42 mm)	1-2.5	2D (SG) + SNV	10	36		0.71		0.698	5		0.81	5.90	1.60	4.78		1.30										
(Liu et al., 2010a)	Switchgrass (0.42 mm)	1-2.5	None	10	36		0.44		0.75	5		0.88	5.43	1.48	4.78		1.30										
(Liu et al., 2010a)	Corn-stover (0.42 mm)	1-2.5	SNV	9	35		0.61		2.52	5		1.25	11.41	3.13	14.26		3.91										
(Liu et al., 2010a)	Corn-stover (0.42 mm)	1-2.5	MSC	9	35		0.59		2.72	5		1.19	11.98	3.29	14.26		3.91										
(Liu et al., 2010a)	Corn-stover (0.42 mm)	1-2.5	1D (SG)	9	35		0.85		1.49	5		1.05	13.58	3.72	14.26		3.91										
(Liu et al., 2010a)	Corn-stover (0.42 mm)	1-2.5	2D (SG)	10	35		0.83		1.27	5		0.87	16.39	4.49	14.26		3.91										
(Liu et al., 2010a)	Corn-stover (0.42 mm)	1-2.5	EMSC	9	35		0.85		1.12	5		0.79	18.05	4.95	14.26		3.91										
(Liu et al., 2010a)	Corn-stover (0.42 mm)	1-2.5	1D (SG) + SNV	9	35		0.83		1.52	5		0.97	14.70	4.03	14.26		3.91										
(Liu et al., 2010a)	Corn-stover (0.42 mm)	1-2.5	SNV + 2D (SG)	10	35		0.85		1.49	5		1.02	13.98	3.83	14.26		3.91										
(Liu et al., 2010a)	Corn-stover (0.42 mm)	1-2.5	2D (SG) + SNV	9	35		0.90		1.20	5		0.81	17.60	4.83	14.26		3.91										
(Liu et al., 2010a)	Corn-stover (0.42 mm)	1-2.5	None	11	35		0.49		2.83	5		1.92	7.43	2.04	14.26		3.91										
(Ono et al., 2003)	Leaf litter (74 μm)	0.4-2.5	2D (MLR used)	4	85	0.94		4.0		44	0.87	5.0	9.62	2.10	48.1	30.2	10.1										
DU																											
(Jones et al., 2006)	Pine strips	1.1-2.5	2D	4	28	0.85		0.48	0.92	12	0.51	1.21		1.43													
(Nkansah et al., 2010)	Yellow poplar wood blocks	0.8-2.5	None	8	37	0.83		6.00		17	0.64	9.05															
(Nkansah et al., 2010)	Yellow poplar wood blocks	1.3-1.8	None	9	37	0.85		5.50		17	0.38	10.97															
(Nkansah et al., 2010)	Yellow poplar wood blocks	0.8-2.5	1D (SG)	4	37	0.81		6.41		17	0.66	8.95															
(Nkansah et al., 2010)	Yellow poplar wood blocks	1.3-1.8	1D (SG)	3	37	0.68		8.17		17	0.77	7.24															

Table B-10: Statistics for NIRS calibration equations for the extractives contents of lignocellulosic feedstocks.

Reference	Feedstock(s)	Wav. Range (10 ³ nm)	Pretreatment	F	Calibration Set					Validation Set					Const. Data		
					n	r ²	r _{CP} ²	SEC	SECV	n	r ²	SEP	RER	RPD	R	M	SD
Extractives																	
DG																	
(Sanderson et al., 1996)	Various crops and residues (1 mm)	1.1-2.5	SNVDT		96	0.95		1.10	1.31	20	0.96	1.10		4.97		7.88	5.47
(Sanderson et al., 1996)	Various crops and residues (1 mm)	1.4-2.5	SNVDT		95	0.94		1.02	1.31	20	0.96	1.13		4.84		7.88	5.47
(Schimleck et al., 1997)*b	Eucalyptus globulus wood (milled)	1.1-2.5	2D	1	11	0.88		0.44									
(Vavrova et al., 2008)*c	Plant litter (1 mm)	0.78-2.5	D-2,4,4,1	4	56	0.96	0.94	0.86	0.97				14.70	4.25	14.29	6.82	4.12
(Vavrova et al., 2008)*d	Plant litter (1 mm)	0.78-2.5	D-2,4,4,1	5	56	0.94	0.88	0.48	0.78				9.80	2.91	7.68	2.37	2.04
(Vavrova et al., 2008)	Plant litter (1 mm)	0.78-2.5	D-2,4,4,1	5	56	0.94	0.86	0.64	0.99				10.80	2.70	10.66	4.09	2.67
(Vavrova et al., 2008)*b	Plant litter (1 mm)	0.78-2.5	D-2,4,4,1	6	56	0.98	0.96	0.45	0.69				14.40	4.86	9.95	5.02	3.35
(Vavrova et al., 2008)*e	Plant litter (1 mm)	0.78-2.5	D-2,4,4,1	4	56	0.96	0.94	2.11	2.58				12.60	4.03	32.41	18.29	10.41
(Labbé et al., 2008)*f	Switchgrass (5 mm)	1-2.5	MSC	4	72	0.96	0.91	0.46	0.68								
(Labbé et al., 2008)*g	Switchgrass (5 mm)	1-2.5	MSC	3	72	0.91	0.82	0.22	0.31								
(Ono et al., 2003)*h	Plant litter (74 µm)	0.4-2.5	2D (MLR used)	4	88	0.91		2.9		44	0.82	3.4	9.88	2.10	33.6	11.3	7.3
DU																	
(Poke, 2006)	Eucalyptus strips			6	40	0.84		1.37		9	0.87						
(Kelley et al., 2004a)	Pine strips	0.5-2.4	None	4	45	0.93		2.3		27	0.85	2.3					
(Kelley et al., 2004a)	Pine strips	0.65-1.15	None	5	45	0.93		2.4		27	0.88	2.2					
(Nkansah et al., 2010)	Yellow poplar wood blocks	0.8-2.5	None	8 *i	37	0.86		2.68		17	0.84	2.99					
(Nkansah et al., 2010)	Yellow poplar wood blocks	1.3-1.8	None	8 *i	37	0.83		2.92		17	0.78	3.42					
(Nkansah et al., 2010)	Yellow poplar wood blocks	0.8-2.5	1D (SG)	5	37	0.83		2.89		17	0.78	3.42					
(Nkansah et al., 2010)	Yellow poplar wood blocks	1.3-1.8	1D (SG)	6	37	0.81		3.11		17	0.83	2.95					
WU																	
(Axrup et al., 2000)*a	Wood chips	0.8-1.1	None	6	83			2.0		35		2.7	2.22		6		
(Axrup et al., 2000)*a	Wood chips	0.8-1.1	MSC	5	83			1.0		35		1.2	5.00		6		

*a – these are acetone extractives; *b – these are hot water extractives; *c – nonpolar (dichloromethane) extractives; *d - acetone extractives; *e – total extractives (sum of ethanol extractives + *b + *c + *d); *f – starch (calculated as the glucose liberated on enzymatic digestion); *g – soluble sugars obtained under extraction with 85% ethanol; *h - benzene:ethanol (2:1) solution; *i – principal components (PCR was used for these models).

Table B-11: Statistics for NIRS calibration equations for the nitrogen and carbon contents of lignocellulosic feedstocks.

Reference	Feedstock(s)	Wav. Range (10 ³ nm)	Pretreatment	F	Calibration Set					Validation Set					Const. Data			
					n	r ²	r _{CP} ²	SEC	SECV	n	r ²	SEP	RER	RPD	R	M	SD	
Nitrogen																		
DG																		
(Montes et al., 2009)	Maize stover (1 mm particle size)	9.6-1.69		11	242	0.94	0.92	0.05	0.06									
(Huang et al., 2007)	Various animal manures	1-2.5	1D, SNV	9	90	0.99		0.24		30	0.97	0.36	34.44	6.95	12.4			
(Sanderson et al., 1996)	Various crops and residues (1 mm)	1.1-2.5	SNVDT		80	0.95		0.06	0.07	18	0.77	0.17		1.76		0.63	0.30	
(Sanderson et al., 1996)	Various crops and residues (1 mm)	1.4-2.5	SNVDT		82	0.94		0.06	0.08	18	0.90	0.09		3.33		0.63	0.30	
(McTiernan et al., 2003)	Decomposing Pinus sylvestris needles (1 mm)				121	0.92		0.08	0.11			0.08	10.00	3.13	0.80	1.07	0.25	
(Beining et al., 2000)	Freeze dried peat tablets	1.25-2.5	1D			0.979									2.09			
(Huang et al., 2009)	Rice and wheat straws (1 mm)	1.1-2.5			166	0.93		0.07	0.08	56	0.87	0.10	12.30	2.76	1.23	0.78	0.27	
(Reeves III and Van Kessel, 2002)	Dried dairy manure (0.85 mm)	0.4-2.5			96	0.728												
(Reeves III and Van Kessel, 2002)	Dried dairy manure (0.85 mm) – ash free DM basis	0.4-2.5			95	0.857												
(Vavrova et al., 2008)	Plant litter (1 mm)	0.78-2.5	SNVDT	9	76	0.92	0.88	0.06	0.08					12.80	2.91	1.02	0.56	0.20
DU																		
(Montes et al., 2009)	Maize stover, 8 cm av. size (custom NIR unit)	9.6-1.69			80	0.52	0.41	0.17	0.18									
(Montes et al., 2009)	Maize stover, 6 cm av. size (custom NIR unit)	9.6-1.69			80	0.57	0.44	0.16	0.18									
(Montes et al., 2009)	Maize stover, 4 cm av. size (custom NIR unit)	9.6-1.69			80	0.66	0.50	0.14	0.17									
(Montes et al., 2009)	Maize stover, 0.5 cm av. size (custom NIR unit)	9.6-1.69			80	0.69	0.51	0.13	0.17									
(Montes et al., 2009)	Maize stover, 0.5 cm av. size (custom NIR unit)	9.6-1.69		9	242	0.61	0.44	0.14	0.16									
WU																		
(Huang et al., 2007)	Various animal manures	1-2.5	1D, SNV	8	90	0.99		0.16		30	0.97	0.396	31.31	6.11	12.4			
(Vergnoux et al., 2009)	Sewage sludge compost	1-2.222	1D	5	24	0.94		0.099		7	0.98	0.109	19.91		2.17			
Carbon																		
DG																		
(Huang et al., 2007)*a	Various animal manures (1 mm)	1-2.5	1D	9	90	0.81		3.78		30	0.85	3.30	10.49	2.56	34.63			
(Sanderson et al., 1996)	Various crops and residues (1 mm)	1.1-2.5	SNVDT		75	0.91		0.37	0.43	17	0.73	0.91		1.84		47.89	1.67	
(Sanderson et al., 1996)	Various crops and residues (1 mm)	1.4-2.5	SNVDT		76	0.90		0.41	0.45	17	0.75	0.81		2.06		47.89	1.67	
(Beining et al., 2000)	Freeze dried peat tables (FTIR)	1.25-2.5	1D			0.975									8.29			
(Huang et al., 2009)	Rice and wheat straws (1 mm)	1.1-2.5			166	0.98		0.30	0.34	56	0.97	0.37	22.20	5.64	8.21	39.45	2.11	
(Reeves III and Van Kessel, 2002)	Dried dairy manure (0.85 mm)	0.4-2.5			95	0.832												
(Reeves III and Van Kessel, 2002)	Dried dairy manure (0.85 mm) – ash free DM	0.4-2.5			94	0.796												
(Vavrova et al., 2008)	Plant litter (1 mm)	0.78-2.5	SNVDT	5	78	0.96	0.95	0.79	0.91					14.20	4.43	12.9	50.53	4.03
WU																		
(Huang et al., 2007)*a	Various animal manures	1-1.432, 1.45-1.925 1.955-2.5	1D	5	90	0.96		1.65		30	0.91	2.63			34.63			
(Vergnoux et al., 2009)*a	Sewage sludge compost	1-2.222	MSC	9	24	0.99		0.62		7	0.99	0.524	30.53		16			
(Fagan et al., 2011)*b	Miscanthus and willows (< 3 mm)	1.1-2.5	2D (SG)	4	44		0.88		0.57					10.40	4.60	6.7		

*a – TOC measured; *b – since all particles were less than 3mm diameter a more appropriate classification here may be WG (i.e. wet and ground)

Table B-12: Statistics for NIRS calibration equations for the moisture, hydrogen, oxygen, and crude protein contents of lignocellulosic feedstocks.

Reference	Feedstock(s)	Wav. Range (10 ³ nm)	Pretreatment	F	Calibration Set					Validation Set					Const. Data				
					n	r ²	r _{CP} ²	SEC	SECV	n	r ²	SEP	RER	RPD	R	M	SD		
Moisture																			
DU																			
(Liu and Chen, 2007)	Rice straw (0.425-0.25mm, air dried)	1.3-2.44	D-1,5,10,1	12	34	0.91			0.89	9	0.8871	1.01	5.54	1.29	5.6	6.9	1.3		
WU																			
(Axrup et al., 2000)	Wood chips	0.8-1.1	None	6	262			1.0		110		1.1	31.27				34.4		
(Axrup et al., 2000)	Wood chips	0.8-1.1	MSC	3	262			0.6		110		0.8	43.00				34.4		
(Axrup et al., 2000)	Bark	0.8-1.1	None	8	359			6.1		180		7.2	5.26				37.9		
(Axrup et al., 2000)	Bark	0.8-1.1	MSC	5	359			1.2		180		1.9	19.95				37.9		
(Cozzolino et al., 2006a)	Whole maize silage	0.4-2.5	2D, SNVDT	6	90	0.85			2.74					2.40			30		
(Cozzolino et al., 2006a)	Whole maize silage	0.5-1.1	2D, SNVDT	4	90	0.90			2.55					2.50			30		
(Malley et al., 2007)	Peat	0.6-1.69	None	10	151	0.83				76		1.80	18.00	2.38			32.4		
(Fagan et al., 2011)*a	Miscanthus and willows (< 3 mm)	1.1-2.5	MSC, 1D (SG)		164		0.99		0.90					39.30	13.50		35.3		
Hydrogen																			
DG																			
(Sanderson et al., 1996)	Various crops and residues (1 mm)	1.1-2.5	SNVDT		75	0.19		0.21	0.22	17	0.12	0.22		1.00			5.75	0.22	
(Sanderson et al., 1996)	Various crops and residues (1 mm)	1.4-2.5	SNVDT		76	0.34		0.18	0.20	17	0.30	0.20		1.10			5.75	0.22	
(Beining et al., 2000)	Freeze dried peat tables (FTIR)	1.25-2.5	2D														1.03		
(Huang et al., 2009)	Rice and wheat straws (1 mm)	1.1-2.5			166	0.85		0.13	0.16	56	0.77	0.17	7.24	2.07	1.23		5.41	0.35	
Oxygen																			
DG																			
(Sanderson et al., 1996)	Various crops and residues (1 mm)	1.1-2.5	SNVDT		45	0.59		0.87	0.99	12	0.65	1.23		1.69			40.55	2.08	
(Sanderson et al., 1996)	Various crops and residues (1 mm)	1.4-2.5	SNVDT		44	0.55		0.94	1.02	12	0.51	1.43		1.45			40.55	2.08	
Crude Protein																			
DG																			
(Hames et al., 2003)	Corn stover (1 mm)	0.4-2.5			47				0.828					11.1			9.2		
(Hames et al., 2003)	Pretreated corn stover (1mm)	0.4-2.5			96				0.479					7.9			3.8		
(Sinnaeve et al., 1994)	Grass-hay (1 mm)	1.3-2.4	WMSC + D-1,5,5,1	14	817		0.99		0.70			0.65 *b	41.7				29.2	14.7	6.17
(Sinnaeve et al., 1994)	Tropical forages (1mm)	1.1-2.9	WMSC + d-1,5,5,1	14	857		0.95		1.17			1.07 *b	27.5				32.2	13.6	5.34
(Sinnaeve et al., 1994)	Maize (whole plants) (1 mm)	1.1-2.5	SNVDT + D-1,5,5,1	15	902		0.91		0.36			0.33 *b	23.6				8.5	8.0	1.2
WU																			
(Cozzolino et al., 2006a)	Whole maize silage	0.4-2.5	2D, SNVDT	7	90	0.91			0.65					4.8			20.56		
(Cozzolino et al., 2006a)	Whole maize silage	0.5-1.1	2D, SNVDT	4	90	0.90			0.77					4.1			20.56		

*a – since all particles were less than 3mm diameter a more appropriate classification here may be WG (i.e. wet and ground); *b = using a separate local calibration

Table B-13: Statistics for NIRS calibration equations for the ash contents of lignocellulosic feedstocks

Reference	Feedstock(s)	Wav. Range (10 ³ nm)	Pretreatment	F	Calibration Set					Validation Set					Const. Data			
					n	r ²	r _{cv} ²	SEC	SECV	n	r ²	SEP	RER	RPD	R	M	SD	
DG																		
(Montes et al., 2009)	Maize stover (1 mm particle size)	9.6-1.69		10	242	0.84	0.74	0.39	0.50									
(Sanderson et al., 1996)	Various crops and residues (1 mm)	1.1-2.5	SNVD		96	0.96		0.36	0.44	20	0.93	0.60		3.75		3.64	2.25	
(Sanderson et al., 1996)	Various crops and residues (1 mm)	1.4-2.5	SNVD		96	0.94		0.46	0.53	20	0.95	0.51		4.41		3.64	2.25	
(McTiernan et al., 2003)	Decomposing Pinus sylvestris needles (1 mm)				115	0.97		0.32	0.46			0.31	34.84	5.48	10.8	2.1	1.7	
(Ye et al., 2008)	Corn stover fractions (0.42 mm)	1-2.5	EMSC		35		0.90		1.08				7.05			4.48		
(Liu et al., 2010b)	Switchgrass fractions (0.42 mm)	1-2.5	EMSC		36		0.96		0.26				14.56			2.95		
(Liu et al., 2010b)	Switchgrass and corn stover fraction (0.42 mm)	1-2.5	EMSC		71		0.88		0.62	15		0.43	11.73			3.02		
(Hames et al., 2003)	Corn stover (1 mm)	0.4-2.5			47				1.029				13.20			13.6		
(Hames et al., 2003)	Pretreated corn stover (1 mm)	0.4-2.5			96				1.467				11.00			16.1		
(Liu and Chen, 2007)	Rice straw (0.425-0.25 mm)	1.33-2.44	D-1,5,10,1	10	34	0.86			0.72	9	0.857	0.8	9.75	3.50	7.8	11.8	2.8	
(Liu and Chen, 2007)*a	Rice straw (0.425-0.25 mm)	1.3-2.44	D-1,5,10,1	10	34	0.87			0.75	9	0.8813	0.6	15.70	4.67	9.4	7.8	2.8	
(Beining et al., 2000)	Freeze dried peat tables (FTIR)	1.25-2.5	1D			0.931										6.38		
(Reeves III and Van Kessel, 2002)	Dried dairy manure (0.85 mm)	0.4-2.5			96	0.819												
(Vavrova et al., 2008)	Plant litter (1 mm)	0.78-2.5	D-2,4,4,1	7	56	0.95	0.88	0.22	0.35				13.30	2.83	4.64	1.33	0.99	
(Liu et al., 2010a)	Switchgrass (0.42 mm)	1-2.5	SNV	9	36		0.91		0.278	5		0.44	8.66	2.09	3.81		0.92	
(Liu et al., 2010a)	Switchgrass (0.42 mm)	1-2.5	MSC	9	36		0.90		0.29	5		0.61	6.25	1.51	3.81		0.92	
(Liu et al., 2010a)	Switchgrass (0.42 mm)	1-2.5	1D (SG)	9	36		0.90		0.299	5		0.56	6.80	1.64	3.81		0.92	
(Liu et al., 2010a)	Switchgrass (0.42 mm)	1-2.5	2D (SG)	10	36		0.86		0.355	5		0.64	5.95	1.44	3.81		0.92	
(Liu et al., 2010a)	Switchgrass (0.42 mm)	1-2.5	EMSC	9	36		0.92		0.26	5		0.57	6.68	1.61	3.81		0.92	
(Liu et al., 2010a)	Switchgrass (0.42 mm)	1-2.5	1D (SG) + SNV	9	36		0.89		0.31	5		0.58	6.57	1.59	3.81		0.92	
(Liu et al., 2010a)	Switchgrass (0.42 mm)	1-2.5	SNV + 2D (SG)	10	36		0.84		0.368	5		0.68	5.60	1.35	3.81		0.92	
(Liu et al., 2010a)	Switchgrass (0.42 mm)	1-2.5	2D (SG) + SNV	10	36		0.85		0.307	5		0.64	5.95	1.44	3.81		0.92	
(Liu et al., 2010a)	Switchgrass (0.42 mm)	1-2.5	None	10	36		0.71		0.38	5		1.05	3.63	0.88	3.81		0.92	
(Liu et al., 2010a)	Corn-stover (0.42 mm)	1-2.5	SNV	9	35		0.86		0.91	5		0.44	17.30	4.05	7.61		1.78	
(Liu et al., 2010a)	Corn-stover (0.42 mm)	1-2.5	MSC	9	35		0.79		1.07	5		0.39	19.51	4.56	7.61		1.78	
(Liu et al., 2010a)	Corn-stover (0.42 mm)	1-2.5	2D (SG)	10	35		0.81		0.96	5		0.58	13.12	3.07	7.61		1.78	
(Liu et al., 2010a)	Corn-stover (0.42 mm)	1-2.5	EMSC	9	35		0.81		0.90	5		0.24	31.71	7.42	7.61		1.78	
(Liu et al., 2010a)	Corn-stover (0.42 mm)	1-2.5	2D (SG) + SNV	9	35		0.81		1.05	5		0.62	12.27	2.87	7.61		1.78	
DU																		
(Montes et al., 2009)	Maize stover, 8cm av. size (custom NIR unit)	9.6-1.69			80	0.51	0.34	0.88	1.02									
(Montes et al., 2009)	Maize stover, 4cm av. size (custom NIR unit)	9.6-1.69			80	0.53	0.38	0.87	1.00									
(Nkansah et al., 2010)	Yellow poplar wood blocks	0.8-2.5	None, PCA	10	37	0.99		0.44		17	0.87	1.22						
(Nkansah et al., 2010)	Yellow poplar wood blocks	1.3-1.8	None, PCA	10	37	0.98		0.49		17	0.83	1.48						
(Nkansah et al., 2010)	Yellow poplar wood blocks	0.8-2.5	1D (SG), PCA	5	37	0.96		0.75		17	0.82	1.33						
(Nkansah et al., 2010)	Yellow poplar wood blocks	1.3-1.8	1D (SG), PCA	6	37	0.98		0.56		17	0.92	0.99						
WU																		
(Fagan et al., 2011)*b	Miscanthus and Willow (<3 mm)	0.75-1.1	MSC 1D (SG)	5	44		0.58		0.42				7.73	3.60	3.2			

*a – acid insoluble ash; *b – since all particles were less than 3mm diameter a more appropriate classification here may be WG (i.e. wet and ground)

Appendix C Figures and Tables for Chapter 12: Analysis of Sugarcane Bagasse

Table C-1: Extractives-free data (% of dry matter on an extractives-free basis) including standard deviation of the duplicates (SD) for the 30 bagasse samples analysed.

NIR #	UNIV. CODE	EIA_EF		AIR_EF		AIA_EF		KL_EF		ASL_EF		ARA_EF_SRS		GAL_EF_SRS		RHA_EF_SRS		GLU_EF_SRS		XYL_EF_SRS		MAN_EF_SRS		TOT_EF_SRS	
		AV	SD	AV	SD	AV	SD	AV	SD	AV	SD	AV	SD	AV	SD	AV	SD	AV	SD	AV	SD	AV	SD	AV	SD
50001	SBOZ1	1.98	(0.13)	21.17		1.98		19.18		2.07		2.42		1.01		0.13		44.09		23.73		0.16		71.40	
50002	SBOZ2	3.37	(0.12)	22.52	(0.04)	3.36	(0.01)	19.16	(0.05)	2.06	(0.11)	2.38	(0.04)	1.02	(0.01)	0.09	(0.00)	42.31	(0.06)	23.32	(0.02)	0.15	(0.01)	69.19	(0.01)
50003	SBOZ3	2.95	(0.50)	21.40	(0.22)	2.17	(0.50)	19.23	(0.28)	2.27	(0.02)	2.41		0.85		0.11		43.72		24.75		0.11		71.84	
50004	SBOZ4	1.74		20.73	(0.09)	1.11	(0.28)	19.62	(0.19)	2.21	(0.19)	2.33	(0.03)	0.96	(0.02)	0.12	(0.00)	44.20	(0.21)	24.11	(0.22)	0.14	(0.00)	71.74	(0.39)
50006	SBOZ6	3.33		21.58	(0.04)	2.39	(0.18)	19.19	(0.14)	2.21	(0.05)	2.33	(0.02)	0.74	(0.00)	0.12	(0.00)	43.35	(0.02)	23.46	(0.07)	0.14	(0.00)	70.03	(0.06)
50007	SBOZ7	5.55	(0.27)	22.64	(0.19)	3.60	(0.13)	19.04	(0.07)	2.20	(0.07)	2.18	(0.03)	0.75	(0.02)	0.11	(0.00)	42.41	(0.09)	23.07	(0.07)	0.13	(0.00)	68.54	(0.12)
50011	SBOZ11	3.25	(0.04)	21.73	(0.18)	2.19	(0.01)	19.53	(0.17)	2.30	(0.08)	2.39	(0.00)	0.89	(0.01)	0.12	(0.00)	43.07	(0.05)	23.28	(0.00)	0.14	(0.01)	69.78	(0.04)
50014	SBOZ14	1.91	(0.12)	20.34	(0.15)	1.35	(0.06)	18.99	(0.08)	2.17	(0.14)	2.40	(0.02)	0.89	(0.01)	0.11	(0.00)	43.81	(0.04)	24.36	(0.15)	0.14	(0.02)	71.61	(0.07)
50015	SBOZ15	3.38	(0.42)	22.52	(0.07)	2.66	(0.18)	19.86	(0.11)	2.04	(0.01)	2.55	(0.02)	0.99	(0.01)	0.12	(0.00)	41.04	(0.10)	24.04	(0.22)	0.15	(0.01)	68.77	(0.16)
50016	SBOZ16	9.09	(0.17)	26.43	(0.06)	6.51	(0.05)	19.93	(0.01)	1.71	(0.01)	2.25	(0.02)	0.78	(0.01)	0.10	(0.00)	37.85	(0.18)	23.04	(0.11)	0.14	(0.03)	64.06	(0.23)
50017	SBOZ17	15.43	(0.35)	30.35	(0.31)	11.90	(0.52)	18.45	(0.21)	1.67	(0.03)	2.30	(0.00)	0.73	(0.01)	0.13	(0.00)	36.07	(0.03)	20.97	(0.07)	0.20	(0.00)	60.28	(0.11)
50019	SBOZ19	8.97	(0.00)	27.69	(0.43)	4.28		18.24	(0.16)	1.80	(0.09)	2.13	(0.02)	0.78	(0.02)	0.10	(0.00)	39.18	(0.59)	22.06	(0.28)	0.14	(0.01)	64.29	(0.83)
50022	SBOZ22	3.36	(0.04)	20.50	(0.17)	2.69	(0.12)	17.81	(0.05)	2.40	(0.01)	2.58	(0.03)	0.92	(0.01)	0.12	(0.00)	43.86	(0.10)	24.36	(0.04)	0.17	(0.00)	71.88	(0.11)
50025	SBOZ25	3.22		22.93		2.72		20.20		1.99		2.61		1.03		0.11		40.54		24.63		0.20		69.00	
50026	SBOZ26	3.22	(0.20)	22.55	(0.60)	3.03	(0.40)	19.52	(0.20)	2.20	(0.04)	2.24	(0.02)	0.91	(0.01)	0.11	(0.00)	42.56	(0.24)	23.78	(0.20)	0.15	(0.01)	69.65	(0.41)
50027	SBOZ27	3.57		21.92	(0.14)	2.78	(0.26)	19.14	(0.40)	2.24	(0.01)	2.35	(0.01)	0.82	(0.01)	0.13	(0.01)	42.89	(0.24)	23.57	(0.00)	0.13	(0.01)	69.76	(0.22)
50031	SBOZ31	2.76	(0.06)	21.80	(0.14)	1.41	(0.57)	20.39	(0.42)	2.11	(0.03)	2.38	(0.00)	0.95	(0.02)	0.11	(0.00)	42.53	(0.13)	23.98	(0.00)	0.13	(0.03)	69.97	(0.13)
50032	SBOZ32	6.02		23.38		4.34		19.04		2.12		2.42		0.91		0.12		41.47		23.44		0.16		68.41	
50033	SBOZ33	2.96		21.32	(0.20)	2.05	(0.11)	19.28	(0.10)	2.32	(0.07)	2.41	(0.01)	0.87	(0.00)	0.12	(0.00)	43.37	(0.20)	24.31	(0.05)	0.14	(0.00)	71.11	(0.25)
50034	SBOZ34	5.91	(0.02)	23.48	(0.28)	5.27	(0.36)	18.21	(0.08)	2.07	(0.03)	2.46	(0.01)	0.89	(0.00)	0.10	(0.00)	40.85	(0.33)	23.45	(0.19)	0.18	(0.02)	67.82	(0.50)
50037	SBOZ37	7.42	(0.51)	25.38		8.07		17.64	(0.24)	2.09	(0.04)	2.44	(0.02)	0.93	(0.02)	0.12	(0.01)	41.90	(0.11)	22.43	(0.19)	0.18	(0.00)	67.88	(0.30)
50038	SBOZ38	2.61	(0.03)	21.79	(0.10)	2.07	(0.02)	19.72	(0.12)	2.25	(0.06)	2.62	(0.00)	0.93	(0.01)	0.12	(0.00)	42.43	(0.03)	23.24	(0.04)	0.20	(0.02)	69.43	(0.06)
50040	SBOZ40	2.62		21.14	(0.09)	1.83	(0.13)	19.31	(0.22)	2.36	(0.03)	2.27	(0.02)	0.76	(0.00)	0.12	(0.00)	44.08	(0.05)	23.49	(0.12)	0.14	(0.00)	70.73	(0.16)
50041	SBOZ41	1.86	(0.01)	21.28	(0.05)	1.23	(0.12)	20.05	(0.16)	2.22	(0.06)	2.25	(0.02)	0.83	(0.01)	0.11	(0.01)	43.86	(0.23)	24.43	(0.00)	0.12	(0.02)	71.49	(0.21)
50042	SBOZ42	8.30	(0.14)	25.17	(0.17)	6.24	(0.76)	18.93	(0.59)	2.11	(0.05)	2.43	(0.01)	0.94	(0.00)	0.11	(0.00)	39.24	(0.04)	23.66	(0.02)	0.16	(0.02)	66.42	(0.08)
50043	SBOZ43	4.87		22.28		3.58		18.71		2.06		2.58		0.99		0.09		41.81		23.41		0.18		68.96	
50044	SBOZ44	2.92	(0.56)	21.26	(0.03)	2.21	(0.29)	19.04	(0.32)	2.29	(0.08)	2.41	(0.00)	0.72	(0.01)	0.12	(0.00)	44.29	(0.12)	24.15	(0.05)	0.12	(0.02)	71.69	(0.19)
50045	SBOZ45	3.36	(0.02)	22.75	(0.28)	2.60	(0.15)	20.14	(0.12)	2.02	(0.09)	2.47	(0.02)	0.99	(0.01)	0.12	(0.00)	41.59	(0.04)	23.93	(0.00)	0.18	(0.01)	69.17	(0.03)
50046	SBOZ46			23.16	(0.31)	3.54	(0.27)	19.61	(0.03)	2.07	(0.01)	2.50	(0.04)	0.98	(0.02)	0.12	(0.00)	41.02	(0.13)	23.69	(0.00)	0.17	(0.01)	68.36	(0.18)
50047	SBOZ47	3.72	(0.12)	19.38	(0.04)	1.18	(0.07)	18.20	(0.11)	2.17	(0.05)	2.40	(0.00)	0.78	(0.00)	0.11	(0.00)	43.36	(0.04)	24.06	(0.01)	0.14	(0.00)	70.76	(0.03)

Table C-2: Whole Mass compositional data (% of dry matter on a whole mass basis), including standard deviation of the duplicates (SD) for some statistics, for the 30 bagasse samples analysed.

NIR #	UNIV. CODE	EXTR_P D_AV	EXTR_P D_SD	EXTR_C V_AV	EXTR_C V_SD	ASH_ AV	ASH_ SD	ESA	ASA	KL	ASL	AIR	AIA	EIA	ARA_ SRS	GAL_ SRS	RHA_ SRS	GLU_ SRS	XYL_ SR S	MAN_ S RS	TOT_ SRS	
50001	SBOZ1	3.48		4.10		2.21	(0.38)	0.31	-0.01	18.51	2.00	20.43	1.92	1.91	2.33	0.97	0.13	42.55	22.90	0.15	68.92	
50002	SBOZ2	3.89		3.75		4.28	(0.18)	1.04	0.01	18.41	1.98	21.64	3.23	3.24	2.28	0.99	0.09	40.67	22.42	0.15	66.50	
50003	SBOZ3	3.17		4.05		2.60	(0.53)	-0.25	0.75	18.62	2.19	20.72	2.10	2.85	2.34	0.82	0.11	42.34	23.96	0.11	69.56	
50004	SBOZ4	4.12		4.27		1.98	(0.32)	0.32	0.60	18.81	2.12	19.88	1.07	1.67	2.23	0.92	0.11	42.38	23.12	0.14	68.79	
50006	SBOZ6	5.93		6.34		3.48	(0.45)	0.35	0.88	18.05	2.08	20.30	2.25	3.13	2.20	0.70	0.12	40.78	22.07	0.13	65.88	
50007	SBOZ7	5.25		5.45		5.78	(0.01)	0.53	1.85	18.04	2.09	21.45	3.41	5.26	2.07	0.71	0.11	40.18	21.86	0.12	64.94	
50011	SBOZ11	3.63	(0.16)	4.08	(0.34)	3.49	(0.08)	0.37	1.01	18.82	2.22	20.94	2.11	3.13	2.31	0.86	0.12	41.50	22.44	0.14	67.25	
50014	SBOZ14	3.53	(0.38)	4.54	(0.34)	1.80	(0.06)	-0.04	0.54	18.32	2.09	19.62	1.30	1.84	2.32	0.86	0.11	42.26	23.50	0.14	69.08	
50015	SBOZ15	3.72	(0.39)	3.93		4.21	(0.13)	0.95	0.70	19.12	1.96	21.68	2.56	3.26	2.45	0.96	0.12	39.51	23.15	0.15	66.22	
50016	SBOZ16	5.19		5.43		9.79	(0.04)	1.17	2.45	18.89	1.62	25.06	6.17	8.62	2.14	0.74	0.09	35.89	21.84	0.13	60.73	
50017	SBOZ17			5.60		17.46	(1.29)															
50019	SBOZ19	5.13		5.27		4.78	(0.91)	-3.73	4.45	17.30	1.70	26.27	4.06	8.51	2.02	0.74	0.10	37.17	20.93	0.13	61.00	
50022	SBOZ22	3.27	(0.02)	4.21	(0.16)	3.94	(0.38)	0.69	0.65	17.23	2.32	19.83	2.61	3.25	2.49	0.89	0.11	42.42	23.56	0.16	69.53	
50025	SBOZ25	4.56		4.36		3.51	(0.11)	0.45	0.47	19.28	1.90	21.88	2.60	3.07	2.49	0.98	0.11	38.69	23.51	0.19	65.86	
50026	SBOZ26	3.75		3.85		5.68	(0.28)	2.58	0.18	18.79	2.12	21.71	2.92	3.10	2.15	0.88	0.10	40.97	22.89	0.15	67.03	
50027	SBOZ27	5.72		5.72		4.27	(0.11)	0.90	0.74	18.05	2.11	20.67	2.62	3.37	2.21	0.77	0.12	40.43	22.22	0.13	65.77	
50031	SBOZ31	3.64		3.72		2.53	(0.07)	-0.13	1.30	19.65	2.03	21.00	1.36	2.66	2.29	0.91	0.10	40.99	23.11	0.13	67.43	
50032	SBOZ32	4.12	(0.38)	4.57	(0.22)	4.50	(0.85)	-1.27	1.61	18.26	2.04	22.42	4.16	5.77	2.32	0.88	0.11	39.77	22.47	0.16	65.59	
50033	SBOZ33	4.63		4.97		3.77	(0.02)	0.95	0.87	18.38	2.21	20.34	1.95	2.82	2.30	0.83	0.11	41.36	23.19	0.13	67.81	
50034	SBOZ34	4.02		4.24		7.14	(0.29)	1.47	0.62	17.48	1.99	22.54	5.06	5.67	2.37	0.85	0.09	39.21	22.50	0.17	65.10	
50037	SBOZ37	3.94		3.99		8.20	(0.97)	1.07	-0.63	16.94	2.00	24.38	7.76	7.12	2.35	0.90	0.11	40.25	21.55	0.17	65.21	
50038	SBOZ38	5.11		5.15		2.71	(0.23)	0.24	0.51	18.71	2.13	20.68	1.96	2.47	2.49	0.89	0.12	40.26	22.05	0.19	65.88	
50040	SBOZ40	5.21		5.39		2.71	(0.19)	0.23	0.75	18.30	2.24	20.04	1.73	2.48	2.15	0.72	0.11	41.78	22.26	0.13	67.05	
50041	SBOZ41	4.21		4.64		2.22	(0.52)	0.44	0.61	19.21	2.13	20.39	1.18	1.78	2.16	0.79	0.11	42.01	23.40	0.11	68.48	
50042	SBOZ42	3.26		3.89		9.42	(0.33)	1.39	1.99	18.31	2.04	24.35	6.04	8.03	2.35	0.91	0.11	37.97	22.89	0.16	64.26	
50043	SBOZ43	3.06		3.47		4.23	(0.24)	-0.49	1.25	18.14	1.99	21.60	3.47	4.72	2.50	0.96	0.09	40.53	22.70	0.17	66.85	
50044	SBOZ44	4.80		5.50		3.34	(0.12)	0.56	0.68	18.13	2.18	20.23	2.11	2.78	2.29	0.68	0.11	42.17	22.99	0.12	68.25	
50045	SBOZ45	3.38	(0.13)	4.25	(0.22)	4.06	(0.31)	0.81	0.73	19.46	1.95	21.98	2.51	3.24	2.39	0.96	0.11	40.19	23.12	0.18	66.83	
50046	SBOZ46	4.19		3.96		7.10	(0.07)			18.79	1.98	22.19	3.39		2.39	0.93	0.12	39.30	22.70	0.16	65.50	
50047	SBOZ47	5.90				3.70	(0.01)	0.20	2.39	17.13	2.05	18.24	1.11	3.50	2.26	0.74	0.11	40.80	22.64	0.14	66.58	

Table C-3: Histograms (% whole dry mass basis), with associated statistics, for the concentration values, for a range of constituents, of the bagasse samples. Note, sample 50017 is excluded from WM and AF statistics, but included in EF statistics, for lignocellulosic components in this Table and Table C-4.

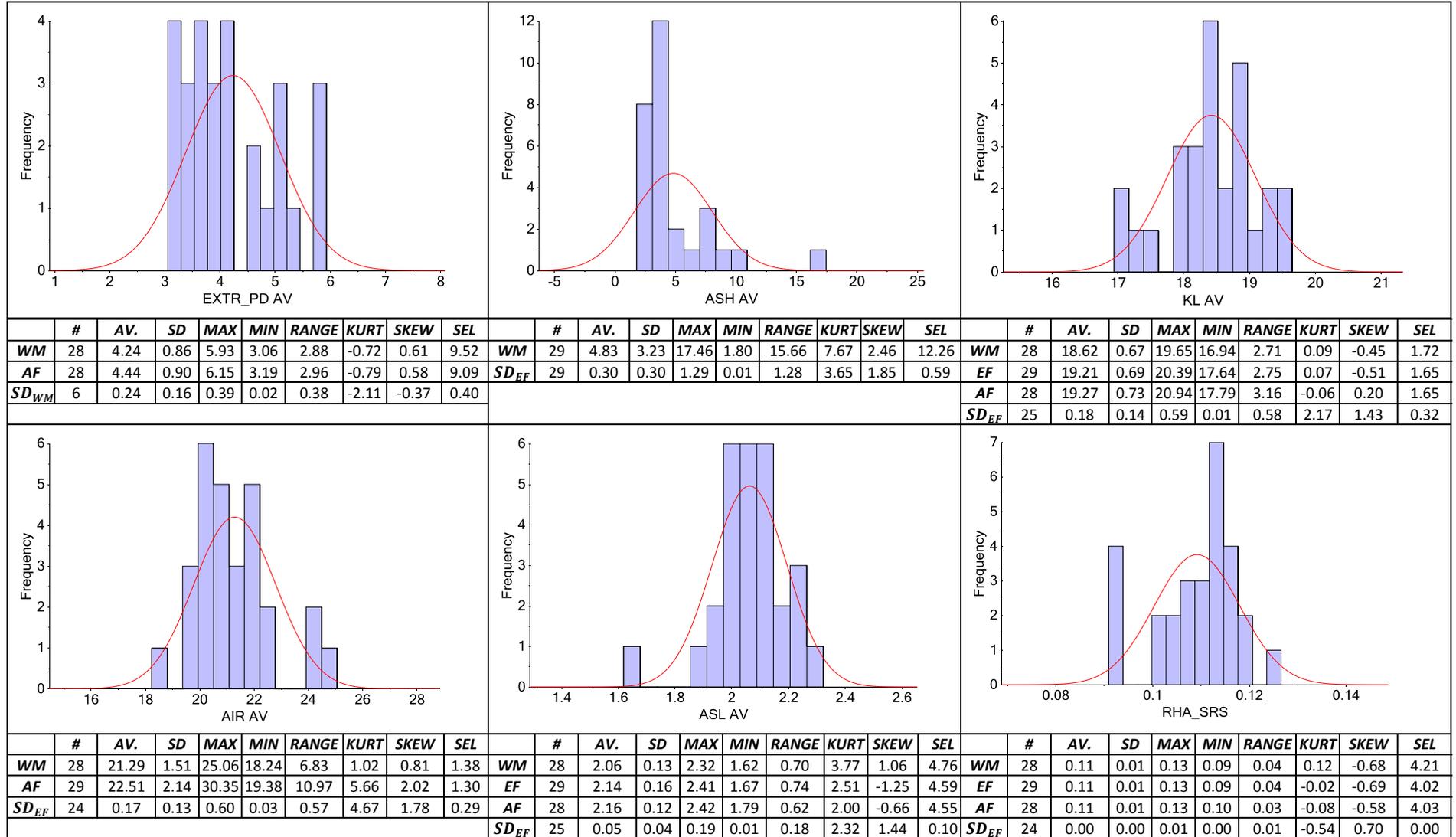


Table C-4: Histograms (% whole mass basis), with associated statistics, for the concentration values, for a range of constituents, of the bagasse samples.

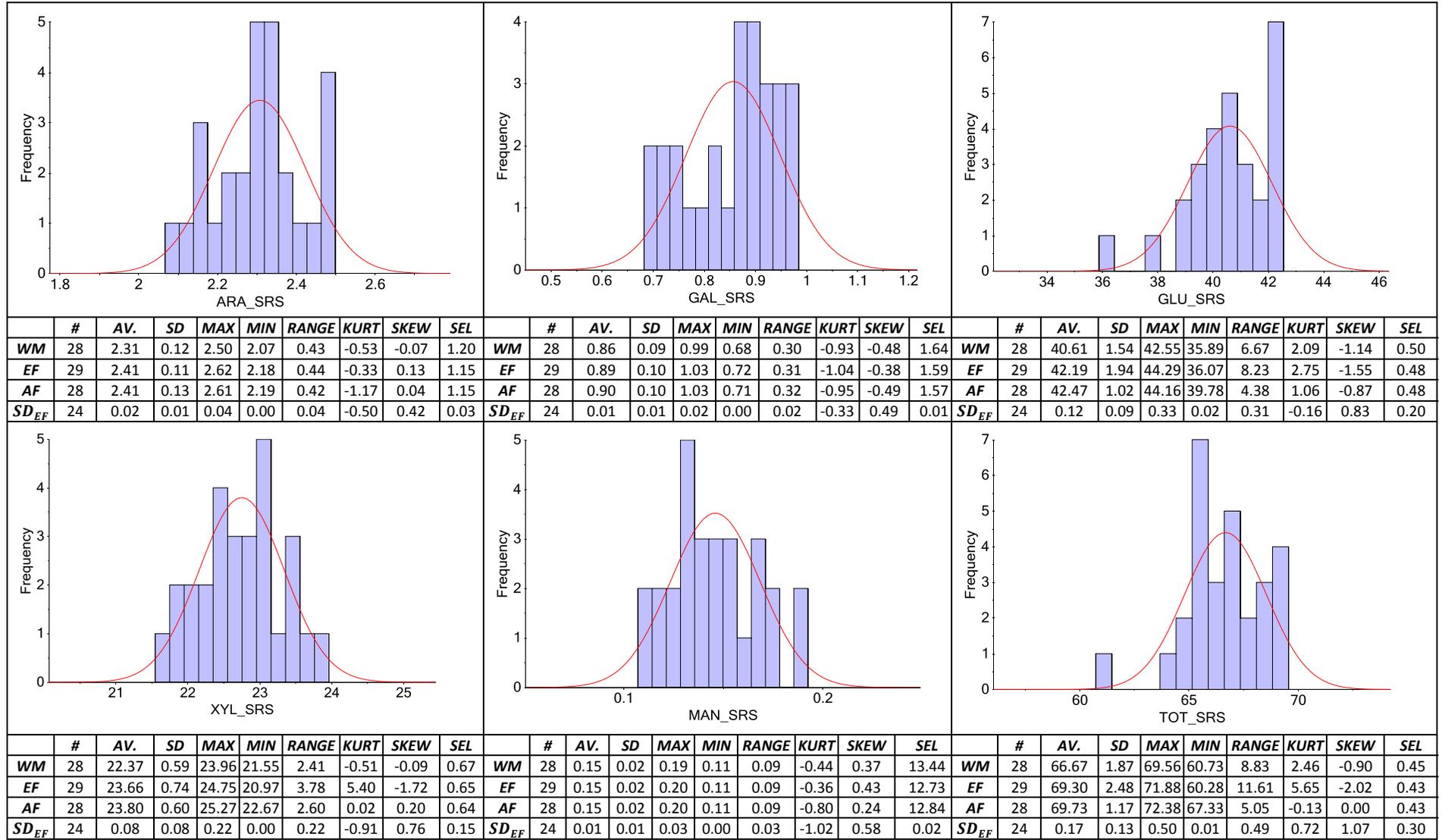


Table C-5: Pearson correlation coefficients between the various constituents for the 29 sugarcane bagasse samples analysed (sample 19 is not included). Absolute values greater than 0.5 are highlighted in bold.

	EXTR_CV	ASH	ESA	ASA	KL	ASL	AIR	AIA	EIA	ARA_SRS	GAL_SRS	RHA_SRS	GLU_SRS	XYL_SRS	MAN_SRS	TOT_SRS
EXTR_PD	0.904	0.009	0.015	0.301	-0.224	-0.060	-0.226	-0.119	0.008	-0.493	-0.715	0.124	-0.182	-0.535	-0.236	-0.387
EXTR_CV		0.160	-0.050	0.246	-0.170	0.091	-0.221	-0.131	-0.025	-0.504	-0.832	0.258	-0.042	-0.402	-0.353	-0.237
ASH			0.534	0.309	-0.260	-0.530	0.852	0.905	0.945	-0.076	-0.022	-0.338	-0.795	-0.485	0.220	-0.805
ESA				-0.221	-0.053	-0.139	0.348	0.346	0.227	-0.203	0.015	-0.185	-0.290	-0.139	0.075	-0.290
ASA					0.011	-0.269	0.096	0.068	0.441	-0.255	-0.408	-0.273	-0.474	-0.132	-0.256	-0.465
KL						-0.251	0.079	-0.353	-0.322	-0.008	0.322	0.040	-0.113	0.360	-0.037	0.037
ASL							-0.646	-0.495	-0.545	-0.001	-0.220	0.440	0.785	0.295	-0.269	0.719
AIR								0.904	0.851	0.054	0.239	-0.323	-0.791	-0.389	0.326	-0.748
AIA									0.925	0.055	0.092	-0.312	-0.685	-0.520	0.323	-0.711
EIA										-0.055	-0.081	-0.403	-0.797	-0.518	0.189	-0.815
ARA_SRS											0.675	0.090	-0.042	0.350	0.719	0.181
GAL_SRS												-0.046	-0.062	0.321	0.648	0.151
RHA_SRS													0.394	0.086	-0.053	0.351
GLU_SRS														0.420	-0.353	0.938
XYL_SRS															-0.115	0.695
MAN_SRS																-0.235

Table C-6: Summary data for the sugar recoveries of the nine hydrolysis batches that were involved in the analysis of the sugarcane bagasse samples.

Batch #	Number of SRS	Average of the Sugar Recoveries for Each Batch (%)						Standard Deviation of Sugar Recoveries for Batch (%)					
		Ara	Gal	Rha	Glu	Xyl	Man	Ara	Gal	Rha	Glu	Xyl	Man
683	3	92.15	95.00	92.97	95.17	86.35	94.70	0.18	0.17	3.09	0.21	0.20	1.04
685	3	92.10	94.74	94.02	95.16	86.37	94.08	0.33	0.38	0.53	0.26	0.36	0.59
688	3	92.24	94.79	94.42	95.07	86.39	92.28	0.11	0.26	0.54	0.12	0.24	1.84
690	3	91.99	94.54	94.53	94.79	86.54	93.32	0.17	0.17	0.34	0.15	0.16	1.80
691	1	91.89	94.79	94.27	94.52	86.38	93.68	-	-	-	-	-	-
694	3	91.73	94.51	94.12	94.45	86.15	93.30	0.45	0.70	0.81	0.32	0.21	0.08
696	2	91.63	94.22	93.26	94.51	86.13	93.06	0.42	0.52	0.38	0.17	0.02	0.74
698	3	91.99	94.37	92.23	94.63	86.63	94.49	1.02	0.88	0.50	0.97	0.76	2.81
808	3	91.98	95.01	94.15	94.17	86.25	96.94	0.04	0.16	0.35	0.13	0.22	1.34
Av		91.97	94.66	93.77	94.72	86.35	93.98	0.34	0.41	0.82	0.29	0.27	1.28
Max		92.24	95.01	94.53	95.17	86.63	96.94	1.02	0.88	3.09	0.97	0.76	2.81
Min		91.63	94.22	92.23	94.17	86.13	92.28	0.04	0.16	0.34	0.12	0.02	0.08
Range		0.61	0.79	2.29	1.01	0.50	4.66	0.98	0.72	2.75	0.85	0.74	2.73
STDEV		0.20	0.27	0.78	0.35	0.16	1.34	0.31	0.27	0.93	0.28	0.22	0.86

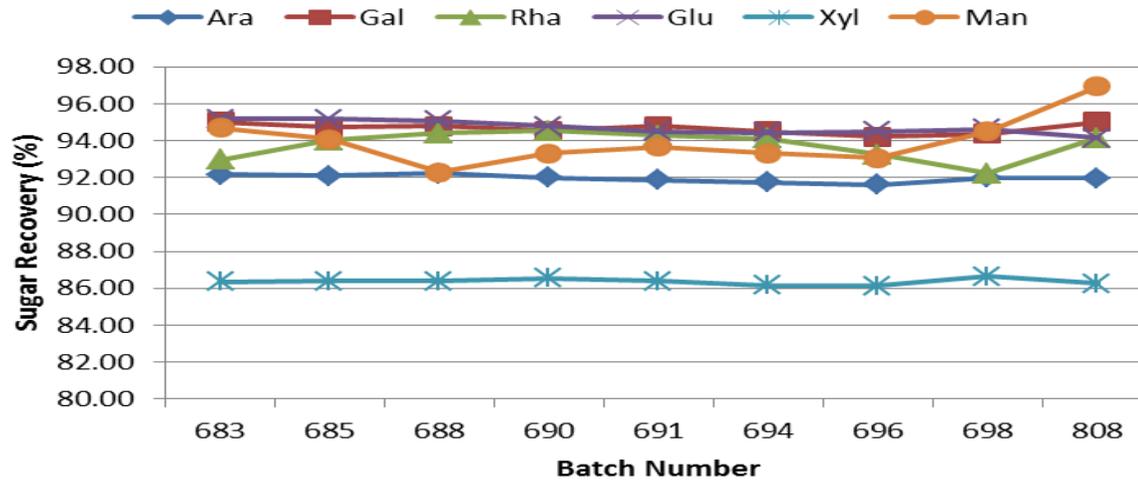


Figure C-1: A chart displaying the sugar recovery values for each batch, determined from the sugar recovery solutions (SRS).

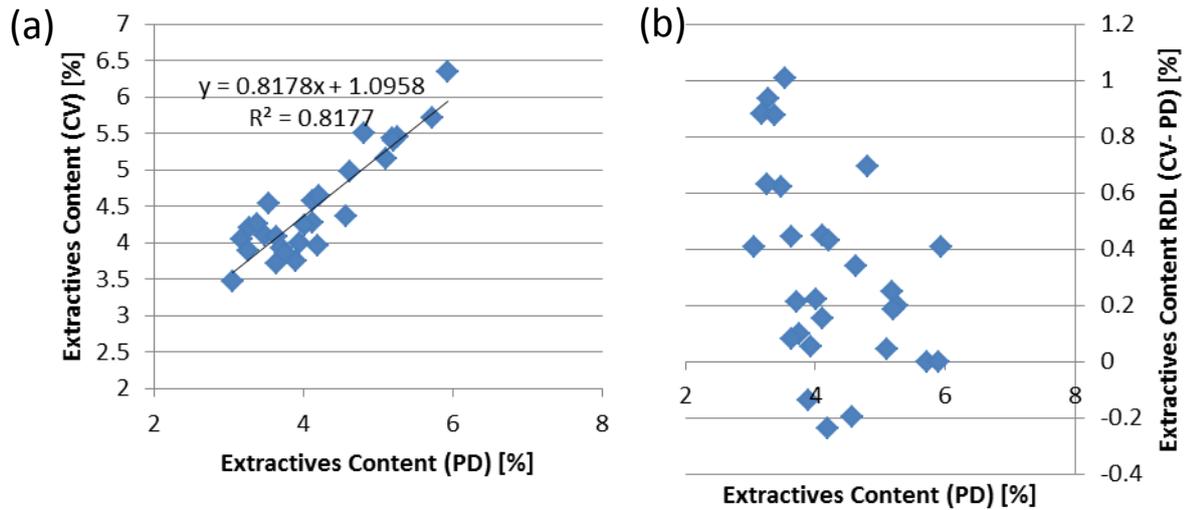


Figure C-2: Bagasse extractives-content plots comparing EXTR_PD and EXTR_CV data. (a) A scatter plot of the extractives content as measured by the petridish method (PD) compared with the extractives content as measured by the collection vial method (CV); (b) A plot of the extractives content residual (determined as EXTR_CV minus EXTR_PD) against EXTR_PD.

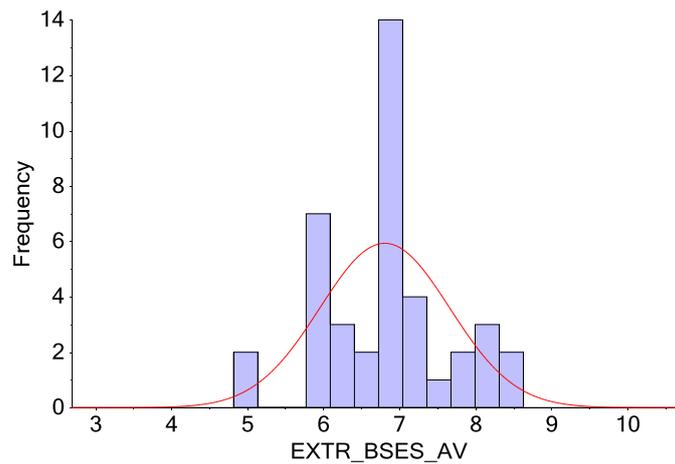


Figure C-3: A histogram for extractives data (% whole dry mass) obtained at the BSES laboratories.

Table C-7: Summary data for the extractives content data as determined at the BSES laboratories and at the UL laboratories (EXTR_PD and EXTR_CV methods).

	#	AV. (%)	SD (%)	MAX (%)	MIN (%)	RANGE (%)	KURT	SKEW	SEL (%)	SEL/AV (%)
BSES DATA										
WM	40	6.81	0.85	8.63	4.82	3.81	0.12	0.07		17.11
SD_{WM}	40	0.68	0.47	1.59	0.00	1.58	-0.95	0.31	1.16	
EXTR_PD DATA										
WM	29	4.27	0.86	5.93	3.06	2.88	-0.87	0.51		9.45
SD_{WM}	6	0.24	0.16	0.39	0.02	0.38	-2.11	-0.37	0.40	
EXTR_CV DATA										
WM	29	4.58	0.74	6.34	3.47	2.88	-0.62	0.63		8.16
SD_{WM}	5	0.25	0.08	0.34	0.16	0.18	-2.48	0.20	0.37	

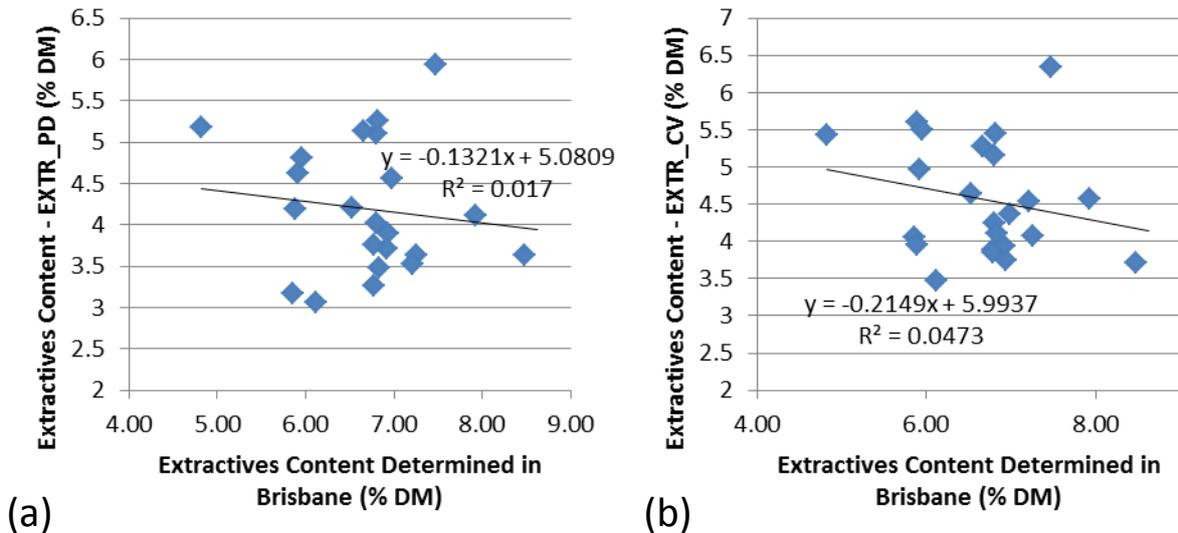


Figure C-4: Bagasse extractives-content plots comparing UL and BSES data. (a) Extractives content of the bagasse samples as determined in the BSES labs and the extractives content as measured by the EXTR_PD method; (b) a scatter plot involving the extractives content of the bagasse samples as determined in the BSES labs and the extractives content as measured by the EXTR_CV method.

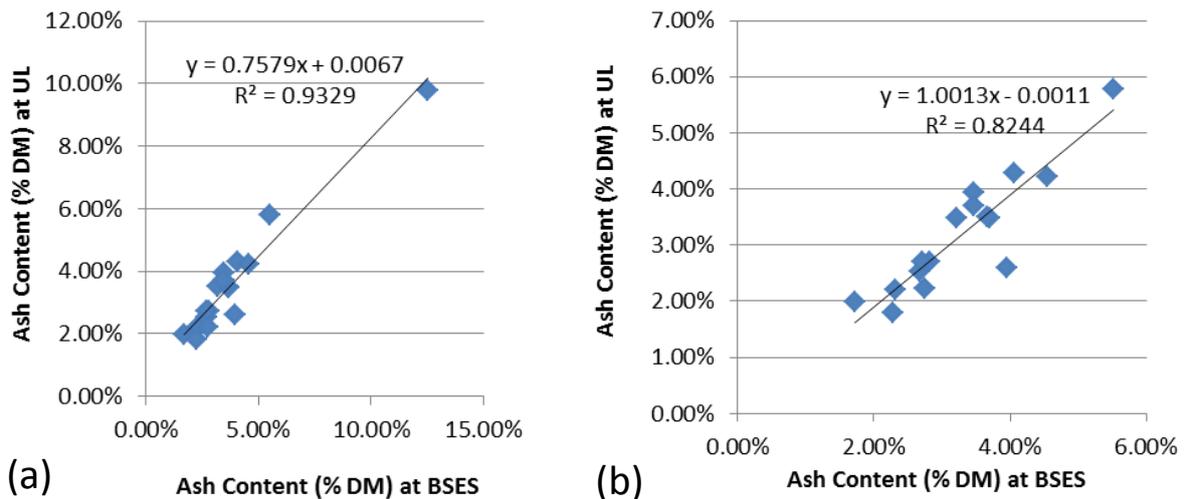


Figure C-5: Bagasse ash-content plots comparing UL and BSES data. (a) UL ash content versus BSES ash content, (b) the same plot as (a) but with the high ash content sample excluded.

Table C-8: Summary statistics for the ash contents and moisture contents of the 47 bagasse samples analysed at BSES. Ash data are included for the DB fraction and the crude ash analysis that took place of the WU fraction.

	#	AV. (%)	SD (%)	MAX (%)	MIN (%)	RANGE (%)	KURT	SKEW	SEL (%)	SEL/AV (%)
DB Fraction Ash Content										
WM	47	5.33	3.50	17.77	1.72	16.05	5.15	2.20		8.91
SD_{WM}	47	0.23	0.24	1.29	0.00	1.29	7.59	2.41	0.47	
WU Fraction Ash Content										
WM	42	5.96	3.39	19.82	2.20	17.63	6.05	2.12		18.37
SD_{WM}	42	0.61	0.48	1.97	0.02	1.95	0.40	1.01	1.10	
WU Fraction Moisture Content										
WB	47	43.63	3.87	48.51	31.01	17.50	1.65	-1.28		4.15
SD_{WM}	47	0.95	0.87	3.76	0.07	3.69	2.95	1.65	1.81	

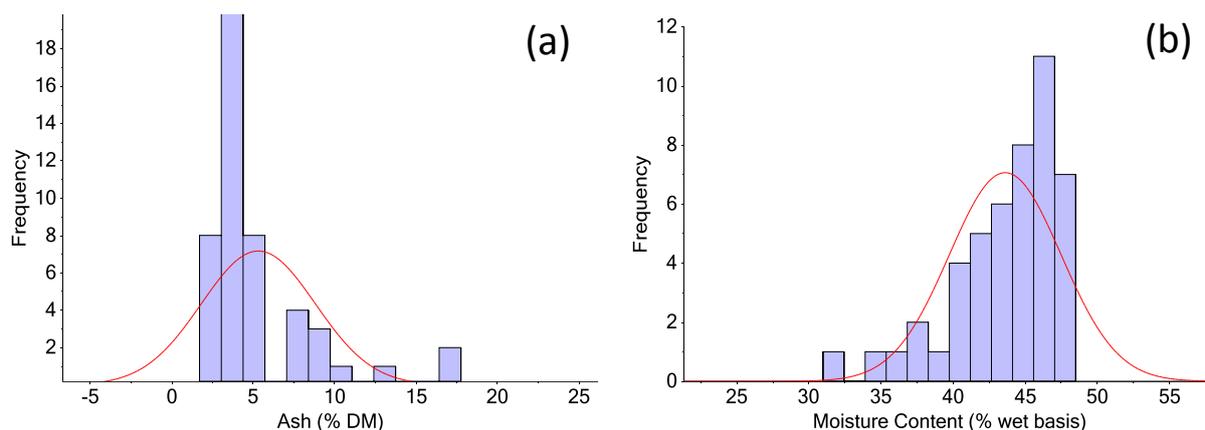


Figure C-6: Ash and moisture content histograms. (a) A histogram of the ash contents (% whole dry mass) of the DB samples as measured at BSES; (b) a histogram of the moisture contents (% wet basis) of the WU samples.

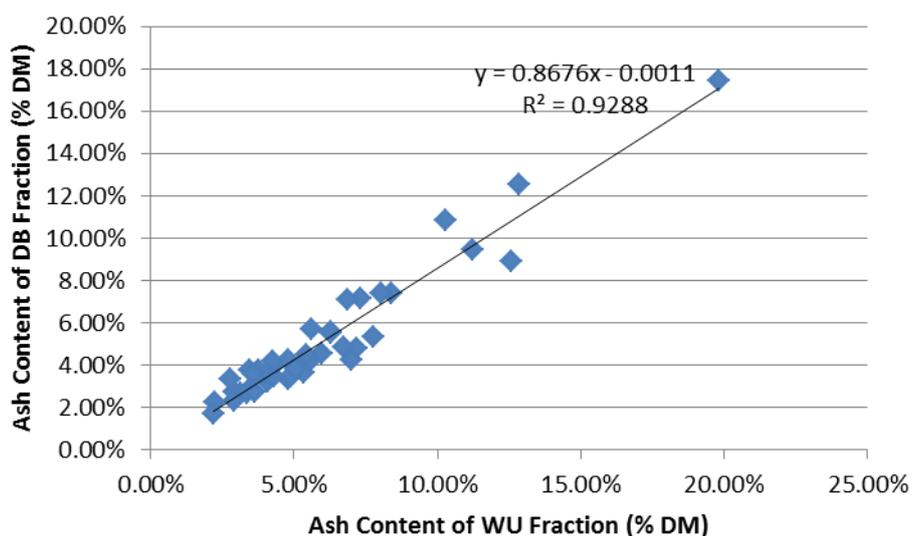


Figure C-7: Plot of the WU ash (% whole dry mass) contents versus the DB ash contents.

Table C-9: Statistics for the PCR for a range of bagasse chemical constituents where the predictive variables are the other constituents in this table.

Component	# PCs	R^2_{cal}	RMSEC (%)	Slope	Offset	R^2_{cv}	RMSECV	Slope	Offset
EXTR_PD	8	0.649	0.498	0.649	1.489	0.349	0.705	0.542	1.934
ASH	3	0.914	0.933	0.943	0.305	0.885	1.12	0.857	0.667
ARA_EF_SRS	9	0.784	0.052	0.780	0.531	0.611	0.072	0.719	0.675
GAL_EF_SRS	9	0.788	0.043	0.825	0.153	0.584	0.062	0.677	0.289
RHA_EF_SRS	6	0.429	0.007	0.428	0.066	0.101	0.009	0.184	0.093
GLU_EF_SRS	5	0.943	0.454	0.958	1.760	0.898	0.630	1.012	-0.548
XYL_EF_SRS	2	0.648	0.430	0.649	8.298	0.575	0.489	0.528	11.194
MAN_EF_SRS	7	0.787	0.011	0.810	0.030	0.622	0.015	0.670	0.050
KL_EF_SRS	9	0.762	0.329	0.763	4.539	0.475	0.506	0.687	6.00
ASL_EF_SRS	2	0.735	0.083	0.732	0.574	0.663	0.096	0.611	0.834

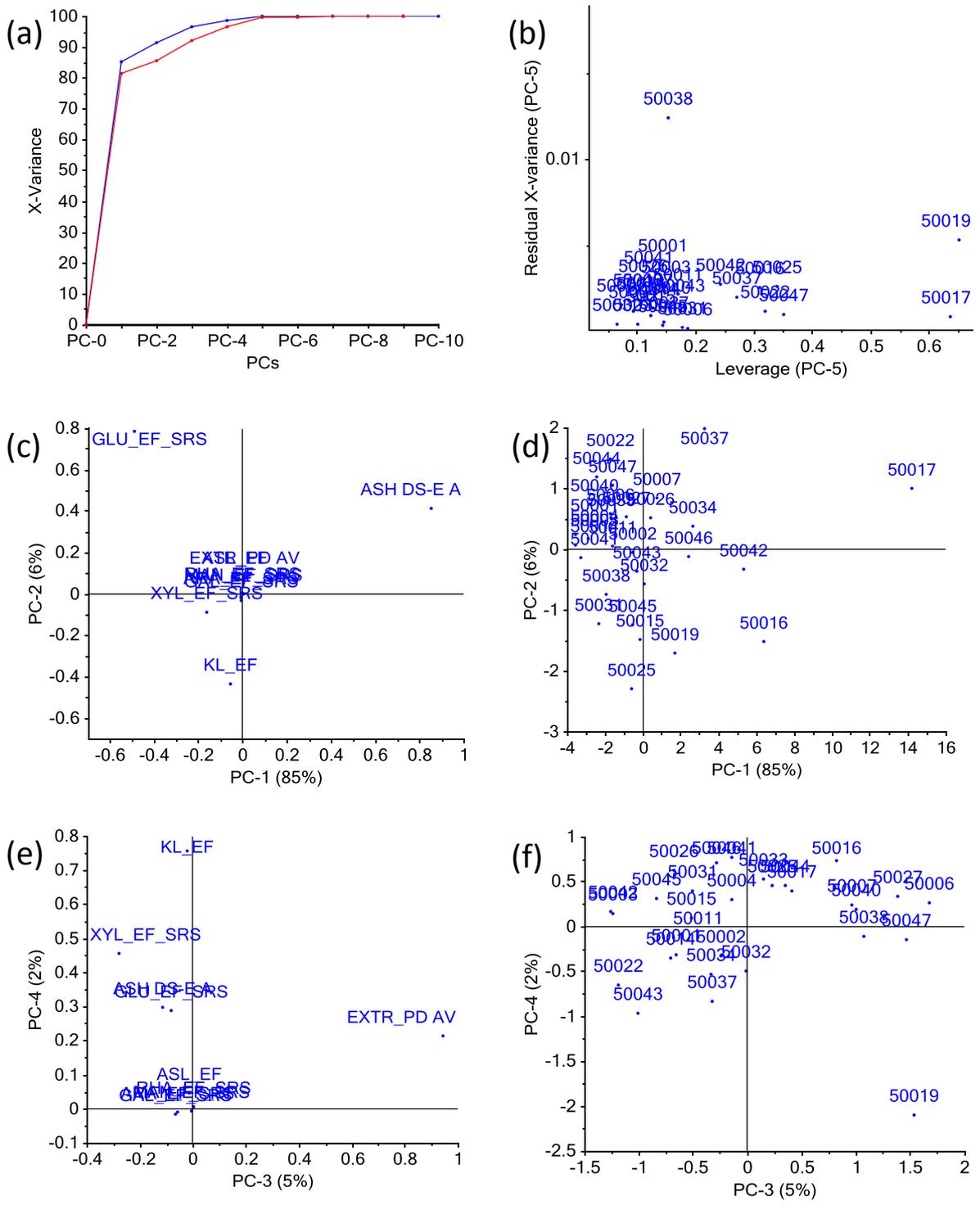


Figure C-8: Data correspond to the PCA involving 10 variables (chemical data) and 30 bagasse samples; (a) an explained variance plot with up to 10 PCs; (b) an influence plot using a 5 PC model; (c) a PC1 vs. PC2 loadings plot; (d) a PC1 vs. PC2 scores plot; (e) a PC3 vs. PC4 loading plot; (f) A PC3 vs. PC4 scores plot.

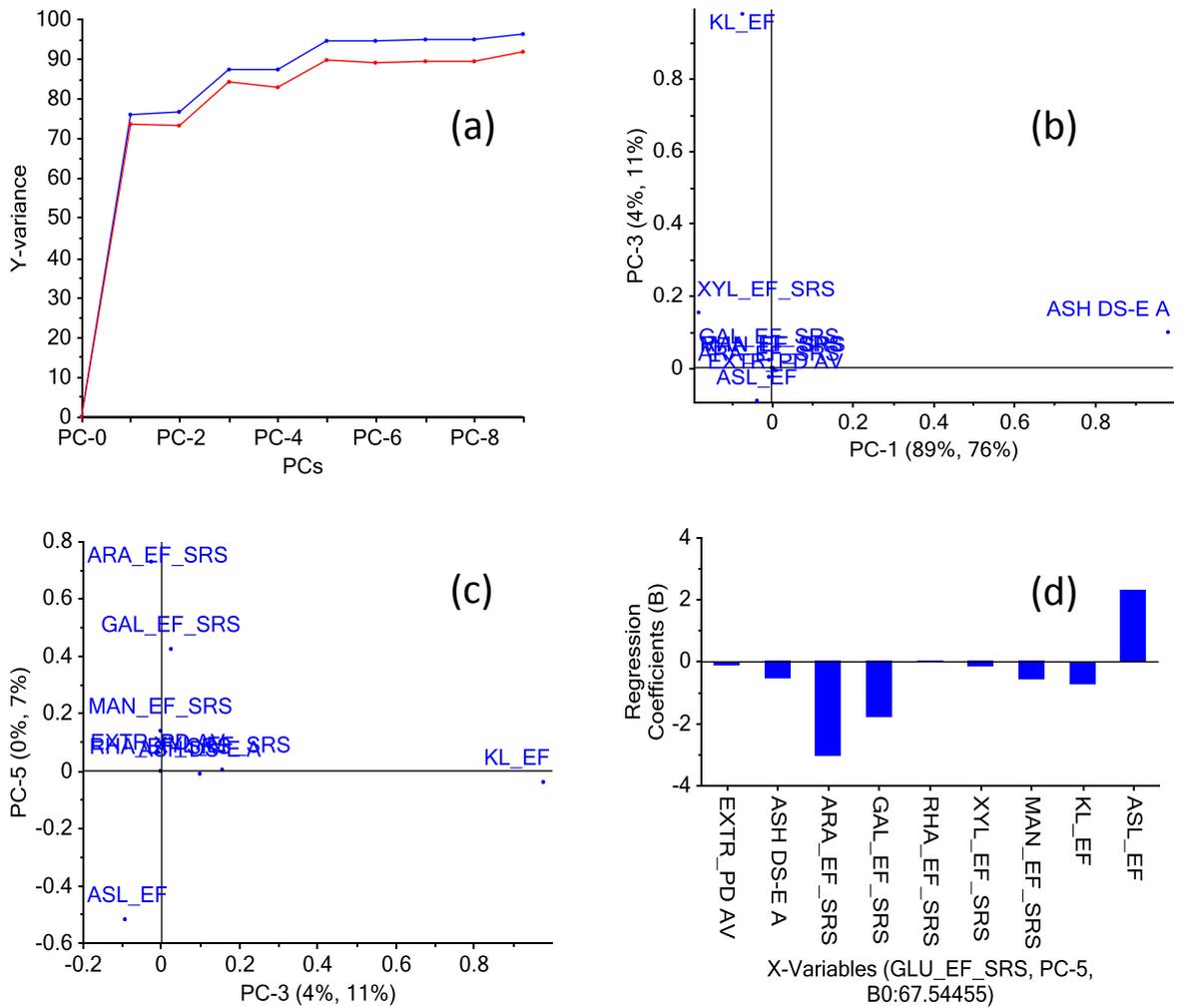


Figure C-9: Figures for a PCR for the glucose content of bagasse samples using the chemical data for these samples. (a) Explained variance plot for a PCR for glucan; (b) PC1 vs. PC3 loadings plot for a PCR for glucan; (c) PC3 vs. PC5 loadings plot for a PCR for glucan; (d) Regression coefficients plot for a 5 PC model for glucan.

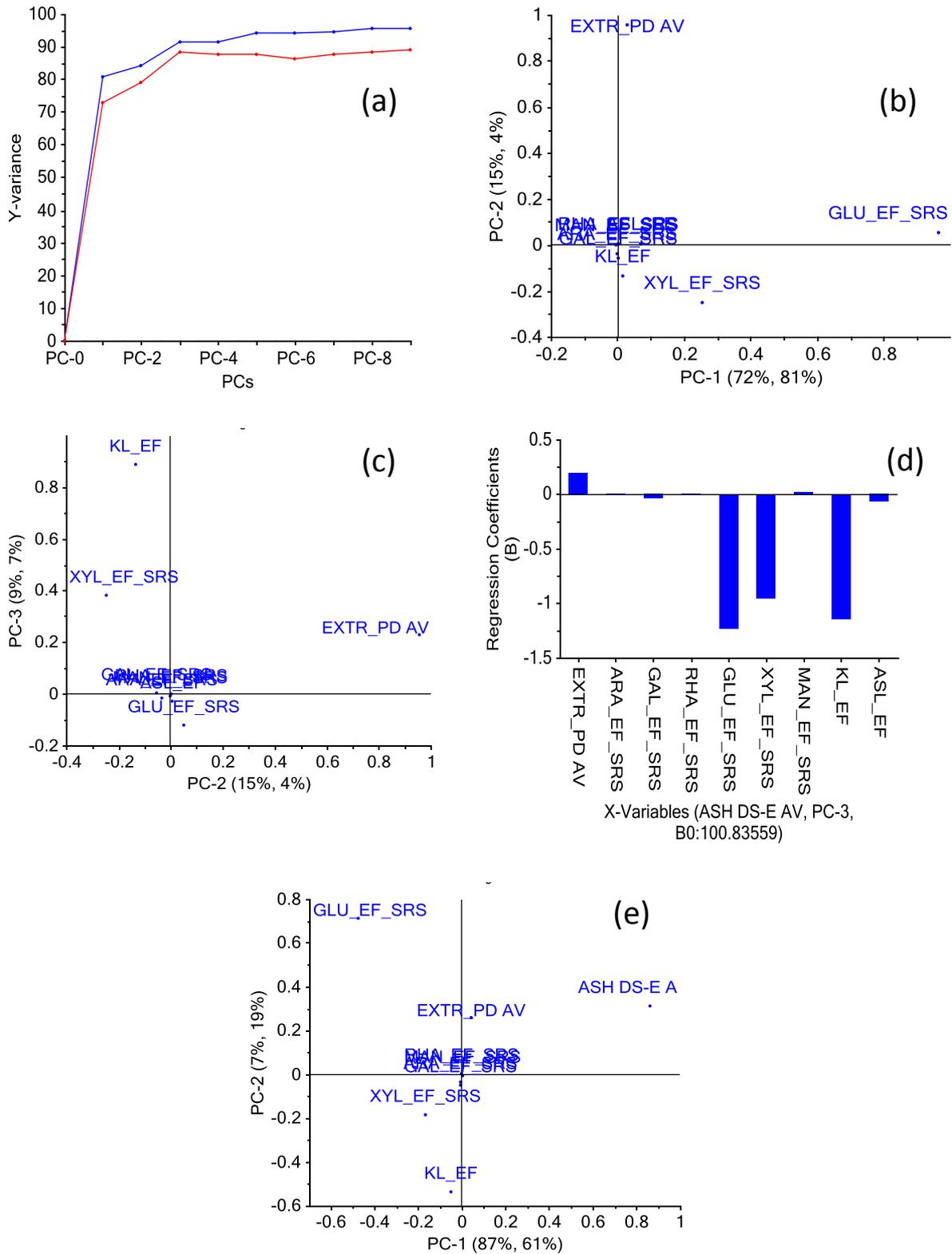


Figure C-10: Plots relating to PCRs using chemical data. (a) Explained variance plot for a PCR for ash; (b) PC1 vs. PC2 loadings plot for a PCR for ash; (c) PC2 vs. PC3 loadings plot for a PCR for ash; (d) Regression coefficients plot for a 3 PC model for ash; (e) PC1 vs. PC2 loadings plot for a PCR for ASL.

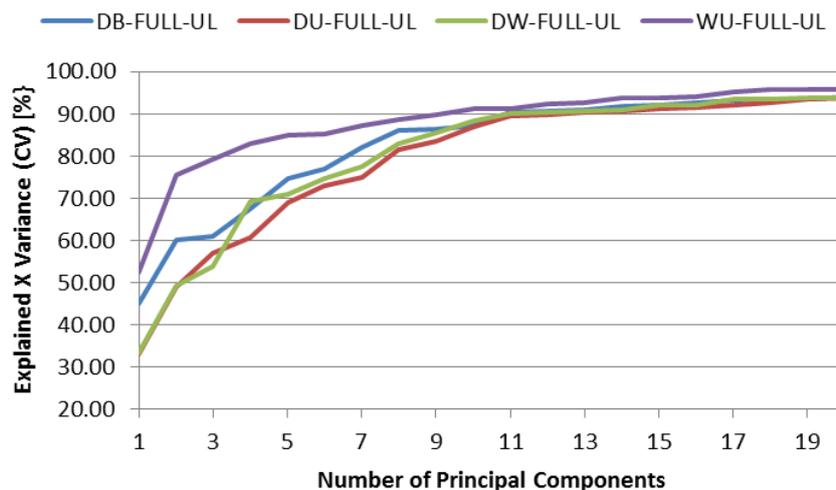


Figure C-11: Total explained X-variance under full cross-validation, according to PCA models using different numbers of principal components, for the 30 samples analysed at UL and their different spectral data sets (DB, DU, DW, WU), using the 400-2500 nm region.

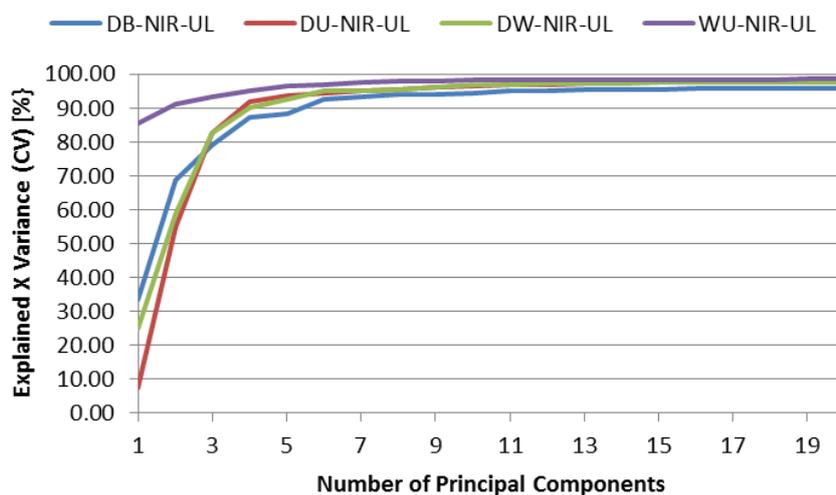


Figure C-12: Same as Figure C-11 but using the 1100-2500 nm spectral region for the PCA.

Table C-10: Explained X variance (in %) for the full cross validation, using up to 20 PCs, for the FULL spectral region (400-2500 nm) or the 1100-2500 nm region (NIR) for all 47 samples (ALL) or the 30 analysed at UL (UL). The DB scans are used for all sets.

Set	PC Number																			
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
FULL-UL	45.14	60.24	61.04	67.43	74.77	77.09	82.08	86.02	86.43	87.22	90.41	90.75	90.92	91.83	91.97	92.55	93.21	93.62	93.63	94.06
FULL-ALL	46.06	63.41	73.07	76.37	78.83	84.28	86.69	88.23	89.71	90.34	91.25	92.53	93.31	94.36	94.88	95.07	95.52	95.71	95.76	95.90
NIR-UL	33.23	68.87	79.21	87.34	88.50	92.69	93.29	93.94	94.07	94.46	94.99	95.23	95.37	95.41	95.54	95.73	95.80	95.83	95.88	95.91
NIR-ALL	51.39	71.67	81.35	88.30	92.15	94.63	95.26	95.90	96.41	96.58	96.80	96.81	97.07	97.16	97.30	97.39	97.41	97.51	97.55	97.61

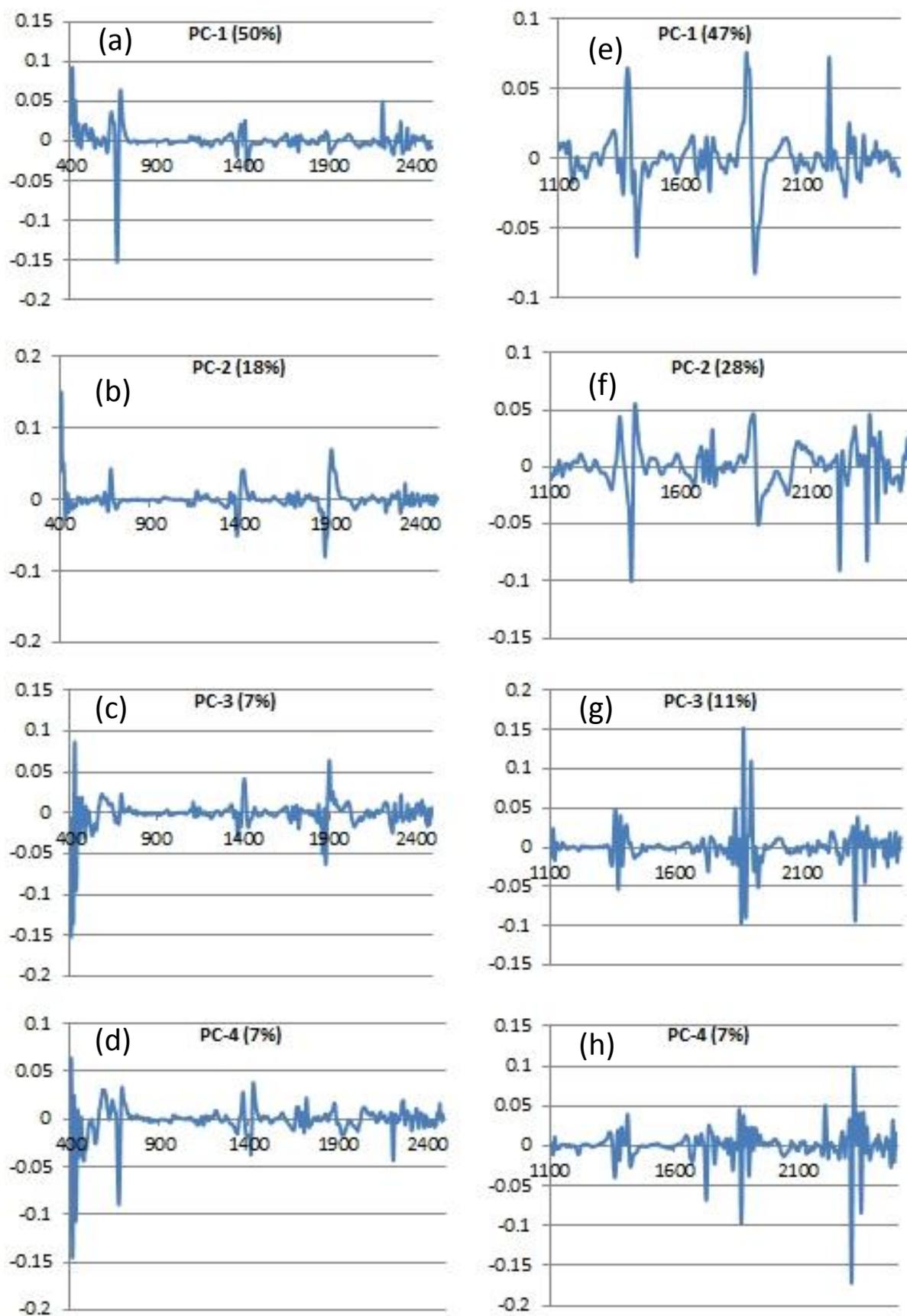


Figure C-13: Loadings plots for the DB bagasse models. (a)-(d) PC1-4 loading plots for a PCA over the 400-2500 nm spectral region using the 30 UL samples; (e)-(h) PC1-4 loading plots for a PCA over the 1100-2500 nm spectral region. Wavelength (nm) on x-axis, loading value on y-axis. Numbers in brackets indicate X-variance in calibration set explained by that PC.

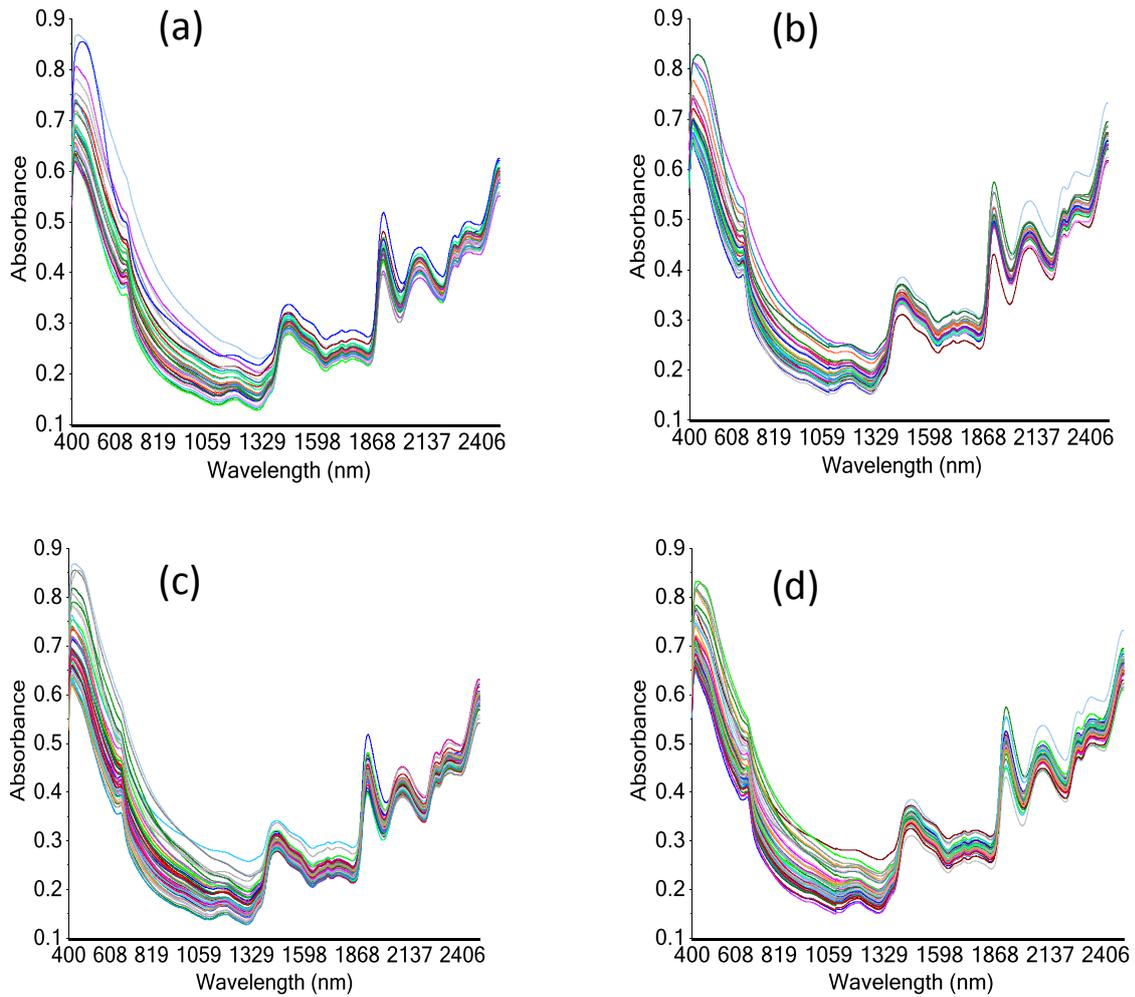


Figure C-14: The DU and DW bagasse spectra. (a) DU Spectra of the 30 samples analysed at UL; (b) DW spectra of these samples; (c) DU spectra of all 47 samples; (d) DW spectra of all 47 samples.

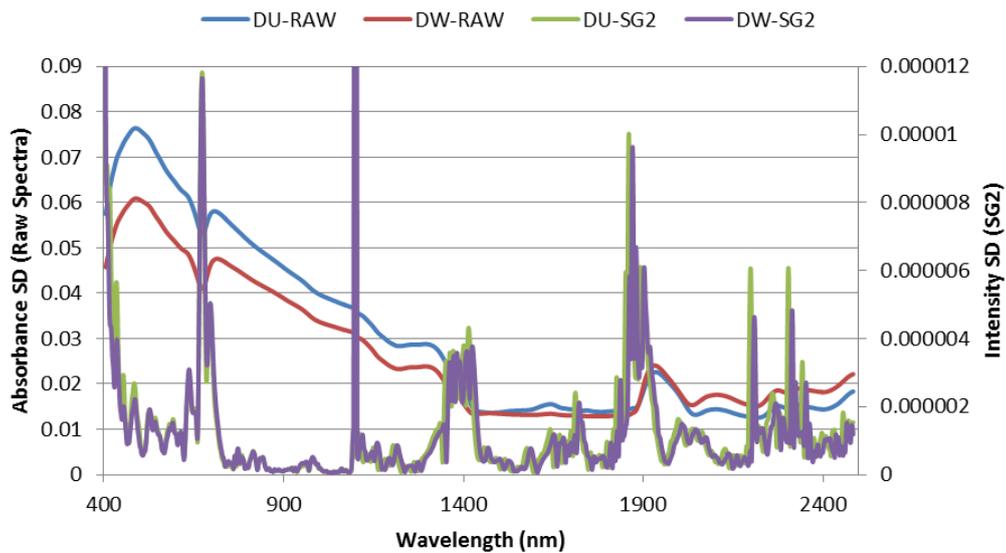


Figure C-15: A plot of standard deviation spectra, over all 47 samples, for the DU and DW datasets and the raw spectra and the second derivative (SG2,2,10,10) spectra.

Table C-11: Explained X variance (in %) for the cross validation, using up to 20 PCs, for the FULL spectral region (400-2500 nm) or the 1100-2500 nm region (NIR) for all 47 samples (ALL) or the 30 analysed at UL (UL). The DU scans and DW scans are represented here.

	Explained X-Variance (%) with Models of Varying PCs																			
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
DU SCANS																				
FULL-UL	33.04	48.99	57.00	60.84	69.02	72.94	75.03	81.59	83.57	86.83	89.39	89.76	90.52	90.68	91.24	91.60	92.08	92.77	93.62	93.82
FULL-ALL	38.69	51.18	63.51	69.77	75.90	78.47	82.76	84.96	86.87	87.29	91.52	92.09	92.64	93.48	94.31	94.90	95.05	95.22	96.00	96.35
NIR-UL	7.49	54.82	82.80	91.91	93.59	94.54	95.19	95.59	96.04	96.42	96.94	97.04	97.40	97.57	97.59	97.70	97.82	97.87	97.92	97.94
NIR-ALL	32.49	59.50	81.91	90.37	93.44	95.26	95.59	95.68	96.76	96.90	97.60	97.70	97.94	98.11	98.17	98.33	98.34	98.35	98.42	98.44
DW SCANS																				
FULL-UL	33.43	49.39	53.75	69.18	70.97	74.63	77.42	82.91	85.67	88.41	90.00	90.40	90.79	91.07	91.99	92.15	93.38	93.45	93.69	93.89
FULL-ALL	40.51	53.05	67.32	75.77	78.85	81.56	81.80	85.93	90.08	91.36	91.65	92.57	93.28	93.86	94.08	94.35	95.07	95.75	95.87	96.20
NIR-UL	24.72	58.65	82.80	90.21	92.71	95.07	95.22	95.47	96.31	96.84	96.96	97.15	97.23	97.37	97.44	97.53	97.57	97.61	97.67	97.70
NIR-ALL	31.95	59.90	84.12	90.35	93.15	95.67	95.94	96.30	96.44	97.13	97.71	97.74	97.97	97.98	98.17	98.26	98.31	98.31	98.37	98.42

Table C-12: Explained X variance (in %), for the WU scans, for the cross validation, using up to 20 PCs, for the FULL spectral region (400-2500 nm) or the 1100-2500 nm region (NIR) for all 47 samples (ALL) or the 30 analysed at UL (UL).

	Explained X-Variance (%) with Models of Varying PCs																			
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
FULL-UL	52.45	75.43	79.30	83.03	85.06	85.27	87.29	88.75	89.81	91.14	91.25	92.48	92.61	93.78	93.84	94.12	95.26	95.70	95.77	95.84
FULL-ALL	47.61	74.86	80.15	84.03	86.26	88.15	89.75	90.36	91.60	92.29	93.28	94.18	94.76	94.93	95.01	95.96	96.10	96.49	96.86	97.06
NIR-UL	85.41	91.16	93.19	95.24	96.66	96.96	97.54	97.85	98.04	98.15	98.21	98.31	98.34	98.38	98.39	98.42	98.44	98.50	98.51	98.59
NIR-ALL	79.25	84.28	93.24	95.46	96.38	97.28	97.97	98.11	98.27	98.38	98.40	98.43	98.50	98.57	98.62	98.66	98.68	98.75	98.79	98.82

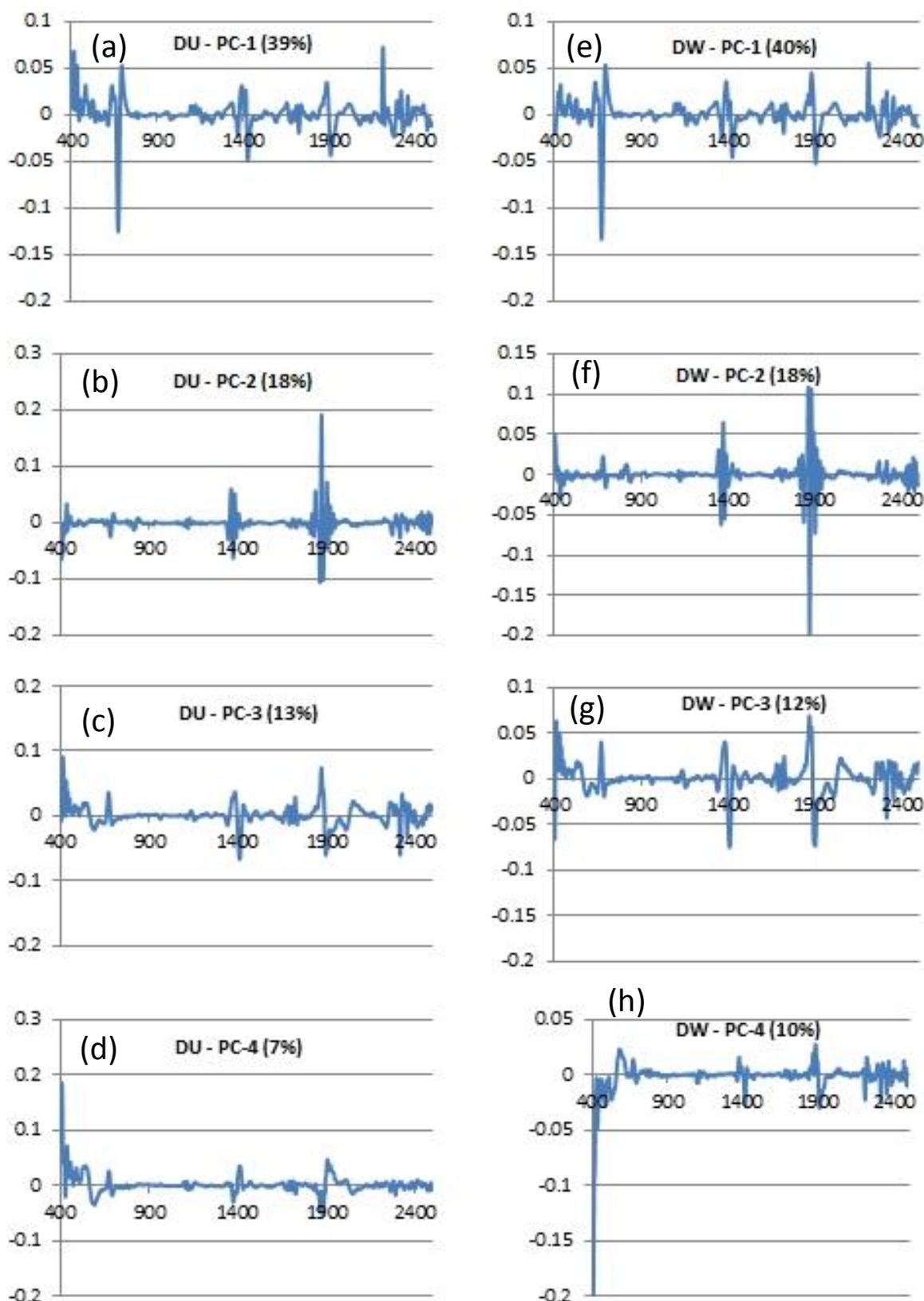


Figure C-16: Loadings plots for the DU and DW models. (a)-(d) PC1-4 loading plots for a PCA over the 400-2500 nm region using the DU scans of the 30 UL samples; (e)-(h) PC1-4 loading plots for a PCA over the same region using the same samples and the DW scans. Wavelength (nm) on x-axis, loading value on y-axis. Numbers in brackets indicate X-variance in calibration set explained by that PC.

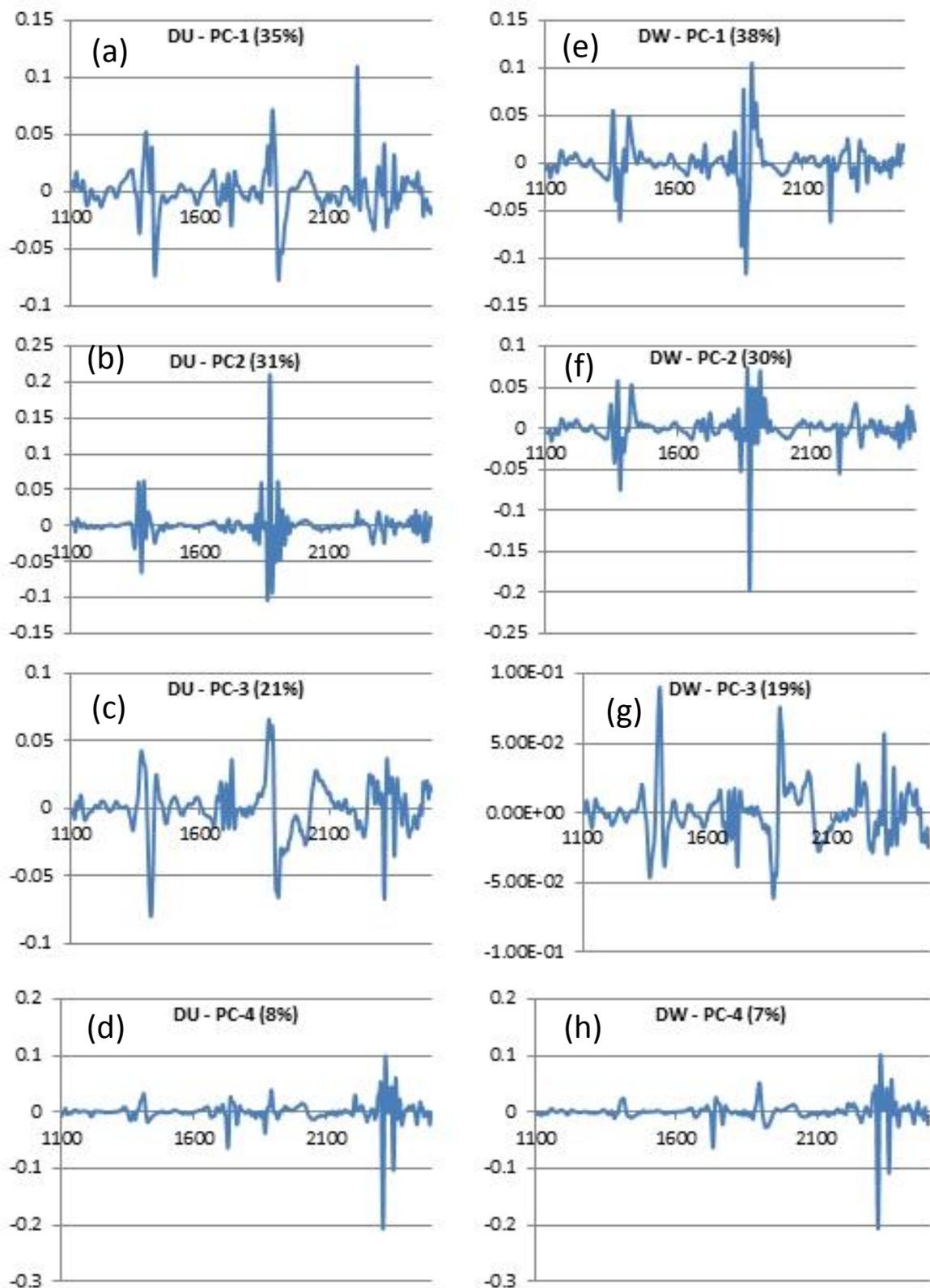


Figure C-17: Loadings plots for the DU and DW NIR models. (a)-(d) PC1-4 loading plots for a PCA over the 1100-2500 nm spectral region using the DU scans of the 30 UL samples; (e)-(h) PC1-4 loading plots for a PCA over the same spectral region using the same samples and the DW scans. Wavelength (nm) on x-axis, loading value on y-axis. Numbers in brackets indicate X-variance in calibration set explained by that PC.

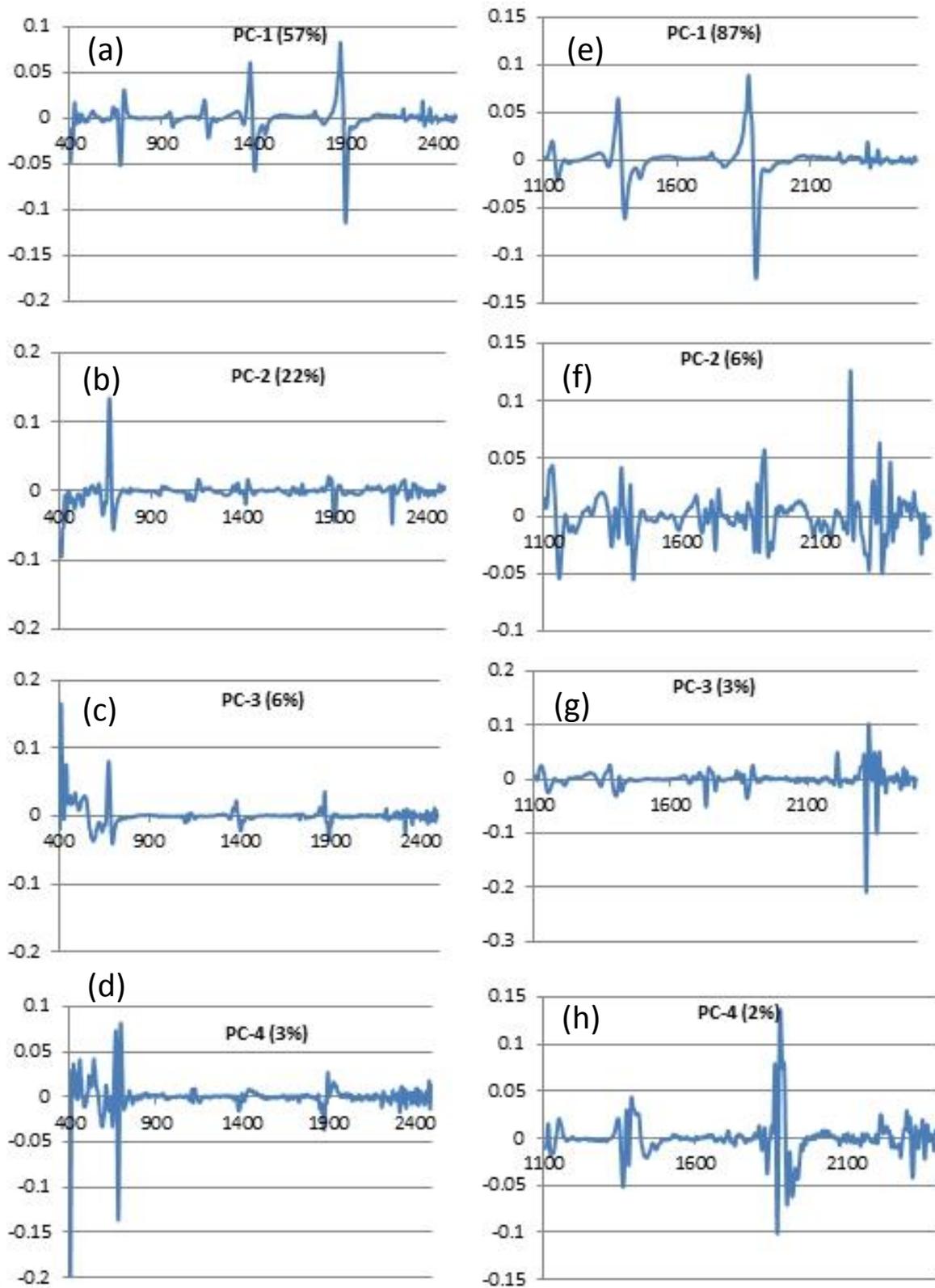


Figure C-18: Loadings plots for the WU models. (a)-(d) PC1-4 loading plots for a PCA over the 400-2500 nm spectral region using the WU scans of the 30 UL samples; (e)-(h) PC1-4 loading plots for a PCA over the 1100-2500 nm spectral region. Wavelength (nm) on x-axis, loading value on y-axis. Numbers in brackets indicate X-variance in calibration set explained by that PC.

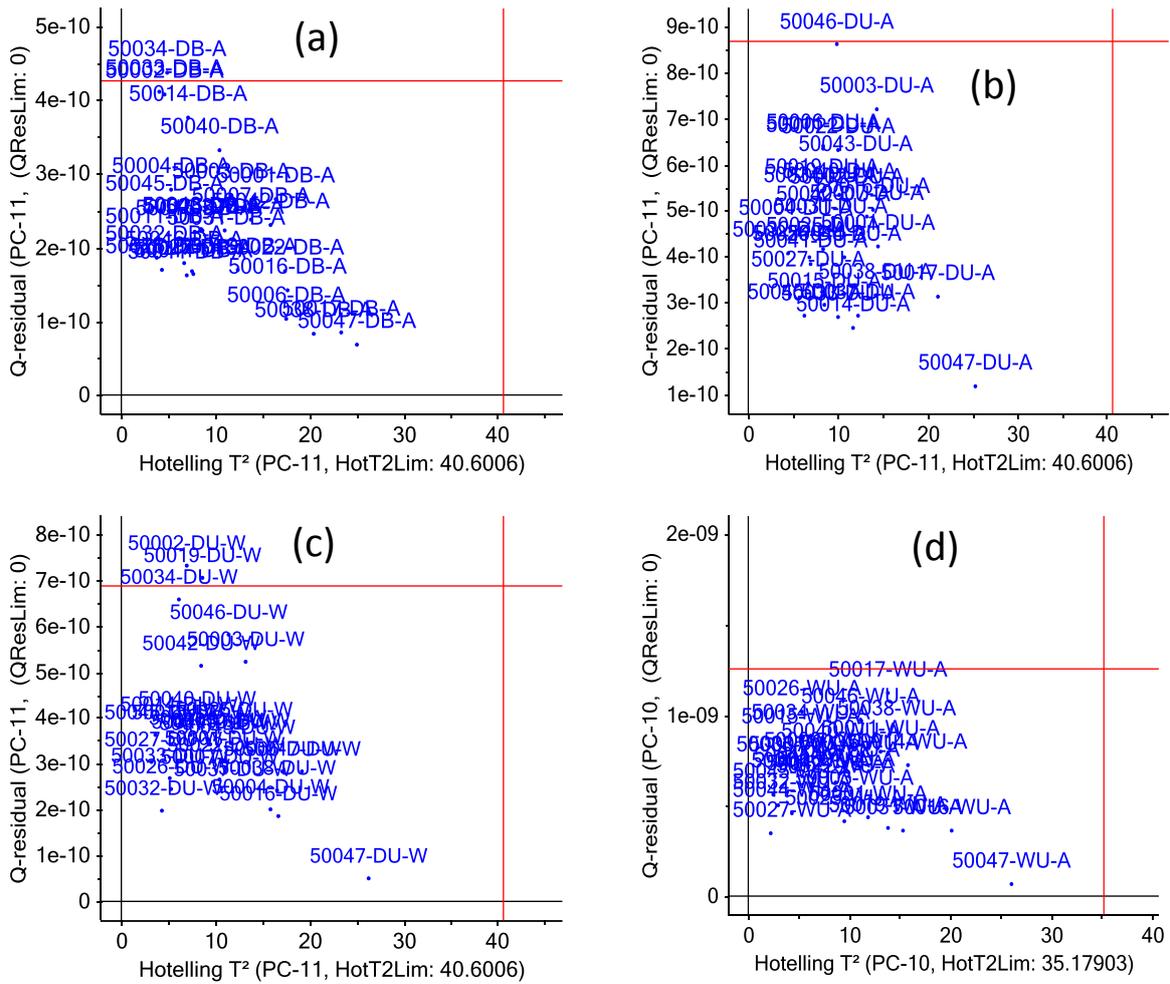


Figure C-19: (a) to (d) represent the influence plots for the DB, DU, DW, and WU data sets, respectively, involving PCAs over the 400-2500 nm spectral region.

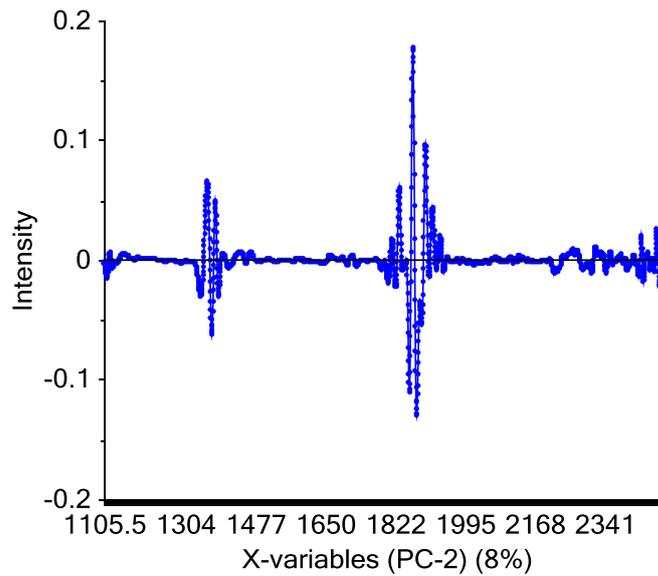


Figure C-20: A loadings plot for PC2 of the 47-sample 1100-2500 nm model for the WU data set.

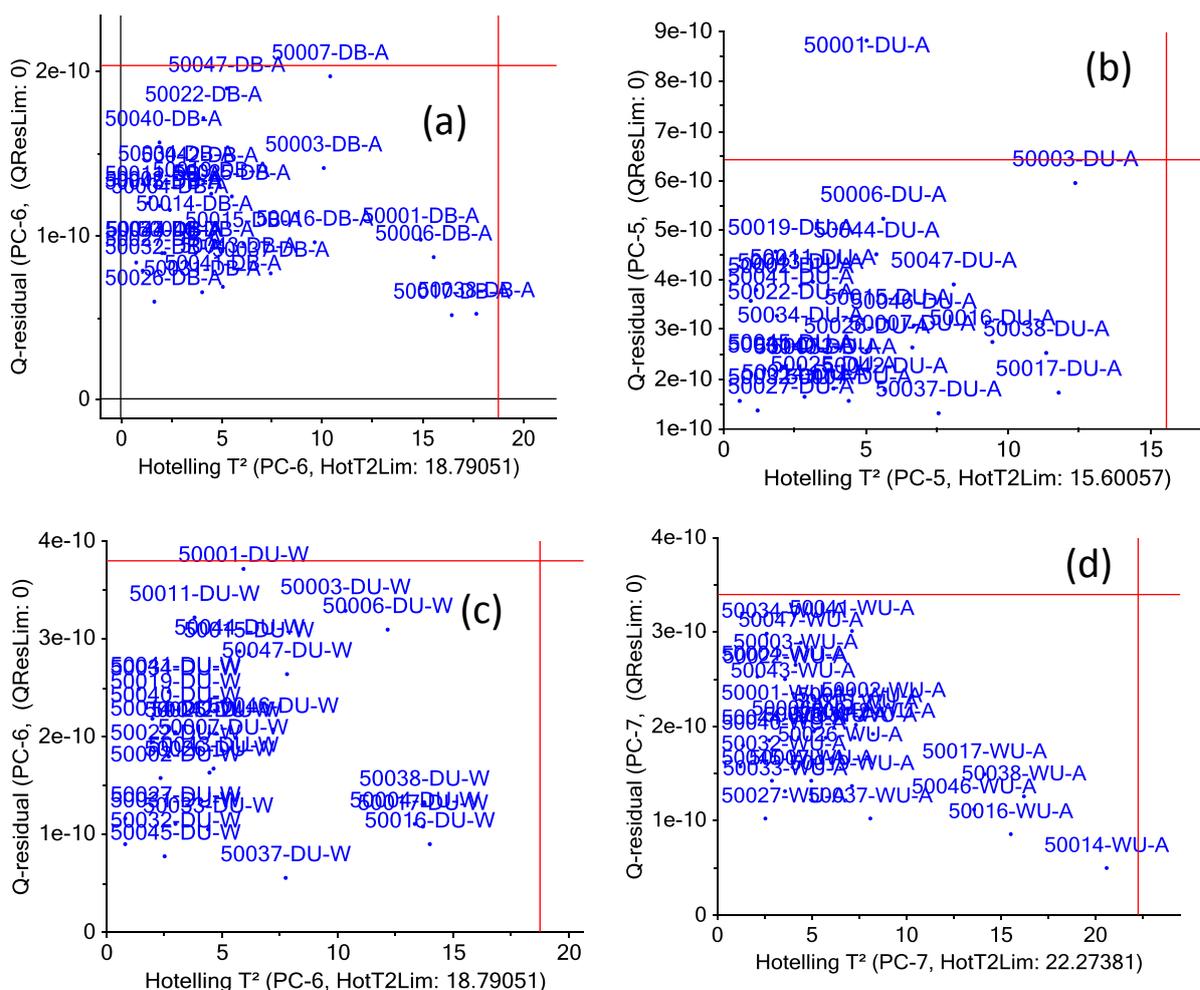


Figure C-21: (a) to (d) represent the influence plots for the DB, DU, DW, and WU data sets, respectively, involving PCAs over the 1100-2500 nm spectral region.

Table C-13: Summary statistics for the moisture contents (% wet basis) for the different calibration sets.

Samples	#	Av (%)	SD (%)	Max (%)	Min (%)	Range (%)
DS-E	21	7.29	0.30	7.83	6.72	1.11
DS-E Dishes	25	7.60	1.07	9.33	6.08	3.25
E-H	20	6.94	0.70	8.48	6.20	2.28

Table C-14: Regression statistics for the chosen calibrations for moisture content (% wet basis) in the extractives/hydrolysis analysis. Pre. = spectral pretreatment.

Samples	Pre.	λ (nm)	# Factors	R^2_{cal}	RMSEC (%)	RMSECV (%)	RPD_{CV}	RER_{CV}
DS-E	MSC	1100-2500	2	0.797	0.142	0.140	2.13	7.92
DS-E Dishes	MSC	1100-2500	2	0.975	0.176	0.186	5.76	17.47
E-H	MSC	1100-2500	3	0.883	0.260	0.297	2.35	7.69

Table C-15: Regression statistics for the PLS calibration for moisture content (% wet basis) of bagasse DS samples prior to the removal of their extractives.

Pretreatment	MSC	SNVDT (2)	SNV	D-1,4,8,1	D-1,4,8,1	D-1,4,16,1	D-1,4,16,1	D-1,4,32,1	D-1,4,32,1	D-2,4,8,1	D-2,4,8,1	D-2,4,16,1	D-2,4,16,1	D-2,4,32,1	D-2,4,32,1	MSC + D-1,4,16,1
Spectral Region (x 10 ³ nm)	1.1-2.5	1.1-2.5	1.1-2.5	0.4-2.5	1.1-2.5	0.4-2.5	1.1-2.5	0.4-2.5	1.1-2.5	0.4-2.5	1.1-2.5	0.4-2.5	1.1-2.5	0.4-2.5	1.1-2.5	1.1-2.5
Factors	2	2	2	4	3	4	3	4	3	4	4	4	4	4	4	3
R^2_{calib}	0.797	0.782	0.797	0.859	0.843	0.858	0.842	0.859	0.840	0.867	0.873	0.867	0.855	0.851	0.844	0.851
RMSEC (%)	0.142	0.147	0.141	0.125	0.128	0.125	0.129	0.125	0.129	0.121	0.119	0.121	0.127	0.129	0.132	0.125
RMSECV (%)	0.140	0.162	0.153	0.177	0.167	0.177	0.168	0.177	0.165	0.191	0.199	0.191	0.184	0.179	0.189	0.163
RPD (CV)	2.13	1.84	1.95	1.69	1.78	1.69	1.78	1.69	1.81	1.56	1.50	1.56	1.62	1.66	1.57	1.83
RER (CV)	7.92	6.84	7.26	6.29	6.64	6.28	6.63	6.28	6.73	5.82	5.58	5.82	6.04	6.20	5.86	6.82

Table C-16: Regression statistics for the moisture content (% wet basis) prediction of different data sets using various PLS moisture content models.

Model	DS-E	DS-E	DS-E Dishes	DS-E Dishes	E-H	E-H
Predicting	DS-E Dishes	E-H	DS-E	E-H	DS-E	DS-E Dishes
R^2_{Pred}	0.975	0.783	0.808	0.808	0.664	0.967
Slope	0.810	0.662	0.943	0.806	0.937	1.055
Intercept	1.508	2.695	0.283	1.560	0.105	-0.590
RMSEP (%)	0.250	0.483	0.190	0.367	0.402	0.267
Bias (%)	0.060	0.352	-0.134	0.215	-0.352	-0.176
SEP (%)	0.247	0.338	0.138	0.305	0.200	0.205
RPD	4.326	2.059	2.162	2.282	1.493	4.920
RER	13.131	6.747	8.053	7.477	5.560	15.198

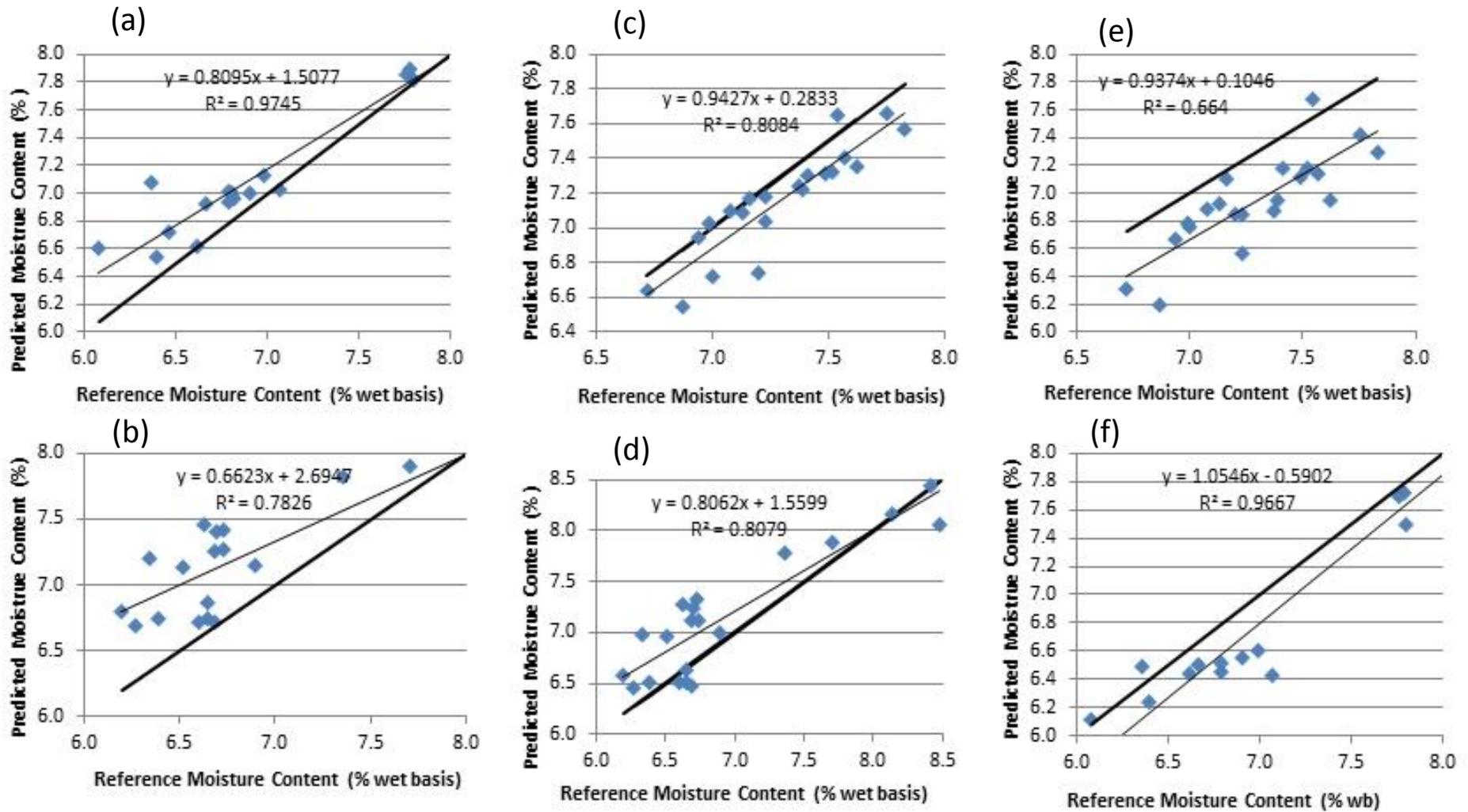


Figure C-22: The moisture content models (E-H, DS-E, DS-E Dishes) predicted the moisture contents of the samples in the other models: (a) DS-E predicting DS-E Dishes; (b) DS-E predicting E-H; (c) DS-E Dishes predicting DS-E; (d) DS-E Dishes predicting E-H; (e) E-H predicting DS-E; (f) E-H predicting DS-E Dishes.

Table C-17: PLSR statistics for models based on the DS scans and the GLU_EF_SRS values (sample 50019 excluded). Refer to Appendix A for descriptions of abbreviations.

Number	*1	*2	*3	*4	*5	*6	*7	*8	*9	*10	*11	*12	*13	*14	*15	*16	*17	*18	*19	*20	*21	*22	*23	
Dataset	DS	DS	DS	DS	DS	DS	DS	DS	DS	DS	DS	DS	DS	DS	DS	DS	DS	DS	DS	DS	DS	DS	DS	
Pre. (1)	NONE	NONE	SNV	SNV	MSC	MSC	MSC	SG	SG	SG	SG	SG	SG	SG	SG	SNVDT	SNVDT	SNVDT	EMSC	EMSC	EMSC	SNVDT	SG	
Specific (1)			0.4-2.5	1.1-2.5	F,0.4-2.5	F,0.4-2.5	F,1.1-2.5	1,2,3,3	1,2,3,3	1,3,7,7	1,3,12,12	2,3,7,7	3,3,7,7	1,3,7,7	3,3,7,7	1,1-2.5,2	1.1-2.5,3	1.1-2.5,4	F,1.1-2.5	A,1.1-2.5	B,1.1-2.5	1.1-2.5,2	1,3,7,7	
Pre. (2)																						SG	SNVDT	
Specific (2)																							1,3,7,7	1.1-2.5,2
PLS-λ 10 ³ nm	0.4-2.5	1.1-2.5	0.4-2.5	1.1-2.5	0.4-2.5	1.1-2.5	1.1-2.5	0.4-2.5	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.2	1.1-2.2	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.2	1.1-2.2	
Sam. Excl.	50019	50019	50019	50019	50019	50019	50019	50019	50019	50019	50019	50019	50019	50019	50019	50019	50019	50019	50019	50019	50019	50019	50019	50019
F-Wold's	1	5	1	4	1	1	4	3	6	6	4	5	2	6	3	4	4	4	4	5	4	8	6	
F-Wold 0.95	1	5	1	4	1	1	4	3	4	4	4	5	2	6	2	4	4	4	1	5	4	5	6	
F-Wold 0.9	1	1	1	4	1	1	4	3	4	4	4	4	1	4	2	4	4	4	1	5	4	5	4	
F-F Test	1	5	1	4	1	1	4	3	4	4	4	5	2	6	2	4	4	4	1	5	4	5	6	
F-Haaland's	7	5	7	3	6	4	3	12	6	6	6	3	1	12	2	4	3	3	4	4	10	11	11	
F.-Min Press	11	5	10	7	9	5	4	16	9	9	5	2	19	3	4	4	4	4	4	5	11	14	13	
F.-UNSCR.	11	5	10	4	6	4	4	3	6	6	9	5	2	6	2	4	4	4	4	5	4	5	6	
R ² _{calib}	0.9336	0.9063	0.9496	0.9079	0.9385	0.9094	0.9078	0.9964	0.9567	0.9573	0.9545	0.8760	0.7742	0.9979	0.8496	0.9194	0.8721	0.9072	0.9238	0.9030	0.9847	0.9972	0.9974	
Offset _{calib}	2.8215	3.9800	2.1388	3.9121	2.6096	3.8452	3.9139	0.1507	1.8383	1.8120	1.9338	5.2667	9.5859	0.0892	6.3864	3.4211	5.4293	3.9397	3.2340	4.1193	0.6502	0.1197	0.1100	
RMSEC (%)	0.3983	0.4731	0.3468	0.4690	0.3831	0.4650	0.4691	0.0921	0.3215	0.3192	0.3298	0.5442	0.7342	0.0709	0.5993	0.4386	0.5526	0.4707	0.4265	0.4813	0.1913	0.0820	0.0786	
R ² _{CV}	0.6966	0.8357	0.7301	0.8512	0.7069	0.8194	0.8531	0.6679	0.8073	0.8080	0.8069	0.6711	0.5236	0.8644	0.5614	0.8839	0.8186	0.8567	0.8644	0.8399	0.8699	0.8595	0.8757	
Slope _{CV}	0.8590	0.8315	0.8452	0.8632	0.8202	0.8011	0.8651	0.8494	0.8376	0.8369	0.8394	0.6656	0.4248	0.9055	0.6165	0.8761	0.8147	0.8711	0.8658	0.8282	0.9267	0.9121	0.9340	
Offset _{CV}	5.8934	7.1669	6.4625	5.7966	7.5495	8.4394	5.7134	6.3474	6.8826	6.9184	6.8107	14.1894	24.4597	4.0724	16.2734	5.2680	7.8510	5.4603	5.6879	7.3082	3.1195	3.8089	2.8390	
RMSECV (%)	0.8839	0.6504	0.8337	0.6191	0.8688	0.6820	0.6151	0.9247	0.7044	0.7032	0.7052	0.9202	1.1076	0.5910	1.0627	0.5468	0.6834	0.6075	0.5908	0.6421	0.5788	0.6015	0.5657	
BIAS _{CV}	-0.0930	0.0117	-0.1116	-0.0130	-0.0872	-0.0065	-0.0142	-0.0480	-0.0112	-0.0092	-0.0092	-0.0110	0.0352	0.0603	-0.0128	0.0052	-0.0177	-0.0143	-0.0122	0.0126	0.0059	0.0766	0.0384	
SECV (%)	0.8957	0.6627	0.8419	0.6307	0.8809	0.6950	0.6267	0.9411	0.7177	0.7166	0.7186	0.9377	1.1281	0.5991	1.0829	0.5572	0.6962	0.6189	0.6020	0.6542	0.5898	0.6079	0.5751	
RPD _{CV}	1.7581	2.3763	1.8703	2.4966	1.7877	2.2658	2.5128	1.6733	2.1941	2.1976	2.1914	1.6793	1.3958	2.6284	1.4542	2.8261	2.2620	2.5445	2.6160	2.4071	2.6699	2.5903	2.7380	
RER _{CV}	7.1960	9.7267	7.6556	10.2191	7.3173	9.2742	10.2852	6.8491	8.9806	8.9951	8.9696	6.8737	5.7134	10.7585	5.9523	11.5678	9.2586	10.4150	10.7076	9.8528	10.9284	10.6024	11.2072	
RMSECV _{MP}	0.8136	0.6504	0.7525	0.5482	0.8544	0.6668	0.5518	0.8497	0.6500	0.6543	0.6384	0.8338	1.0678	0.5391	1.0574	0.5468	0.6453	0.5423	0.5908	0.5828	0.5181	0.5351	0.5075	

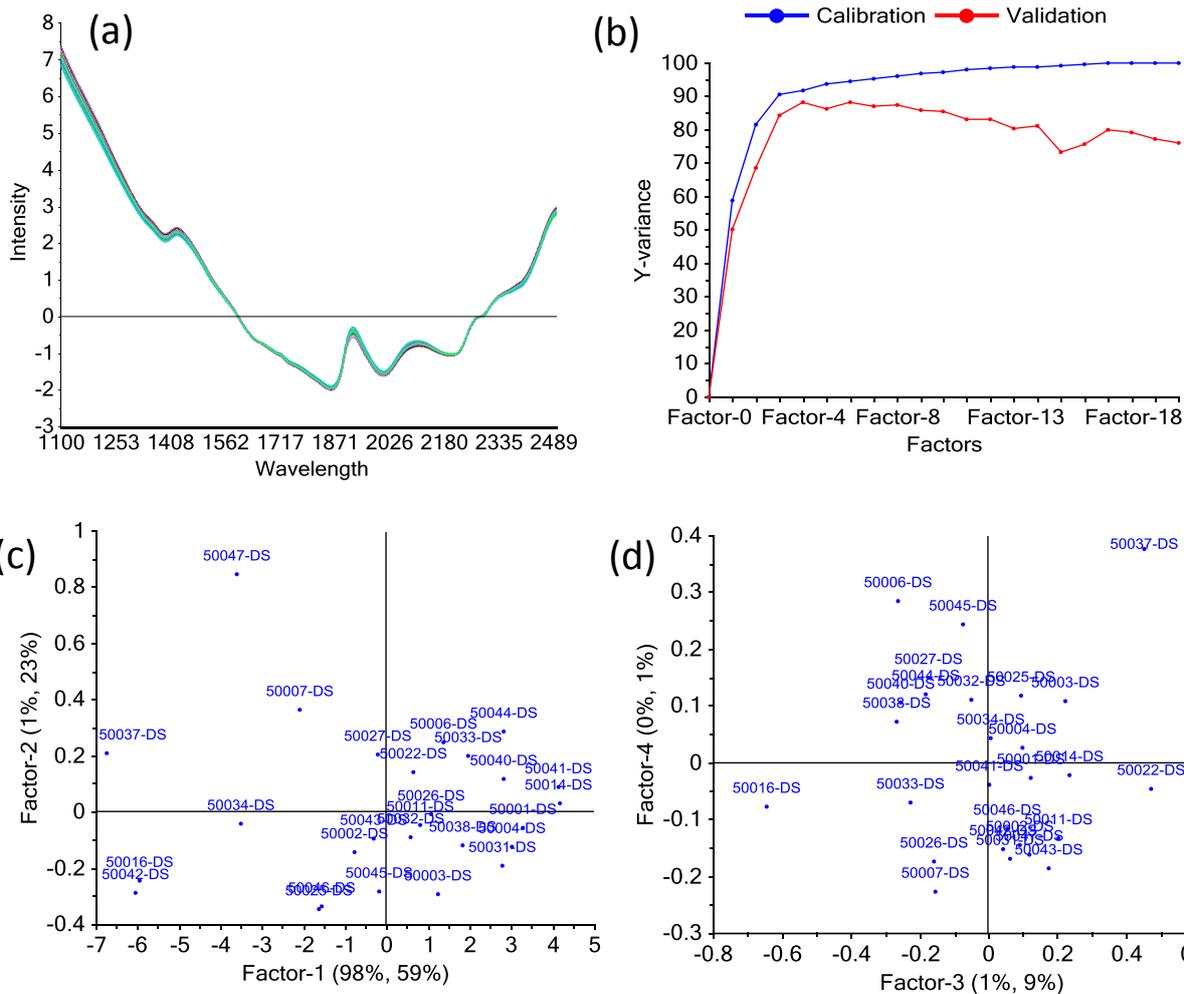


Figure C-24: Plots for GLU_EF_SRS DS PLSR model *16 in Table C-17. (a) The transformed 27 DS spectra after the SNVDT transform was applied; (b) an explained y-variance plot, with differing numbers of PLS factors; (c) a F1 vs. F2 scores; (d) a F3 vs. F4 scores plot.

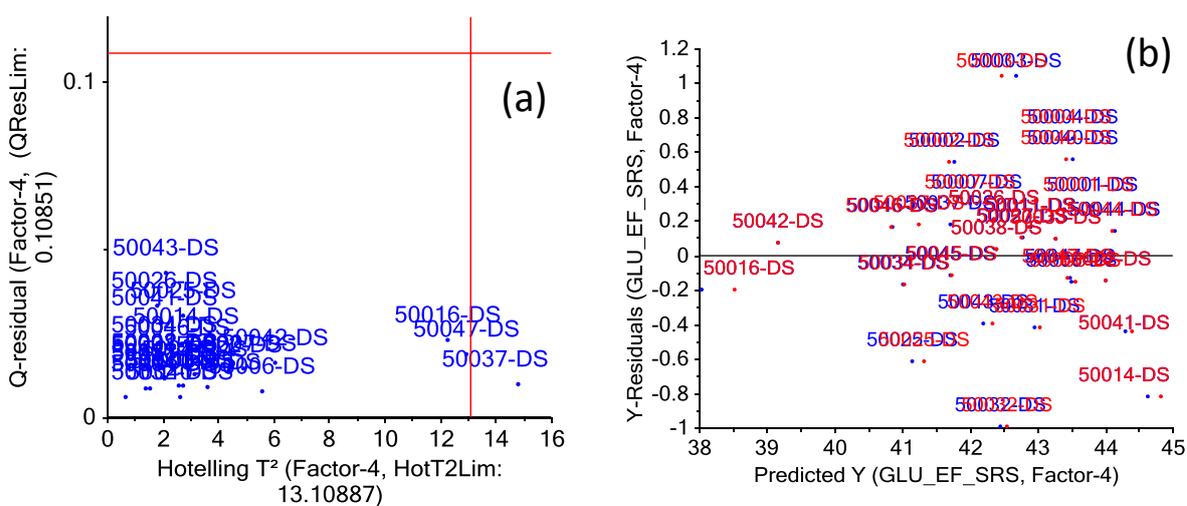


Figure C-25: Further plots for GLU_EF_SRS DS PLSR model *16 in Table C-17. (a) An influence plot; (b) a predicted y vs. y-residuals plot for this regression.

Table C-18: PLSR statistics for models based on the DS scans and the XYL_EF_SRS, ARA_EF_SRS, and RHA_EF_SRS contents. Refer to Appendix A for descriptions of terms.

Number	*1	*2	*3	*4	*5	*6	*7	*8	*9	*10	*11	*12	*13	*14	*15	*16	*17
Dataset	DS																
Constituent	XYL_EF_SRS	XYL_EF_SRS	XYL_EF_SRS	ARA_EF_SRS	RHA_EF_SRS	RHA_EF_SRS	RHA_EF_SRS	RHA_EF_SRS	RHA_EF_SRS	RHA_EF_SRS							
Pre. (1)	NONE	SG	EMSC	MSC	SG	NONE	SNV	SNVDT	EMSC	EMSC	EMSC	NONE	NONE	SG1,3,7,7	SG	SG	MSC
Specific (1)		1,3,7,7	A,1.1-2.5	F,1.1-2.5	1,3,7,7		1.1-2.5	1.1-2.5,2	A,1.1-2.5	B,1.1-2.5	B,1.1-2.5			1,3,14,14	2,3,10,10	F,1.1-2.5	
Pre. (2)		SNV									SG						
Specific (2)										1,3,7,7							
PLS-λ 10 ³ nm	1.1-2.5	1.1-2.2	1.1-2.5	1.1-2.5	1.1-2.2	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.2	0.4-2.5	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5	
Sam. Excl.	50019	50019	50019	50019	50019	50019	50019	50019	50019	50019	50019	50019	50019	50019	50019	50019	
F-Wold's	2	1	1	2	4	4	2	2	1	2	5	1	1	2	2	1	
F-Wold 0.95	2	1	1	2	1	2	2	2	1	2	1	1	1	1	1	1	
F-Wold 0.9	2	1	1	2	1	2	2	2	1	2	1	1	1	1	1	1	
F-F Test	2	1	1	2	1	4	2	2	1	2	5	1	1	1	1	1	
F-Haaland's	1	6	13	6	6	18	6	6	4	4	5	1	8	1	7	1	
F.-Min Press	2	20	14	7	8	20	7	7	6	6	9	1	10	12	13	1	
F.-UNSCR.	2	12	14	7	7	8	6	7	5	6	7	1	10	8	13	1	
R^2_{calib}	0.2427	0.8788	0.9673	0.8217	0.8429	0.9956	0.8177	0.8182	0.6790	0.7428	0.8269	0.2327	0.8630	0.0888	0.8741	0.1277	
$Offset_{calib}$	17.9826	2.8788	0.7776	0.4286	0.3776	0.0093	0.4381	0.4369	0.7717	0.6184	0.4162	0.0874	0.0156	0.1038	0.0143		
RMSEC (%)	0.4630	0.1853	0.0963	0.0464	0.0435	0.0073	0.0469	0.0468	0.0622	0.0557	0.0457	0.0082	0.0035	0.0089	0.0033		
R^2_{cv}	-0.0583	0.3978	0.3586	0.4945	0.4734	0.6081	0.4907	0.4910	0.4116	0.4302	0.4885	0.1650	0.4207	0.0203	0.1705		
$Slope_{cv}$	0.1119	0.5314	0.5820	0.6454	0.6842	0.8981	0.6320	0.6281	0.5390	0.5534	0.6824	0.1735	0.5691	0.0236	0.4432		
$Offset_{cv}$	21.0630	11.1054	9.9558	0.8460	0.7575	0.2300	0.8782	0.8872	1.1087	1.0717	0.7591	0.0941	0.0496	0.1112	0.0641		
RMSECV (%)	0.5684	0.4288	0.4425	0.0811	0.0827	0.0714	0.0814	0.0813	0.0874	0.0861	0.0815	0.0088	0.0074	0.0096	0.0088		
$BIAS_{cv}$	-0.0254	-0.0213	0.0306	-0.0065	-0.0016	-0.0149	-0.0064	-0.0067	0.0003	-0.0020	-0.0043	0.0000	0.0006	0.0000	0.0007		
SEC (%)	0.5787	0.4364	0.4499	0.0823	0.0843	0.0711	0.0827	0.0826	0.0891	0.0877	0.0830	0.0090	0.0075	0.0098	0.0090		
RPD_{cv}	0.9370	1.2425	1.2052	1.3588	1.3273	1.5731	1.3535	1.3544	1.2554	1.2760	1.3483	1.0538	1.2687	0.9729	1.0603		
RER_{cv}	4.0011	5.3055	5.1465	5.3382	5.2142	6.1797	5.3173	5.3209	4.9319	5.0128	5.2967	4.0512	4.8774	3.7402	4.0760		
$RMSECV_{MP}$	0.5380	0.3891	0.4077	0.0801	0.0760	0.0668	0.0811	0.0795	0.0825	0.0771	0.0757	0.0088	0.0070	0.0086	0.0077		

Table C-19: PLSR statistics for models based on the DS scans and the MAN_EF_SRS and GAL_EF_SRS values. Refer to Appendix A for descriptions of abbreviations.

Number	*1	*2	*3	*4	*5	*6	*7	*8	*9	*10	*11	*12	*13	*14	*15	*16	*17	*18	*19	*20	*21
Dataset	DS																				
Constituent	MAN_EF_SRS	GAL_EF_SRS																			
Pre. (1)	NONE	NONE	MSC	SNVDT	SG	SG	SG	NONE	NONE	MSC	SNV	SNVDT	SG	SG	SG	MSC	MSC	MSC	SG	SG	SG
Specific (1)			F,1.1-2.5	2,1.1-2.5	1,3,14,14	2,3,10,10	1,3,14,14			F,1.1-2.5	1.1-2.5	2,1.1-2.5	1,3,7,7	1,3,14,14	2,3,10,10	F,1.1-2.5	F,1.1-2.5	F,0,4-2.5	1,3,7,7	1,3,14,14	2,3,10,10
Pre. (2)																	SG	SG			
Specific (2)																1,3,14,14	1,3,14,14				
PLS-λ 10 ³ nm	0.4-2.5	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5	0.4-2.5	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5	0.4-2.5	0.4-2.5	0.4-2.5	0.4-2.5	0.4-2.5
Sam. Excl.	50019	50019	50019	50019	50019	50019	50019	50019	50019	50019	50019	50019	50019	50019	50019	50019	50019	50019	50019	50019	50019
F-Wold's	2	4	2	2	1	3	2	2	4	2	2	2	5	5	6	3	7	1	6	6	6
F-Wold 0.95	2	3	2	2	1	3	2	1	4	2	2	2	4	4	6	3	6	1	6	3	3
F-Wold 0.9	1	3	1	1	1	3	1	1	4	2	2	2	4	4	4	3	6	1	2	3	3
F-F Test	2	3	2	2	1	3	2	1	4	2	2	2	4	4	6	3	6	1	6	3	3
F-Haaland's	1	3	1	1	4	2	5	7	3	2	2	5	4	4	4	3	6	6	6	5	5
F.-Min Press	8	6	5	5	5	3	6	7	7	6	6	5	7	7	6	5	7	6	6	15	6
F.-UNSCR.	8	6	5	5	5	3	6	7	7	6	6	5	4	4	6	5	6	6	6	6	6
R ² _{calib}	0.0695	0.4180	0.1168	0.1067	0.7389	0.5768	0.8596	0.9035	0.7280	0.6478	0.6483	0.8187	0.8748	0.8721	0.8889	0.7475	0.8885	0.8763	0.8988	0.8647	0.9139
Offset _{calib}	0.1419	0.0887	0.1347	0.1362	0.0398	0.0645	0.0216	0.0859	0.2421	0.3136	0.3131	0.1614	0.1115	0.1139	0.0989	0.2248	0.0993	0.1101	0.0901	0.1205	0.0766
RMSEC (%)	0.0225	0.0178	0.0219	0.0220	0.0119	0.0152	0.0083	0.0278	0.0468	0.0532	0.0532	0.0382	0.0317	0.0321	0.0299	0.0451	0.0299	0.0315	0.0285	0.0330	0.0263
R ² _{CV}	0.0126	0.1527	0.0565	0.0501	0.3910	0.2639	0.4691	0.7883	0.3631	0.5694	0.5646	0.6771	0.7234	0.7218	0.6484	0.6132	0.6528	0.6792	0.6448	0.5443	0.6100
Slope _{CV}	0.0088	0.2654	0.0578	0.0483	0.4573	0.3341	0.5507	0.8429	0.5373	0.5720	0.5698	0.7398	0.7687	0.7693	0.7001	0.6654	0.6831	0.7036	0.7149	0.6377	0.5977
Offset _{CV}	0.1513	0.1113	0.1440	0.1454	0.0823	0.1016	0.0682	0.1419	0.4027	0.3831	0.3850	0.2314	0.2032	0.2026	0.2662	0.2972	0.2859	0.2660	0.2570	0.3268	0.3585
RMSECV (%)	0.0240	0.0223	0.0235	0.0236	0.0189	0.0208	0.0169	0.0428	0.0743	0.0611	0.0614	0.0529	0.0490	0.0491	0.0552	0.0579	0.0549	0.0527	0.0555	0.0629	0.0581
BIAS _{CV}	0.0002	-0.0007	0.0003	0.0003	-0.0004	0.0001	-0.0011	0.0020	-0.0092	0.0021	0.0019	-0.0003	-0.0028	-0.0028	-0.0008	-0.0006	0.0037	0.0021	0.0032	0.0042	0.0003
SECV (%)	0.0245	0.0227	0.0239	0.0240	0.0192	0.0212	0.0172	0.0436	0.0751	0.0622	0.0626	0.0539	0.0498	0.0500	0.0563	0.0590	0.0558	0.0537	0.0565	0.0639	0.0593
RPD _{CV}	0.9691	1.0467	0.9915	0.9881	1.2343	1.1224	1.3147	2.0952	1.2161	1.4683	1.4601	1.6947	1.8340	1.8285	1.6242	1.5484	1.6380	1.7016	1.6184	1.4297	1.5420
RER _{CV}	3.7171	4.0147	3.8028	3.7900	4.7342	4.3048	4.9827	7.0820	4.1104	4.9631	4.9354	5.7282	6.1992	6.1807	5.4900	5.2338	5.5366	5.7517	5.4704	4.8326	5.2123
RMSECV _{MP}	0.0219	0.0206	0.0212	0.0215	0.0183	0.0189	0.0157	0.0428	0.0675	0.0545	0.0558	0.0529	0.0476	0.0481	0.0514	0.0539	0.0542	0.0527	0.0555	0.0559	0.0547

Table C-20: PLSR statistics for models based on the DS scans and the ash contents. Refer to Appendix A for descriptions of abbreviations.

Number	*1	*2	*3	*4	*5	*6	*7	*8	*9	*10	*11	*12	*13	*14	*15	*16	*17
Dataset	DS	DS															
Pre. (1)	NONE	NONE	NONE	NONE	MSC	MSC	MSC	SG	SG	SG	SG	SG	MSC	MSC	MSC	MSC	MSC
Specific (1)					F,1.1-2.5	F,1.1-2.5	F,1.1-2.5	1,3,7,7	2,3,10,10	1,3,14,14	1,3,14,14	1,3,14,14	F,0.4-2.5	F,0.4-2.5	F,1.1-2.5	F,1.1-2.5	F,1.1-2.5
Pre. (2)														SG		SG	SG
Specific (2)														1,3,14,14		1,3,14,14	1,3,14,14
PLS-λ 10 ³ nm	0.4-2.5	1.1-2.5	0.4-1.1	0.4-0.75	1.1-2.5	1.1-2.5	1.1-2.5	0.4-2.5	0.4-2.5	0.4-2.5	0.4-2.5	0.4-0.75	0.4-2.5	0.4-2.5	0.4-2.5	0.4-2.5	0.4-2.5
Sam. Excl.						50047	50047								50007		50007
F-Wold's	3	3	2	3	1	6	3	4	2	4	7	3	2	6	2	4	2
F-Wold 0.95	3	3	2	3	1	2	3	4	2	4	7	3	2	6	2	4	2
F-Wold 0.9	2	2	2	3	1	1	3	4	2	4	6	3	2	6	2	3	2
F-F Test	3	3	2	3	1	2	3	4	2	4	7	3	2	6	2	4	2
F-Haaland's	5	2	2	7	1	4	5	3	4	3	6	6	4	5	5	3	6
F.-Min Press	6	3	7	9	1	6	6	4	5	4	7	8	4	6	5	4	6
F.-UNSCR.	6	3	4	9	1	5	6	4	5	4	6	8	4	6	5	4	6
R^2_{calib}	0.8681	0.6926	0.7738	0.8940	0.7753	0.9493	0.9714	0.8515	0.8342	0.8581	0.9554	0.8711	0.8908	0.9105	0.9159	0.8942	0.9572
$Offset_{calib}$	0.5804	1.3529	0.9955	0.4679	0.9891	0.2243	0.1251	0.6535	0.7109	0.6244	0.1939	0.5675	0.4807	0.3894	0.3702	0.4658	0.1861
RMSEC (%)	0.7722	1.1789	1.0113	0.6924	1.0080	0.4865	0.3696	0.8194	0.3463	0.8009	0.4536	0.7636	0.7027	0.6428	0.6166	0.6918	0.4444
R^2_{cv}	0.7500	0.6753	0.5708	0.7112	0.7500	0.8634	0.9303	0.7587	0.4031	0.7737	0.8679	0.6717	0.7943	0.8158	0.8369	0.8219	0.8887
$Slope_{cv}$	0.8086	0.7379	0.6337	0.8157	0.7422	0.9084	0.9303	0.7457	0.4532	0.7423	0.8733	0.8296	0.8155	0.7998	0.8966	0.8049	0.8551
$Offset_{cv}$	0.8804	1.1798	1.6219	0.8157	1.1203	0.3492	0.2763	1.0987	2.3067	1.1128	0.6152	0.7270	0.8306	0.7954	0.4813	0.8628	0.6298
RMSECV (%)	1.1027	1.2567	1.4448	1.1851	1.1027	0.8399	0.6284	1.0471	0.6814	1.0490	0.8821	1.2636	1.0003	1.0472	0.8906	0.9308	0.7685
$BIAS_{cv}$	0.0382	0.0261	0.0095	0.0046	-0.0146	-0.0563	-0.0285	-0.0207	-0.0386	-0.0216	0.0641	-0.0232	0.0186	-0.0753	0.0260	0.0043	-0.0004
SECV (%)	1.1222	1.2795	1.4712	1.2069	1.1229	0.8540	0.6402	1.0661	0.6928	1.0680	0.8966	1.2866	1.0185	1.0643	0.9066	0.9478	0.7831
RPD_{cv}	1.9296	1.6925	1.4719	1.7943	1.9286	2.5789	3.4816	2.0313	1.2501	2.0276	2.4421	1.6832	2.1263	2.0571	2.3887	2.2847	2.7959
RER_{cv}	7.1223	6.2472	5.4328	6.6228	7.1185	9.3598	12.4854	7.4975	4.1519	7.4841	8.9151	6.2126	7.8481	7.5098	8.8169	8.4328	10.2067
$RMSECV_{MP}$	1.0565	1.2219	1.3607	1.0779	1.1027	0.7377	0.5713	0.9748	0.6342	0.9897	0.8396	1.1563	1.0003	0.9729	0.8906	0.8924	0.7685

Table C-21: PLSR statistics for models based on the DS scans and the EIA and AIA_EF contents. Refer to Appendix A for descriptions of abbreviations.

Number	*1	*2	*3	*4	*5	*6	*7	*8	*9	*10	*11	*12	*13
Dataset	DS	DS	DS	DS	DS	DS	DS	DS	DS	DS	DS	DS	DS
Constituent	EIA	EIA	EIA	EIA	AIA_EF	AIA_EF	AIA_EF	AIA_EF	AIA_EF	AIA_EF	AIA_EF	AIA_EF	AIA_EF
Pre. (1)	NONE	NONE	MSC	SNVDT	NONE	NONE	MSC	MSC	SNVDT	SNVDT	SG	SG	SG
Specific (1)			F,1.1-2.5	2,1.1-2.5			F,1.1-2.5	F,1.1-2.5	2,1.1-2.5	2,1.1-2.5	1,3,14,14	2,3,20,20	2,3,20,20
Pre. (2)													
Specific (2)													
PLS-λ 10 ³ nm	0.4-2.5	1.1-2.5	1.1-2.5	1.1-2.5	0.4-2.5	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5	0.4-2.5	1.1-2.5	1.1-2.5
Sam. Excl.	50019 50046	50019 50046	50019 50046	50019 50046	50019	50019	50019	50019 50037 50047	50019	50019 50047	50019	50019	50019 50037 50047
F-Wold's	3	3	1	1	4	3	1	2	1	1	3	1	7
F-Wold 0.95	2	2	1	1	4	3	1	2	1	1	3	1	7
F-Wold 0.9	2	2	1	1	4	3	1	1	1	1	3	1	7
F-F Test	2	2	1	1	4	3	1	2	1	1	3	1	7
F-Haaland's	6	7	1	1	3	17	1	1	1	1	3	1	7
F.-Min Press	6	7	10	6	4	20	11	2	10	20	3	1	12
F.-UNSCR.	6	7	1	6	4	18	1	2	1	1	3	1	6
R^2_{calib}	0.8786	0.9164	0.7818	0.7780	0.8529	0.9982	0.6997	0.8592	0.7070	0.8787	0.8560	0.4799	0.9860
$Offset_{calib}$	0.4753	0.3285	0.8549	0.8697	0.4439	0.0029	0.9065	0.4070	0.8845	0.3749	0.4346	1.5700	0.0406
RMSEC (%)	0.6742	0.5596	0.9041	0.9120	0.6549	0.0734	0.9359	0.5275	0.9245	0.5926	0.6480	1.2317	0.1665
R^2_{CV}	0.7671	0.7575	0.7540	0.7504	0.7098	0.8135	0.6477	0.8374	0.6575	0.8619	0.6945	0.3467	0.8945
$Slope_{CV}$	0.7923	0.7747	0.7436	0.7439	0.7568	0.7379	0.6440	0.8042	0.6520	0.8507	0.6944	0.3925	0.9155
$Offset_{CV}$	0.8023	0.7807	0.9858	0.9892	0.7695	0.6883	1.0555	0.5499	1.0326	0.4479	0.9117	1.8322	0.2612
RMSECV (%)	0.9712	0.9911	0.9984	1.0056	0.9554	0.7659	1.0527	0.5923	1.0379	0.6548	0.9804	1.4335	0.4298
$BIAS_{CV}$	-0.0111	-0.1020	-0.0184	-0.0138	0.0353	-0.1030	-0.0192	-0.0159	-0.0180	-0.0133	-0.0108	-0.0018	0.0171
SECV (%)	0.9904	1.0054	1.0180	1.0254	0.9729	0.7734	1.0726	0.6043	1.0575	0.6677	0.9990	1.4608	0.4383
RPD_{CV}	1.9927	1.9630	1.9388	1.9248	1.7889	2.2504	1.6226	2.3742	1.6458	2.5983	1.7422	1.1914	3.2731
RER_{CV}	7.4228	7.3122	7.2220	7.1696	7.1543	9.0003	6.4895	8.9270	6.5820	10.4252	6.9678	4.7648	12.3068
$RMSECV_{MP}$	0.9712	0.9911	0.9594	0.8978	0.8825	0.6738	0.9608	0.5726	0.9642	0.6033	0.9804	1.4335	0.3996

Table C-22: PLSR statistics for models based on the DS scans and the KL_EF and ASL_EF contents. Refer to Appendix A for descriptions of abbreviations.

Number	*1	*2	*3	*4	*5	*6	*7	*8	*9	*10	*11	*12	*13	*14	*15
Dataset	DS	DS	DS	DS	DS	DS	DS	DS	DS	DS	DS	DS	DS	DS	DS
Constituent	KL_EF	KL_EF	KL_EF	KL_EF	KL_EF	KL_EF	KL_EF	KL_EF	KL_EF	ASL_EF	ASL_EF	ASL_EF	ASL_EF	ASL_EF	ASL_EF
Pre. (1)	NONE	NONE	MSC	EMSC	SNVDT	SG	SG	MSC	EMSC	NONE	NONE	SG	SG	MSC	SNVDT
Specific (1)			F,1.1-2.5	F,1.1-2.5	1.1-2.5,2	1,3,12,12	1,3,7,7	F,1.1-2.5	B,1.1-2.5			1,3,7,7	1,3,14,14	F,1.1-2.5	2,1.1-2.5
Pre. (2)							SNVDT	SG							
Specific (2)							1.1-2.5,2	1,3,7,7							
PLS- λ 10 ³ nm	0.4-2.5	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.2	1.1-2.5	0.4-2.5	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5
Sam. Excl.	50019	50019	50019	50019	50019	50019	50019	50019	50019	50019	50019	50019	50019	50019	50019
F-Wold's	1	8	6	6	6	7	7	7	6	1	1	2	2	1	1
F-Wold 0.95	1	5	6	6	6	7	7	7	6	1	1	2	2	1	1
F-Wold 0.9	1	1	6	1	6	6	6	1	6	1	1	2	1	1	1
F-F Test	1	8	6	6	6	7	7	7	6	1	1	2	2	1	1
F-Haaland's	6	7	6	4	6	6	6	6	5	1	1	1	1	1	1
F.-Min Press	8	8	6	9	6	7	7	7	6	1	3	2	2	3	4
F.-UNSCR.	8	8	6	6	6	7	7	7	6	1	3	2	2	3	4
R^2_{calib}	0.7806	0.9206	0.9112	0.8823	0.9096	0.9246	0.9184	0.9044	0.8922	0.3229	0.2131	0.2074	0.2214	0.3034	0.2973
$Offset_{calib}$	4.2165	1.5251	1.7071	2.2620	1.7378	1.4498	1.5673	1.8374	2.0712	1.4615	1.6986	1.7109	1.6090	1.5036	1.5168
RMSEC (%)	0.3149	0.1894	0.2004	0.2307	0.2022	0.1847	0.1920	0.2079	0.2207	0.1125	0.1212	0.1217	0.1153	0.1141	0.1146
R^2_{CV}	0.4717	0.7120	0.7964	0.7379	0.7883	0.7176	0.7614	0.7159	0.7674	0.1999	0.0823	0.1121	0.1233	0.1934	0.1883
$Slope_{CV}$	0.6274	0.8530	0.8574	0.8813	0.8476	0.7404	0.7963	0.7717	0.8073	0.2337	0.1263	0.1312	0.1436	0.2223	0.2192
$Offset_{CV}$	7.1943	2.8198	2.7385	2.2930	2.9243	5.0092	3.9265	4.4030	3.7047	1.6561	1.8867	1.8765	1.7709	1.6799	1.6862
RMSECV (%)	0.5075	0.3747	0.3150	0.3574	0.3212	0.3710	0.3411	0.3721	0.3368	0.1270	0.1360	0.1337	0.1270	0.1275	0.1279
$BIAS_{CV}$	0.0336	-0.0049	-0.0023	0.0127	-0.0049	0.0201	0.0117	0.0163	0.0008	0.0021	0.0009	0.0011	0.0012	0.0012	0.0008
SECV (%)	0.5160	0.3818	0.3210	0.3640	0.3273	0.3775	0.3473	0.3789	0.3432	0.1294	0.1385	0.1363	0.1294	0.1299	0.1303
RPD_{CV}	1.3278	1.7946	2.1342	1.8823	2.0933	1.8148	1.9726	1.8085	1.9965	1.0767	1.0053	1.0220	1.0285	1.0722	1.0688
RER_{CV}	5.3385	7.2153	8.5805	7.5677	8.4163	7.2964	7.9307	7.2712	8.0270	5.3656	5.0097	5.0931	5.4396	5.3434	5.3265
$RMSECV_{MP}$	0.4797	0.3337	0.3150	0.3338	0.3212	0.3550	0.3309	0.3537	0.3076	0.1270	0.1339	0.1247	0.1229	0.1235	0.1218

Table C-23: PLSR statistics for models based on the DS scans and the AIR_EF content. Refer to Appendix A for descriptions of abbreviations.

Number	*1	*2	*3	*4	*5	*6	*7	*8	*9	*10	*11	*12
Dataset	DS	DS	DS	DS	DS	DS	DS	DS	DS	DS	DS	DS
Constituent	AIR_EF	AIR_EF	AIR_EF	AIR_EF	AIR_EF	AIR_EF	AIR_EF	AIR_EF	AIR_EF	AIR_EF	AIR_EF	AIR_EF
Pre. (1)	NONE	NONE	NONE	NONE	MSC	MSC	MSC	SG	SG	SG	SNVDT	SNVDT
Specific (1)					F,1.1-2.5	F,1.1-2.5	F,1.1-2.5	2,3,10,10	2,3,10,10	2,3,20,20	2,1.1-2.5	2,1.1-2.5
Pre. (2)												
Specific (2)												
PLS-λ 10 ³ nm	0.4-2.5	0.4-2.5	1.1-2.5	1.1-2.5	0.4-2.5	0.4-2.5	0.4-2.5	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5
Sam. Excl.	50019	50019 50047	50019	50019 50037	50019	50019 50047	50019 50037 50047	50019	50019 50037	50019 50037	50019	50019 50037 50047
F-Wold's	1	2	3	6	2	1	1	2	8	6	3	4
F-Wold 0.95	1	2	3	5	2	1	1	2	6	5	1	4
F-Wold 0.9	1	2	1	1	1	1	1	2	2	5	1	4
F-F Test	1	2	3	5	2	1	1	2	6	5	1	4
F-Haaland's	5	4	10	10	5	4	7	2	4	4	3	3
F.-Min Press	20	5	10	14	5	5	8	7	8	6	8	4
F.-UNSCR.	5	5	10	10	5	5	7	7	6	5	8	4
R^2_{calib}	0.8653	0.8864	0.9772	0.9858	0.9271	0.9278	0.9579	0.6162	0.9381	0.9088	0.8437	0.9282
$Offset_{calib}$	2.9930	2.5360	0.5063	0.3123	1.6211	1.6129	0.9340	8.5294	1.3746	2.0256	3.4730	1.5956
RMSEC (%)	0.5686	0.4964	0.2339	0.1725	0.4184	0.3959	0.2805	0.9598	0.3402	0.4130	0.6125	0.3666
R^2_{CV}	0.5941	0.7490	0.7531	0.8839	0.6462	0.8157	0.8565	0.2671	0.8114	0.8032	0.6492	0.9195
$Slope_{CV}$	0.6647	0.7119	0.7363	0.8602	0.6858	0.8358	0.9188	0.3791	0.7953	0.7653	0.6799	0.8811
$Offset_{CV}$	7.5016	6.4329	5.8124	3.0884	7.0651	3.7036	1.7922	13.8168	4.5458	5.1747	7.1304	2.6549
RMSECV (%)	1.0250	0.7608	0.7995	0.5220	0.9570	0.6625	0.5240	1.3773	0.6434	0.6494	0.9530	0.4619
$BIAS_{CV}$	0.0512	-0.0018	-0.0482	-0.0007	0.0835	0.0360	-0.0110	0.0188	0.0005	-0.0382	0.0163	0.0149
SECV (%)	1.0432	0.7759	0.8132	0.5323	0.9715	0.6746	0.5347	1.4034	0.6567	0.6616	0.9710	0.4712
RPD_{CV}	1.5134	1.9362	1.9414	2.7724	1.6251	2.2268	2.6105	1.1250	2.1254	2.1095	1.6260	2.9622
RER_{CV}	6.7617	7.8513	8.6739	13.2510	7.2606	9.0295	11.3932	5.0261	9.2763	9.2070	7.2646	12.9285
$RMSECV_{MP}$	0.9363	0.6745	0.7995	0.4811	0.9570	0.6086	0.5050	1.2278	0.5645	0.5864	0.8520	0.4071

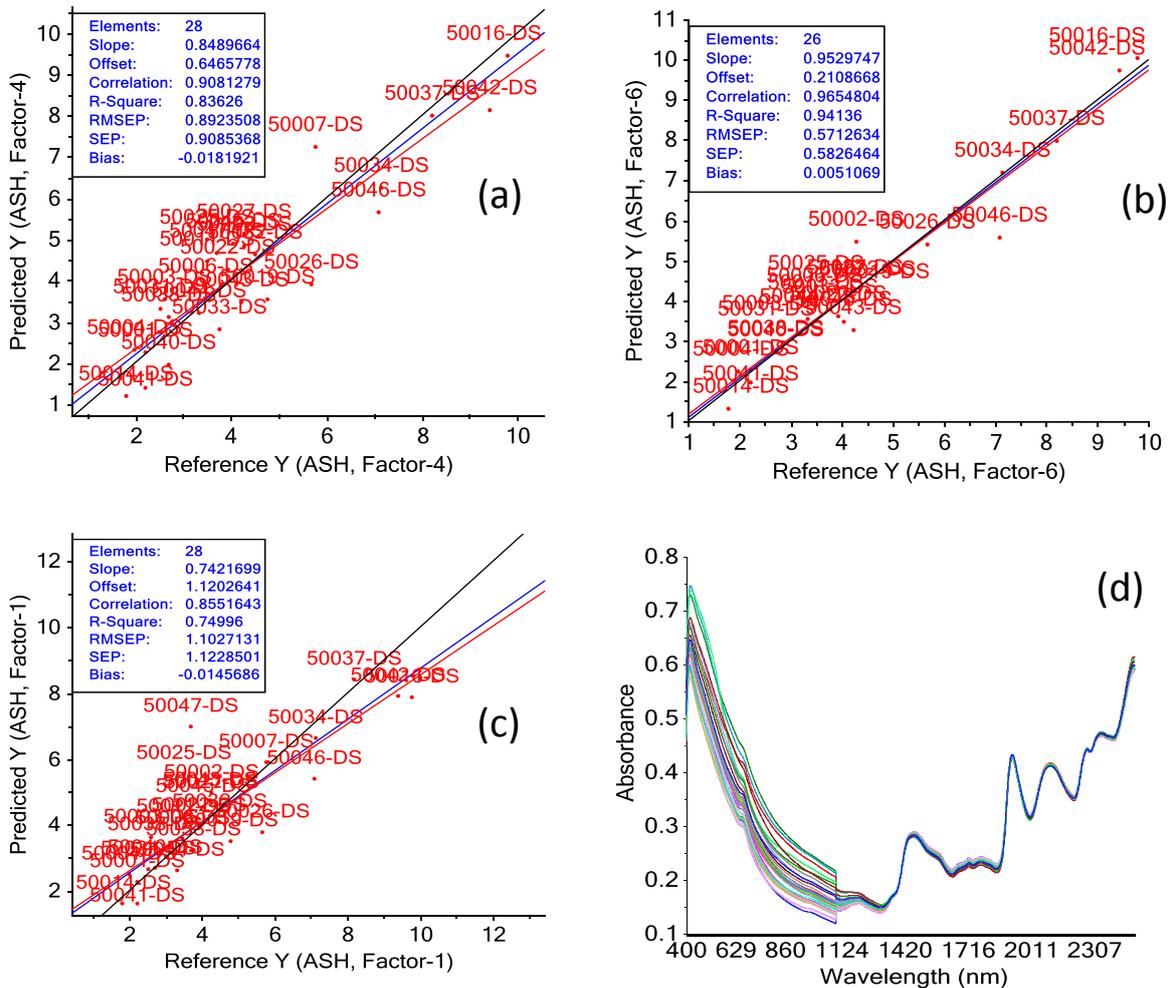


Figure C-26: Plots for ash DS PLSR models. (a) Predicted y vs. reference y for calibration *16 in Table C-23; (b) the same plot for calibration *7; (c) the same plot for calibration *5; (d) the pretreated spectra in calibration *15 (this is equivalent to *16 prior to the derivatisation of the spectra).

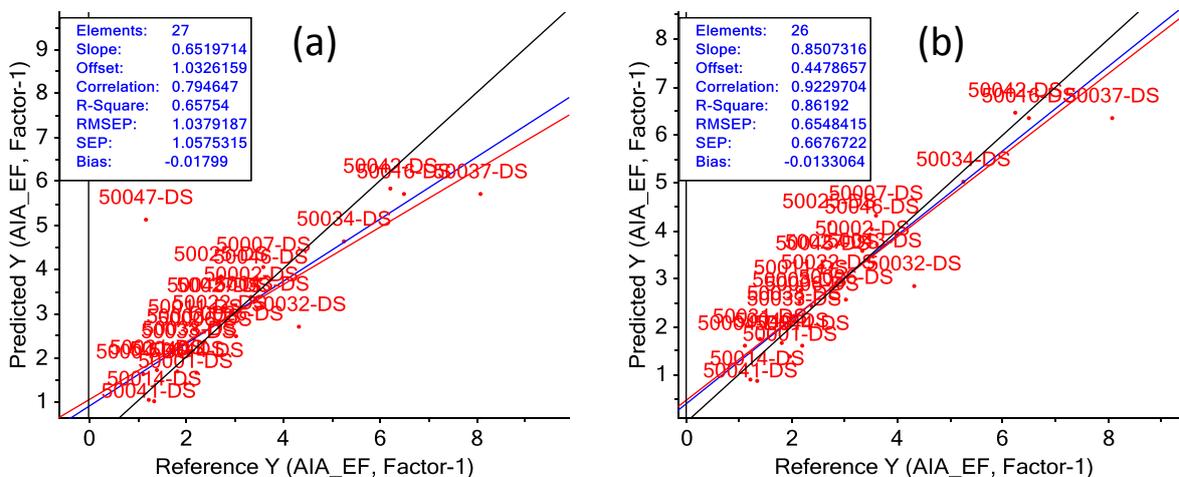


Figure C-27: Plots for AIA_EF DS PLSR models. (a) Predicted vs. reference for *9 in Table C-21 (note the outlier 50047); (b) predicted vs. reference for *10 in Table C-21

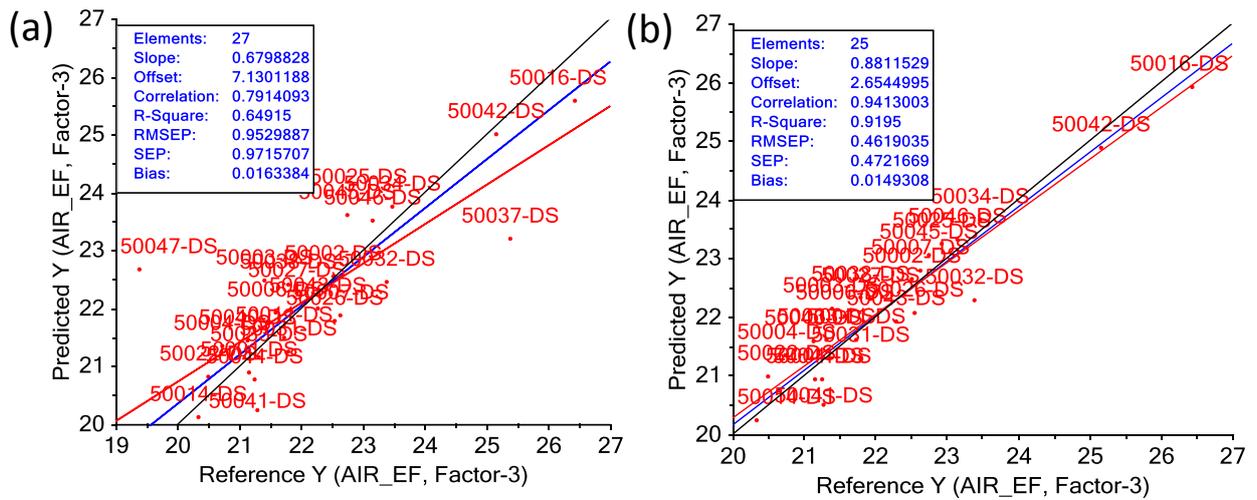


Figure C-28: Plots for AIR_EF DS PLSR models. (a) Predicted vs. reference for *11 in Table C-23; (b) predicted vs. reference for *12 in Table C-23.

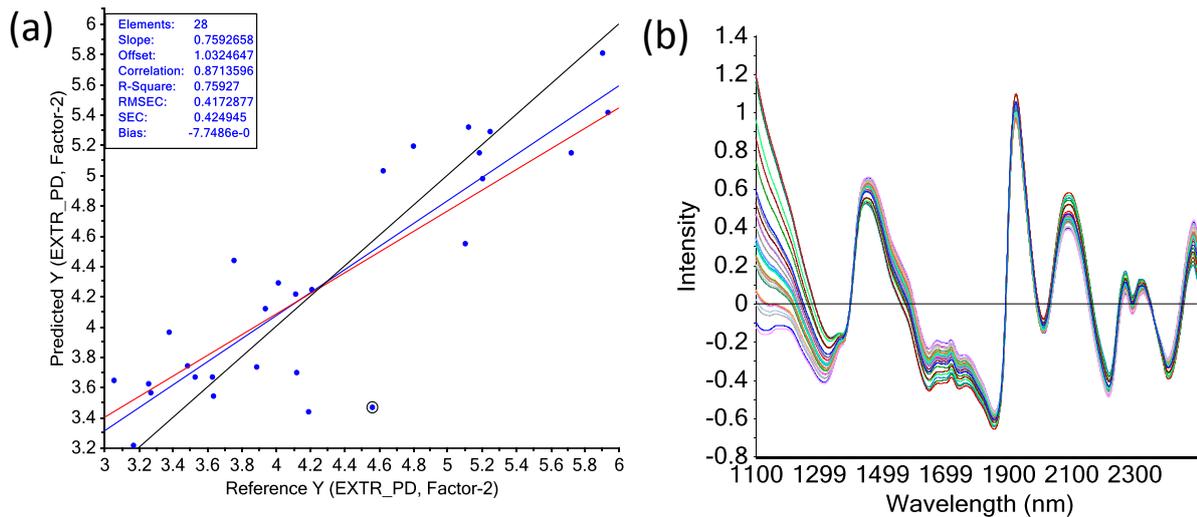


Figure C-29: Plots for EXTR_PD DS PLSR models. (a) The predicted vs. reference plot for calibration *8 in Table C-24 (the blue dot with the circle around it in the middle lower part of the plot represents outlying sample 50025); (b) the 28 DS spectra after the SNVDT treatment over the wavelength region 1100-2500 nm and using a 3rd-order polynomial.

Table C-24: PLSR statistics for models based on the DS scans and the EXTR_PD and EXTR_CV values. Refer to Appendix A for descriptions of abbreviations.

Number	*1	*2	*3	*4	*5	*6	*7	*8	*9	*10	*11	*12	*13	*14	*15	*16	*17	*18	*19	*20	*21	*22
Dataset	DS	DS	DS	DS	DS	DS	DS	DS	DS	DS	DS	DS	DS	DS	DS	DS	DS	DS	DS	DS	DS	DS
Constituent	EXTR_PD																		EXTR_CV			
Pre. (1)	NONE	NONE	SG	SG	SG	SG	SG	MSC	MSC	MSC	EMSC	EMSC	EMSC	SG	EMSC	SNV	SNVDT	SNVDT	SNV	SNV	SNVDT	SNVDT
Specific (1)			1,3,7,7	1,3,7,7	1,3,7,7	1,3,7,7	1,3,7,7	F,1.1-2.5	F,1.1-2.5	F,1.1-2.5,NO 50025	F,1.1-2.5	F,1.1-2.5	A,1.1-2.5,NO 50025	2,3,10,10	F,1.1-2.5,NO 50025	1.1-2.5	2,1.1-2.5	3,1.1-2.5	1.1-2.5	1.1-2.5	3,1.1-2.5	3,1.1-2.5
Pre. (2)															SG							
Specific (2)															2,3,10,10							
PLS-λ 10 ³ nm	0.4-2.5	1.1-2.5	0.4-2.5	1.1-2.5	0.4-0.75	0.4-0.75	0.4-2.5	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5	1.2-2.5	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5
Sam. Excl.						50025	50025		50025	50025	50025	50025	50025	50025	50025	50025	50019	50025	50047	50025	50047	50026
F-Wold's	1	4	6	5	1	1	1	2	3	2	2	2	2	6	8	2	2	2	2	2	2	2
F-Wold 0.95	1	4	6	4	1	1	1	2	3	2	2	2	2	6	7	2	2	2	2	2	2	2
F-Wold 0.9	1	4	5	4	1	1	1	2	3	2	2	2	2	4	4	2	2	2	2	2	2	2
F-F Test	1	4	6	4	1	1	1	2	3	2	2	2	2	6	7	2	2	2	2	2	2	2
F-Haaland's	6	3	5	4	1	1	6	2	2	2	2	2	2	6	5	2	2	2	2	2	2	2
F.-Min Press	6	4	6	12	6	1	6	17	3	2	10	10	9	9	8	2	2	2	2	2	11	2
F.-UNSCR.	6	4	6	5	6	1	6	2	3	2	2	2	2	8	7	2	2	2	2	2	2	2
R ² _{cauib}	0.7940	0.7495	0.8181	0.8367	0.0957	0.0905	0.8669	0.7593	0.8216	0.8231	0.8012	0.8013	0.8013	0.9756	0.9462	0.8231	0.8184	0.8199	0.8008	0.8199	0.7870	0.8372
Offset _{cauib}	0.8833	1.0744	0.7801	0.7005	3.8785	3.8914	0.5696	1.0324	0.7635	0.7569	0.8504	0.8502	0.8500	0.1045	0.2303	0.7569	0.7770	0.7707	0.9088	0.8229	0.9714	0.7469
RMSEC (%)	0.3860	0.4257	0.3627	0.3437	0.8088	0.8244	0.3154	0.4173	0.3652	0.3636	0.3854	0.3853	0.3853	0.1351	0.2005	0.3636	0.3684	0.3669	0.3217	0.3112	0.3326	0.2908
R ² _{CV}	0.6345	0.6015	0.6050	0.7075	-0.0416	-0.2247	0.6054	0.7054	0.7447	0.7825	0.7567	0.7571	0.7452	0.7707	0.7281	0.7807	0.7693	0.7932	0.7547	0.7762	0.7399	0.7947
Slope _{CV}	0.6205	0.5924	0.6654	0.7147	-0.0381	-0.0859	0.6038	0.6808	0.7045	0.7508	0.7503	0.7507	0.7239	0.7119	0.7310	0.7493	0.7366	0.7765	0.7432	0.7650	0.7342	0.7787
Offset _{CV}	1.6040	1.7229	1.4415	1.2145	4.4210	4.5923	1.6849	1.3571	1.2419	1.0530	1.0694	1.0672	1.1652	1.2548	1.1408	1.0586	1.1097	0.9547	1.1626	1.0628	1.2056	1.0084
RMSEC _{CV} (%)	0.5332	0.5568	0.5543	0.4770	0.9002	0.9585	0.5808	0.4787	0.4536	0.4187	0.4428	0.4425	0.4531	0.4299	0.4681	0.4204	0.4312	0.4083	0.3707	0.3609	0.3817	0.3461
BIAS _{CV}	-0.0238	-0.0252	0.0064	-0.0090	-0.0314	-0.0539	-0.0103	-0.0118	-0.0225	-0.0132	0.0008	0.0005	-0.0162	0.0223	-0.0102	-0.0142	-0.0171	-0.0017	-0.0088	-0.0111	-0.0069	-0.0072
SEC _{CV} (%)	0.5425	0.5664	0.5645	0.4857	0.9161	0.9752	0.5918	0.4874	0.4616	0.4265	0.4512	0.4509	0.4614	0.4375	0.4769	0.4281	0.4390	0.4160	0.3776	0.3678	0.3889	0.3529
RPD _{CV}	1.5965	1.5291	1.5344	1.7832	0.9454	0.9033	1.4885	1.7772	1.9082	2.0656	1.9524	1.9537	1.9091	2.0136	1.8473	2.0576	2.0065	2.1173	1.9448	2.0333	1.8884	2.0831
RER _{CV}	5.3025	5.0784	5.0961	5.9226	3.1399	2.9496	4.8606	5.9024	6.2312	6.7452	6.3755	6.3798	6.2341	6.5753	6.0323	6.7191	6.5522	6.9140	7.6196	7.8227	7.3984	8.1550
RMSEC _{MP}	0.5332	0.5140	0.5305	0.4368	0.8071	0.9585	0.5808	0.4641	0.4265	0.4187	0.4353	0.4353	0.4217	0.3938	0.4155	0.4204	0.4312	0.4083	0.3707	0.3609	0.3791	0.3461

Table C-25: PLSR statistics for models based on the DS scans and the lignocellulosic constituents, with the data presented on a whole dry mass basis. Refer to Appendix A for descriptions of abbreviations.

Number	*1	*2	*3	*4	*5	*7	*8	*9	*10	*11	*12	*13	*13	*14
Dataset	DS	DS	DS	DS	DS	DS	DS	DS	DS	DS	DS	DS	DS	DS
Constituent	GLU_SRS	ARA_SRS	RHA_SRS	XYL_SRS	XYL_SRS	KL	GAL_SRS	GAL_SRS	MAN_SRS	AIA	AIR	ASL	EIA	TOT_SRS
Pre. (1)	SNVDT	NONE	NONE	SG	SG	SNVDT	NONE	SG	SG	SNVDT	SNVDT	SNVDT	SNVDT	SNVDT
Specific (1)	1.1-2.5,2			1,3,7,7	1,3,7,7	1.1-2.5,2		1,3,14,14	1,3,14,14	2,1.1-2.5	2,1.1-2.5	1.1-2.5,2	1.1-2.5,2	1.1-2.5,2
Pre. (2)				SNVDT	SNVDT									
Specific (2)				1.1-2.5,2	1.1-2.5,2									
PLS- λ 10 ³ nm	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.2	1.1-2.5	1.1-2.5	0.4-2.5	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5
Sam. Excl.	50019	50019	50019	50019	50019	50019	50019	50019	50019	50019	50019	50019	50019	50019
F-Wold's	4	4	1	3	3	6	1	4	6	1	4	1	1	4
F-Wold 0.95	4	4	1	1	3	6	1	4	2	1	1	1	1	3
F-Wold 0.9	4	1	1	1	1	6	1	4	1	1	1	1	1	3
F-F Test	4	4	1	3	3	6	1	4	6	1	1	1	1	3
F-Haaland's	4	7	8	11	7	5	7	4	5	1	4	1	1	3
F.-Min Press	6	8	10	20	7	6	7	7	6	1	11	4	6	4
F.-UNSCR.	6	8	10	12	7	6	7	4	5	1	4	3	1	3
R^2_{calib}	0.9133	0.8500	0.8650	0.9873	0.9550	0.8958	0.9084	0.8824	0.8587	0.8787	0.9281	0.3094	0.7759	0.8955
$Offset_{calib}$	3.5255	0.3453	0.0147	0.2886	1.0231	1.9173	0.0781	0.1003	0.0209	0.3592	1.5289	1.4273	0.8402	6.9705
RMSEC (%)	0.4440	0.0438	0.0032	0.0661	0.1245	0.2126	0.0275	0.0311	0.0081	0.5686	0.3506	0.1086	0.8773	0.6037
R^2_{cv}	0.8693	0.5489	0.3689	0.5929	0.6379	0.7470	0.7981	0.7512	0.5835	0.8414	0.8743	0.1912	0.7481	0.8547
$Slope_{cv}$	0.8733	0.7712	0.6866	0.6782	0.7081	0.8439	0.8506	0.7933	0.6094	0.8517	0.8927	0.2283	0.7408	0.8505
$Offset_{cv}$	5.1538	0.5132	0.0357	7.3109	6.6446	2.8566	0.1291	0.1736	0.0567	0.4149	2.2851	1.5953	0.9578	9.9478
RMSECV (%)	0.5661	0.0790	0.0072	0.3889	0.3668	0.3440	0.0423	0.0470	0.0150	0.6355	0.4437	0.1220	0.9674	0.7393
$BIAS_{cv}$	0.0028	-0.0135	0.0016	-0.0049	0.0081	-0.0159	0.0017	-0.0027	-0.0009	-0.0242	0.0027	0.0006	-0.0143	-0.0218
SECV (%)	0.5769	0.0793	0.0072	0.3962	0.3736	0.3502	0.0431	0.0478	0.0152	0.6477	0.4528	0.1243	0.9865	0.7531
RPD_{cv}	2.6634	1.4551	1.2414	1.5094	1.6006	1.9165	2.1450	1.9338	1.4499	2.5702	2.9476	1.0708	1.9158	2.5271
REr_{cv}	11.5609	5.4518	4.9623	6.0934	6.4617	7.7334	7.0246	6.3330	5.3573	10.3268	12.0095	5.6632	7.0471	11.7234
$RMSECV_{MP}$	0.5161	0.0758	0.0070	0.3534	0.3668	0.3050	0.0423	0.0466	0.0146	0.6355	0.4062	0.1174	0.8665	0.7345

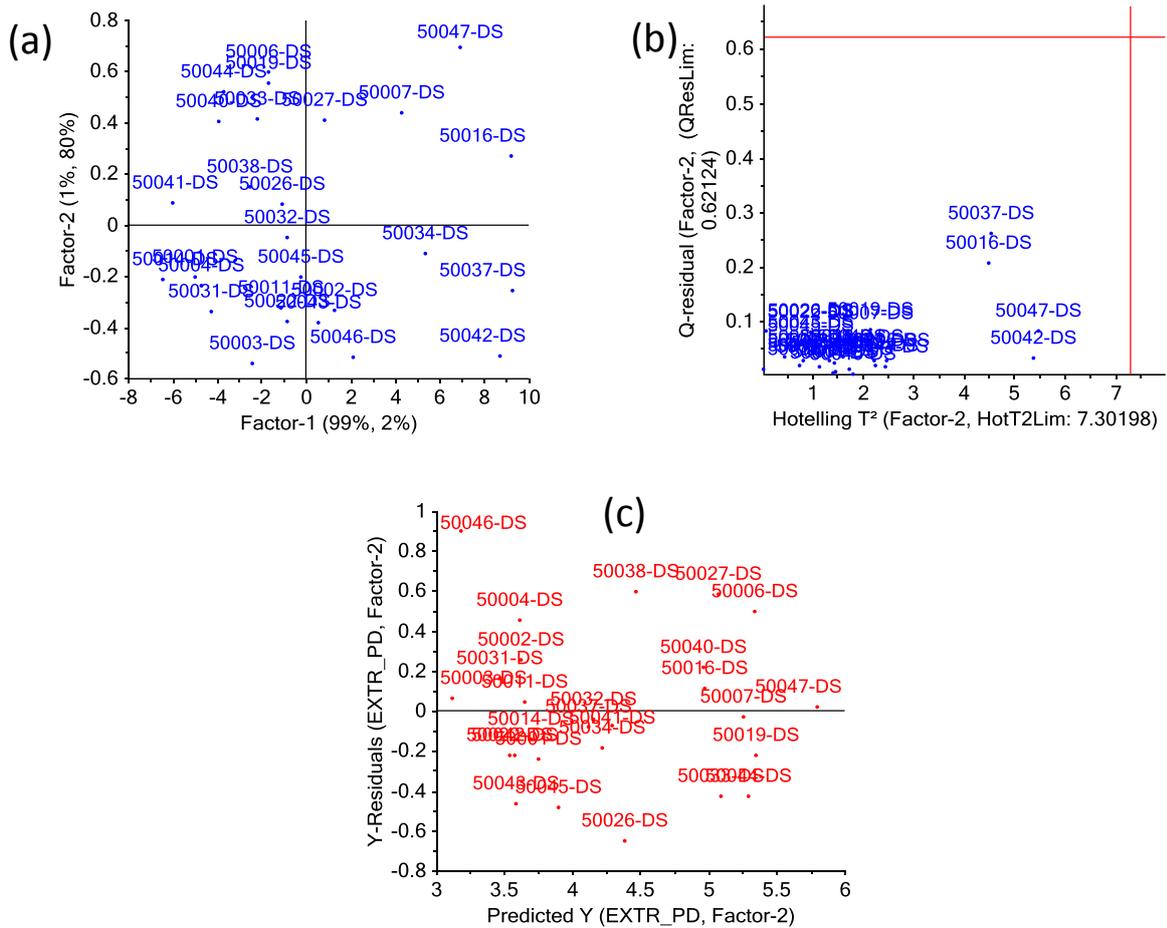


Figure C-30: Plots for EXTR_PD DS model *18 in Table C-24. (a) A F1 vs. F2 plot; (b) an influence plot; (c) a y-residuals vs. predicted y plot.

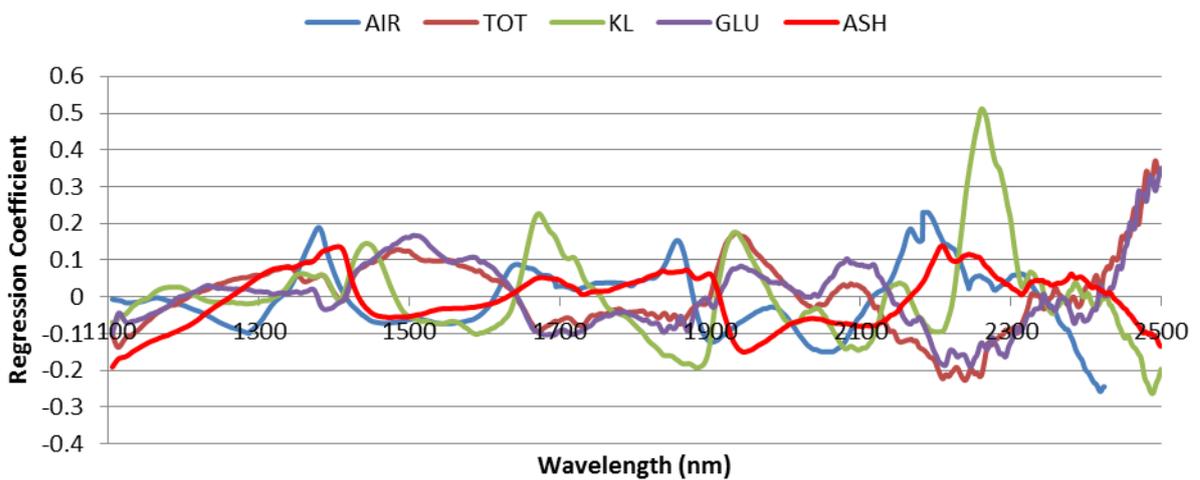


Figure C-31: A regression coefficients plot for models based on the DS dataset and the AIR, TOT_SRS (TOT), KL, GLU_SRS (GLU), and ash constituents. The DS spectra were pretreated with the SNVDT (2nd order polynomial) over the 1100-2500 nm region prior to PLSR over this same region.

Table C-26: PLSR statistics for models based on the DB scans and various chemical constituents. Refer to Appendix A for descriptions of abbreviations.

Number	*1	*2	*3	*4	*5	*6	*7	*8	*9	*10	*11	*12	*13	*14
Dataset	DB	DB	DB	DB	DB	DB	DB	DB	DB	DB	DB	DB	DB	DB
Constituent	GLU_SRS	AIA	EIA	KL	ASL	AIR	ARA_SRS	GAL_SRS	MAN_SRS	ASH	EXTR_PD	XYL_SRS	XYL_SRS	TOT_SRS
Pre. (1)	SNVDT	SNVDT	SNVDT	SNVDT	SNVDT	SNVDT	NONE	SG	SG	MSC	SNVDT	SG	SG	SNVDT
Specific (1)	1.1-2.5,2	1.1-2.5,2	1.1-2.5,2	1.1-2.5,2	1.1-2.5,2	1.1-2.5,2		4,4,30,30	1,3,14,14	F,1.1-2.5	1.1-2.5,3	1,3,7,7	1,3,7,7	1.1-2.5,2
Pre. (2)										SG		SNV	SNV	
Specific (2)										1,3,14,14		1.1-2.5	1.1-2.5	
PLS-λ 10 ³ nm	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5	1.1-1.8	1.1-2.5	0.4-2.5	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5
Sam. Excl.	50019	50019 50047	50019 50046	50019	50019	50019 50047	50019	50019	50003 50019			50019	50019 50007	50019
F-Wold's	4	2	2	7	3	4	2	4	2	5	3	4	6	1
F-Wold 0.95	4	2	2	6	1	3	2	1	1	1	1	4	6	1
F-Wold 0.9	4	2	2	1	1	3	1	1	1	1	1	2	6	1
F-F Test	4	2	2	6	1	3	2	1	2	5	1	4	6	1
F-Haaland's	3	2	2	10	3	2	7	3	1	5	1	12	6	3
F.-Min Press	4	2	4	13	3	4	8	4	8	18	3	13	18	4
F.-UNSCR.	4	2	4	11	3	3	8	4	2	5	1	13	6	4
R^2_{calib}	0.9208	0.8728	0.8645	0.9471	0.7294	0.8769	0.8454	0.5955	0.2626	0.9347	0.2346	0.9960	0.9363	0.9101
$Offset_{calib}$	3.2002	0.4125	0.5580	0.9738	0.5537	2.6537	0.3561	0.3441	0.1099	0.3481	3.3017	0.0905	1.4450	5.9613
RMSEC (%)	0.5336	0.7853	0.9905	0.1552	0.0810	0.6583	0.0453	0.0602	0.0192	0.8851	0.7540	0.0496	0.1980	0.7604
R^2_{CV}	0.8844	0.8598	0.7987	0.7839	0.6301	0.8444	0.4442	0.3728	0.2167	0.8593	0.1920	0.5857	0.7434	0.8444
$Slope_{CV}$	0.8755	0.8094	0.7759	0.8554	0.6291	0.7978	0.6270	0.4368	0.1492	0.8778	0.1816	0.6767	0.6699	0.8182
$Offset_{CV}$	5.0234	0.5916	0.8528	2.6689	0.7620	4.3222	0.8516	0.4774	0.1278	0.6149	3.5245	7.3558	7.4740	12.0811
RMSECV (%)	0.6676	0.9105	1.2520	0.3248	0.0980	0.8016	0.0890	0.0776	0.0216	1.3275	0.8014	0.5236	0.4602	1.0365
$BIAS_{CV}$	-0.0058	-0.0263	-0.0703	0.0096	0.0031	-0.0388	-0.0072	-0.0017	0.0009	-0.0364	-0.0057	0.0340	-0.0126	0.0201
SECV (%)	0.6793	0.9268	1.2729	0.3304	0.0997	0.8159	0.0903	0.0790	0.0219	1.3413	0.8151	0.5317	0.4685	1.0546
RPD_{CV}	2.8400	2.4188	2.1528	2.0779	1.5883	2.3438	1.2994	1.2194	1.0405	2.6098	1.0754	1.5032	1.7055	2.4479
RER_{CV}	12.5205	10.9705	10.1328	8.1959	7.5201	11.0625	4.7869	3.8317	3.7204	11.9656	3.5291	7.8351	8.8933	12.0077
$RMSECV_{MP}$	0.6189	0.9105	1.1611	0.3049	0.0980	0.7416	0.0865	0.0769	0.0199	1.2547	0.7959	0.4734	0.4150	0.9435

Table C-27: PLSR statistics for models based on the DB scans and the carbon and nitrogen contents. Refer to Appendix A for descriptions of abbreviations.

Number	*1	*2	*3	*4	*5	*6	*7	*8	*9	*10	*11	*12	*13
Dataset	DB	DB	DB	DB	DB	DB	DB	DB	DB	DB	DB	DB	DB
Constituent	CARBON	CARBON	CARBON	CARBON	CARBON	CARBON	CARBON	CARBON	CARBON	CARBON	NITROGEN	NITROGEN	NITROGEN
Pre. (1)	NONE	NONE	NONE	SG	SNVDT	SNVDT	MSC	EMSC	SG	SG	NONE	EMSC	SG
Specific (1)				1,3,14,14	1.1-2.5,2	1.1-2.5,2	F,1.1-2.5	A,1.1-2.5	2,3,14,14	2,3,14,14		A,1,1-2.5	2,3,14,14
Pre. (2)													
Specific (2)													
PLS- λ 10 ³ nm	0.4-2.5	1.1-2.5	1.1-2.5	1.1-2.5	0.4-2.5	1.1-2.5	1.1-2.5	1.1-2.5	0.4-2.5	1.1-2.5	0.4-2.5	1.1-2.5	1.1-2.5
Sam. Excl.			50007										
F-Wold's	1	2	3	3	3	2	3	6	1	2	1	2	2
F-Wold 0.95	1	2	2	2	1	1	2	3	1	2	1	1	1
F-Wold 0.9	1	1	2	2	1	1	1	2	1	2	1	1	1
F-F Test	1	2	2	3	2	2	2	3	1	2	1	2	1
F-Haaland's	1	1	2	2	1	1	1	2	1	2	1	8	10
F.-Min Press	1	2	5	3	6	6	3	6	1	2	1	11	20
F.-UNSCR.	1	2	5	3	1	1	2	6	1	2	1	11	12
R^2_{calib}	0.7214	0.5858	0.7287	0.6499	0.7104	0.6938	0.6704	0.6580	0.7000	0.6963	0.1870	0.6504	0.8795
Offset _{calib}	12.5359	18.6389	12.2115	15.7522	13.0338	13.7805	14.8338	15.3897	13.5004	13.6657	0.1993	0.0857	0.0295
RMSEC (%)	0.9169	1.1180	0.9137	1.0278	0.9349	0.9613	0.9974	1.0159	0.9515	0.9573	0.0188	0.0123	0.0072
R^2_{cv}	0.6897	0.5067	0.6759	0.5694	0.6771	0.6581	0.6102	0.5915	0.6548	0.6212	0.1518	0.2648	0.1784
Slope _{cv}	0.6861	0.5457	0.6901	0.5955	0.6697	0.6527	0.6047	0.5972	0.6406	0.6324	0.1491	0.4186	0.3675
Offset _{cv}	14.1451	20.3978	13.9531	18.2152	14.8803	15.6472	17.8126	18.1478	16.1943	16.5519	0.2085	0.1422	0.1530
RMSECv (%)	0.9888	1.2465	1.0032	1.1647	1.0085	1.0377	1.1082	1.1344	1.0428	1.0924	0.0196	0.0182	0.0193
BIAS _{cv}	0.0178	-0.0437	0.0036	0.0144	0.0175	0.0187	0.0243	0.0236	0.0194	0.0080	0.0000	-0.0003	-0.0020
SECV (%)	0.9993	1.2592	1.0143	1.1772	1.0193	1.0488	1.1199	1.1464	1.0539	1.1041	0.0198	0.0184	0.0194
RPD _{cv}	1.7571	1.3944	1.7485	1.4916	1.7227	1.6742	1.5679	1.5317	1.6661	1.5903	1.0627	1.1416	1.0855
RER _{cv}	7.7955	6.1863	7.6800	6.6176	7.6427	7.4276	6.9561	6.7952	7.3917	7.0552	4.5446	4.8819	4.6422
RMSECV _{MP}	0.9888	1.2073	0.9269	1.1463	0.9863	1.0191	1.0546	1.0464	1.0428	1.0924	0.0196	0.0174	0.0180

Table C-28: PLSR statistics for models based on the DW scans and various constituents. Refer to Appendix A for descriptions of abbreviations.

Number	*1	*2	*3	*4	*5	*6	*7	*8	*9	*10	*11	*12	*13
Dataset	DW	DW	DW	DW	DW	DW	DW	DW	DW	DW	DW	DW	DW
Constituent	GLU_SRS	GLU_SRS	GLU_SRS	GLU_SRS	GLU_SRS	GLU_SRS	KL	KL	KL	XYL_SRS	TOT_SRS	TOT_SRS	TOT_SRS
Pre. (1)	SNVDT	NONE	SG	MSC	SG	SG	SNVDT	MSC	SG	SG	SNVDT	NONE	NONE
Specific (1)	1.1-2.5,2		1,3,14,14	F,1.1-2.5	1,3,14,14	2,3,14,14	1.1-2.5,2	F,1.1-2.5	1,3,14,14	1,3,14,14	1.1-2.5,2		
Pre. (2)					SNVDT					SNV			
Specific (2)					1.1-2.5,2					1.1-2.5			
PLS-λ 10 ³ nm	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5	0.4-2.5
Sam. Excl.	50019	50019,5001	50019	50019	50019	50019	50019	50019	50019	50019	50019	50019	50019
F-Wold's	4	1	6	4	4	6	1	3	2	10	1	1	2
F-Wold 0.95	1	1	6	4	4	1	1	2	1	2	1	1	2
F-Wold 0.9	1	1	6	4	4	1	1	1	1	2	1	1	1
F-F Test	1	1	6	4	4	3	1	3	1	2	1	1	2
F-Haaland's	3	5	5	3	3	9	10	3	4	10	3	5	8
F.-Min Press	4	5	6	4	4	13	12	13	8	10	4	5	9
F.-UNSCR.	4	5	6	4	4	10	12	13	8	10	4	5	9
R_{calib}^2	0.9210	0.9549	0.9530	0.9223	0.9114	0.9868	0.8419	0.4243	0.5424	0.9473	0.8682	0.9545	0.9634
$Offset_{calib}$	3.1922	1.8196	1.8975	3.1370	3.5767	0.5336	2.9083	10.5869	8.4150	1.1941	8.7455	3.0181	2.4260
RMSEC (%)	0.5330	0.4026	0.4109	0.5283	0.5642	0.2179	0.2682	0.5119	0.4564	0.1803	0.9211	0.5411	0.4851
R_{CV}^2	0.8669	0.8366	0.8433	0.8734	0.8167	0.8137	0.2327	0.1314	0.2013	0.6803	0.7604	0.9208	0.8227
$Slope_{CV}$	0.8582	0.8406	0.7952	0.8620	0.8087	0.7891	0.6492	0.2703	0.3388	0.6847	0.7514	0.8626	0.8325
$Offset_{CV}$	5.7269	6.3844	8.2550	5.5900	7.7123	8.3747	6.4915	13.4302	12.1747	7.1476	16.5120	9.0925	11.0646
RMSECV (%)	0.7163	0.6692	0.7773	0.6986	0.8406	0.8475	0.6120	0.6512	0.6244	0.4599	1.2861	0.7396	1.1062
$BIAS_{CV}$	0.0018	-0.0443	-0.0156	0.0166	-0.0130	-0.1430	0.0416	0.0112	0.0163	0.0072	0.0176	-0.0228	-0.0454
SECV (%)	0.7290	0.6800	0.7909	0.7107	0.8553	0.8502	0.6214	0.6626	0.6352	0.4680	1.3088	0.7523	1.1249
RPD_{CV}	2.6466	2.8384	2.4394	2.7146	2.2556	2.2693	1.1048	1.0361	1.0808	1.7078	1.9726	3.4315	2.2951
RER_{CV}	11.6682	12.5091	10.7547	11.9679	9.9443	10.0047	4.3577	4.0868	4.2629	8.9020	9.6758	16.8321	11.2578
$RMSECV_{MP}$	0.6679	0.6692	0.7254	0.6343	0.7559	0.7629	0.5528	0.5838	0.5659	0.4599	1.1973	0.7396	0.9784

Table C-29: PLSR statistics for further models based on the DW scans and various constituents. Refer to Appendix A for descriptions of abbreviations.

Number	*1	*2	*3	*4	*5	*6	*7	*8	*9	*10	*11	*12	*13	*14	*15
Dataset	DW	DW	DW	DW	DW	DW	DW	DW	DW	DW	DW	DW	DW	DW	DW
Constituent	ASL	GAL	GAL	AIR	AIA	AIA	EIA	ASH	ASH	EXTR_PD	ARA_SRS	ARA_SRS	MAN_SRS	CARBON	NITR.
Pre. (1)	SNVDT	NONE	SG	SNVDT	SNVDT	NONE	SNVDT	MSC	NONE	SNVDT	NONE	SNVDT	SG	SNVDT	EMSC
Specific (1)	1.1-2.5,2		1,3,14,14	1.1-2.5,2	1.1-2.5,2		1.1-2.5,2	F,1.1-2.5		1.1-2.5,2		1.1-2.5,2	1,3,14,14	1.1-2.5,2	A,1.1-2.5
Pre. (2)								SG							
Specific (2)								1,3,14,14							
PLS-λ 10 ³ nm	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5	0.4-2.5	0.4-2.5	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5	0.4-2.5	1.1-2.5
Sam. Excl.	50019	50019	50019	50019 50037 50047	50019 50047	50019 50037 50047	50019 50046	50020			50019	50019	50019		
F-Wold's	1	2	1	3	2	3	3	1	2	2	1	4	1	4	2
F-Wold 0.95	1	1	1	3	2	1	3	1	2	1	1	1	1	4	1
F-Wold 0.9	1	1	1	3	1	1	3	1	1	1	1	1	1	1	1
F-F Test	1	1	1	3	2	1	3	1	2	1	1	4	1	4	2
F-Haaland's	3	15	3	3	1	5	3	7	9	14	4	3	5	4	11
F.-Min Press	3	16	18	3	2	5	3	10	11	14	17	4	20	4	13
F.-UNSCR.	3	16	18	3	2	5	3	9	11	14	6	5	17	4	13
R^2_{calib}	0.7098	0.9803	0.4757	0.9060	0.7471	0.9042	0.8418	0.9456	0.9376	0.9492	0.4514	0.5477	0.6854	0.8149	0.7284
$Offset_{calib}$	0.6014	0.0158	0.4460	2.0273	0.8197	0.2944	0.6516	0.2835	0.3325	0.2177	1.2633	1.0402	0.0465	8.3272	0.0666
RMSEC (%)	0.0858	0.0133	0.0685	0.5754	1.1070	0.6376	1.0704	0.7939	0.8654	0.1974	0.0854	0.0780	0.0131	0.7473	0.0109
R^2_{CV}	0.5493	0.7001	0.2011	0.8311	0.6783	0.7683	0.7310	0.8320	0.8043	0.5871	0.2107	0.2280	0.0852	0.7200	0.1860
$Slope_{CV}$	0.5738	0.6768	0.3184	0.8251	0.6582	0.7446	0.7460	0.8649	0.8858	0.8545	0.3407	0.2723	0.3406	0.7433	0.4013
$Offset_{CV}$	0.8725	0.2737	0.5838	3.7990	1.0284	0.7357	1.0859	0.6923	0.6621	0.6382	1.5242	1.6745	0.0986	11.5207	0.1461
RMSECV (%)	0.1070	0.0537	0.0876	0.9014	1.2662	1.0975	1.4473	1.4718	1.5652	0.5832	0.1061	0.1099	0.0231	0.9391	0.0192
$BIAS_{CV}$	-0.0020	-0.0013	0.0040	0.0278	-0.0796	-0.0494	0.0397	-0.0117	0.0532	0.0098	0.0059	0.0011	0.0013	-0.0291	-0.0007
SECV (%)	0.1091	0.0546	0.0891	0.9182	1.2869	1.1173	1.4733	1.4880	1.5812	0.5934	0.1078	0.1119	0.0235	0.9488	0.0194
RPD_{CV}	1.5214	1.7635	1.0813	2.0826	1.7421	1.8787	1.8600	2.3126	2.2139	1.5027	1.0884	1.0560	1.0110	1.8506	1.0855
RER_{CV}	6.8716	5.5414	3.3978	9.8299	7.9011	9.1001	8.7547	10.7859	10.1503	4.8473	4.0097	3.8631	3.6390	8.2100	4.6418
$RMSECV_{MP}$	0.1070	0.0485	0.0776	0.9014	1.2093	1.0975	1.4473	1.3530	1.4826	0.5832	0.1006	0.0995	0.0204	0.9391	0.0181

Table C-30: PLSR statistics for models based on the DU scans and various constituents. Refer to Appendix A for descriptions of abbreviations.

Number	*1	*2	*3	*4	*5	*6	*7	*8	*9	*10	*11	*12	*13	*14	*15	
Dataset	DU	DU	DU	DU	DU	DU	DU	DU	DU	DU	DU	DU	DU	DU	DU	
Constituent	GLU_SRS	KL	TOT	XYL_SRS	ASL	AIR	AIA	EIA	ASH	EXTR_PD	ARA_SRS	MAN_SRS	GAL_SRS	CARBON	NITROGEN	
Pre. (1)	SNVDT	SNVDT	NONE	SG	SNVDT	SNVDT	SNVDT	SNVDT	MSC	SNVDT	SNVDT	SG	N/A	SNVDT	EMSC	
Specific (1)	1.1-2.5,2	1.1-2.5,2		1,3,14,14	1.1-2.5,2	1.1-2.5,2	1.1-2.5,2	1.1-2.5,2	F,1.1-2.5	1.1-2.5,2	1.1-2.5,2	1,3,14,14		1.1-2.5,2	A,1.1-2.5	
Pre. (2)				SNV					SG							
Specific (2)				1.1-2.5					1,3,14,14							
PLS-λ 10 ³ nm	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5	0.4-2.5	1.1-2.5	1.1-2.5	1.1-2.5			1.1-2.5	1.1-2.5
Sam. Excl.	50019	50019 50042	50019 50001	50019	50019	50019 50037 50047	50019 50037	50019 50046				50019 50019				
F-Wold's	3	1	5	2	3	3	2	2	8	1	4	1			3	2
F-Wold 0.95	3	1	2	2	3	3	2	2	5	1	1	1			1	1
F-Wold 0.9	1	1	2	2	3	3	2	2	2	1	1	1			1	1
F-F Test	3	1	3	2	3	3	2	2	5	1	1	1			3	2
F-Haaland's	3	9	5	9	2	5	2	6	6	13	7	8			1	1
F.-Min Press	3	9	5	10	3	5	9	6	8	14	8	14			3	2
F.-UNSCR.	3	9	5	10	3	5	6	6	4	14	8	14			3	2
R^2_{calib}	0.9384	0.9482	0.9226	0.9449	0.6336	0.9443	0.8161	0.9569	0.9555	0.9217	0.8565	0.8398			0.6874	0.0713
$Offset_{calib}$	2.4897	0.9527	5.1263	1.2483	0.7496	1.1999	0.5525	0.1767	0.4636	0.3395	0.3304	0.0237			14.0660	0.2207
RMSEC (%)	0.4707	0.1562	0.7048	0.1844	0.0942	0.4427	0.8815	0.5586	0.8096	0.2412	0.0437	0.0093			0.9712	0.0177
R^2_{CV}	0.8654	0.7745	0.7513	0.3330	0.5721	0.8620	0.7624	0.8811	0.8697	0.2113	0.4771	0.2950			0.6725	0.0245
$Slope_{CV}$	0.8934	0.8501	0.8253	0.4516	0.5890	0.8625	0.7250	0.8829	0.8613	0.4635	0.6680	0.5294			0.6696	0.0525
$Offset_{CV}$	4.2821	2.7663	11.7793	12.4148	0.8376	3.0104	0.7553	0.3907	0.7949	2.3764	0.7551	0.0710			14.9119	0.2329
RMSECV (%)	0.7204	0.3259	1.2514	0.6643	0.1054	0.7723	1.0945	0.9621	1.3399	0.7918	0.0863	0.0203			1.0157	0.0187
$BIAS_{CV}$	-0.0216	0.0089	0.2094	-0.0049	-0.0033	0.0449	-0.0708	-0.0916	-0.0760	0.0620	-0.0094	0.0016		0.0461	0.0019	
SECV (%)	0.7329	0.3317	1.2564	0.6760	0.1073	0.7857	1.1122	0.9753	1.3632	0.8028	0.0873	0.0206		1.0256	0.0189	
RPD_{CV}	2.6326	2.1071	2.0535	1.1823	1.4768	2.4337	1.8819	2.8098	3.0893	1.0919	1.3433	1.1533		1.7120	1.0003	
RER_{CV}	11.6061	8.1632	10.0791	6.1624	6.9921	11.4868	9.1414	13.2253	11.3624	3.5831	4.9486	4.1512		7.5955	4.7605	
$RMSECV_{MP}$	0.7204	0.3259	1.2514	0.6368	0.0974	0.7723	0.9769	0.9621	1.1757	0.7497	0.0778	0.0185		0.9294	0.0184	

Table C-31: PLSR statistics for models based on the WU scans and the GLU_SRS, KL, AIR, GAL_SRS, and TOT_SRS contents. Refer to Appendix A for descriptions of abbreviations.

Number	*1	*2	*3	*4	*5	*6	*7	*8	*9	*10	*11	*12	*13	*14	*15	*16	*17
Dataset	WU										WU			WU	WU	WU	
Constituent	GLU_SRS										KL			AIR	GAL_SRS	TOT_SRS	
Pre. (1)	NONE	SNV	SNVDT	SNVDT	SG	SG	MSC	MSC	EMSC	EMSC	SNVDT	SNVDT	SNVDT	SNVDT	SG	EMSC	EMSC
Specific (1)		1.1-2.5	1.1-1.8,2	1.1-2.5,2	1,3,14,14	1,3,14,14	F,1.1-2.5	F,1.1-2.5	F,1.1-2.5	F,1.1-2.5	1.1-2.5,2	1.1-2.5,2	1.1-1.8,2	1.1-2.5,2	1,3,14,14	F,1.1-2.5	F,1.1-2.5
Pre. (2)																	
Specific (2)																	
PLS-λ 10 ³ nm	0.4-2.5	1.1-2.5	1.1-1.8	1.1-2.5	1.1-2,5	1.1-1.8	1.1-2.5	1.1-1.8	1.1-2.5	1.1-1.8	1.1-2.5	1.1-1.8	1.1-1.8	1.1-2.5	1.1-2.5	1.1-2.5	1.1-1.8
Sam. Excl.	50019	50019	50019	50019	50019	50019	50019	50019	50019	50019	50019	50019	50019	50019	50019	50019	50019
F-Wold's	2	1	1	5	6	5	1	5	4	3	5	1	1	3	2	4	1
F-Wold 0.95	2	1	1	5	6	4	1	4	4	3	1	1	1	3	1	4	1
F-Wold 0.9	2	1	1	4	3	4	1	1	4	1	1	1	1	3	1	4	1
F-F Test	2	1	1	5	6	4	1	5	4	3	2	1	1	3	1	4	1
F-Haaland's	9	4	3	4	6	3	4	3	4	3	12	6	5	3	9	12	7
F.-Min Press	14	5	4	5	10	5	6	5	16	3	13	7	5	5	10	15	7
F.-UNSCR.	10	5	3	4	6	4	5	3	4	3	13	7	5	5	10	4	7
R ² _{calib}	0.9730	0.9384	0.9239	0.9354	0.9571	0.8757	0.9355	0.9111	0.9405	0.9222	0.9565	0.6527	0.6369	0.8981	0.9302	0.9859	0.9666
Offset _{calib}	1.0924	2.4864	3.0742	2.6076	1.7308	5.1671	2.6031	3.5905	2.4011	3.1421	0.7980	6.3855	6.6771	2.1971	0.0592	0.9374	2.2163
RMSEC (%)	0.3114	0.4704	0.5230	0.4817	0.3924	0.6800	0.4813	0.5653	0.4623	0.5288	0.1407	0.3976	0.4065	0.5990	0.0252	0.3011	0.4637
R ² _{CV}	0.8186	0.8945	0.8812	0.9029	0.8760	0.8180	0.8869	0.8668	0.9014	0.8933	0.6578	0.2075	0.4038	0.8387	0.3304	0.9028	0.9034
Slope _{CV}	0.9423	0.8828	0.8882	0.8967	0.9006	0.8184	0.8779	0.8660	0.9000	0.8924	0.8100	0.4865	0.4720	0.8271	0.5011	0.9460	0.9075
Offset _{CV}	2.3090	4.7787	4.5786	4.1789	4.0319	7.3429	4.9482	5.4252	4.0459	4.3740	3.4759	9.4719	9.7254	3.6966	0.4208	3.6704	6.2181
RMSECV (%)	0.8362	0.6377	0.6769	0.6119	0.6913	0.8523	0.6604	0.7165	0.6164	0.6414	0.4087	0.6220	0.5395	0.8078	0.0806	0.8190	0.8168
BIAS _{CV}	-0.0232	0.0444	0.0650	0.0066	0.0170	0.0158	0.0190	0.0146	0.0083	0.0276	-0.0177	0.0302	0.0164	-0.0321	-0.0025	0.0891	0.0807
SECV (%)	0.8507	0.6474	0.6857	0.6226	0.7034	0.8678	0.6719	0.7291	0.6273	0.6522	0.4156	0.6322	0.5488	0.8226	0.0820	0.8285	0.8272
RPD _{CV}	2.2679	2.9800	2.8138	3.0986	2.7430	2.2420	2.8717	2.6464	3.0756	2.9584	1.6520	1.0859	1.2510	2.3247	1.1856	3.1160	3.1211
RER _{CV}	9.9983	13.1379	12.4051	13.6607	12.0929	9.8020	12.6602	11.6669	13.5594	13.0426	6.5159	4.2831	4.9345	10.9723	3.6907	15.2847	15.3095
RMSECV _{MP}	0.7400	0.5985	0.6625	0.5887	0.6327	0.7616	0.6165	0.6759	0.5633	0.6414	0.3812	0.5679	0.5395	0.7736	0.0791	0.7713	0.8168

Table C-32: PLSR statistics for models based on the WU scans and the XYL_SRS content. Refer to Appendix A for descriptions of abbreviations.

Number	*1	*2	*3	*4	*5	*6	*7	*8	*9	*10	*11	*12	*13	*14	*15	*16	*17	
Dataset	WU																	
Constituent	XYL_SRS																	
Pre. (1)	MSC	EMSCF	SG	SG	SG	SG	SG	SG	EMSC									
Specific (1)	F,1.1-2.5	F,1.1-2.5	1,3,14,14	1,3,14,14	1,3,14,14	2,3,14,14	2,3,14,14	2,3,14,14	3,3,16,16	3,3,25,25	4,4,16,16	4,4,30,30	4,5,50,50	4,5,30,30	4,4,40,40	1,4,30,30	F,1.1-2.5	
Pre. (2)																	SG	
Specific (2)																	4,4,30,30	
PLS- λ 10 ³ nm	1.1-2.5	1.1-1.8	0.4-2.5	1.1-2.5	1.1-1.8	0.4-2.5	1.1-2.5	1.1-1.8	1.1-1.8	1.1-1.8	1.1-1.8	1.1-1.8	1.1-1.8	1.1-1.8	1.1-1.8	1.1-1.8	1.1-1.8	
Sam. Excl.	50019	50019	50019	50019	50019	50019	50019	50019	50019	50019	50019	50019	50019	50019	50019	50019	50019	
F-Wold's	1	1	3	3	3	2	2	2	2	2	2	3	2	3	2	2	3	
F-Wold 0.95	1	1	3	2	2	2	2	2	2	2	2	2	2	2	2	2	1	
F-Wold 0.9	1	1	1	2	2	1	2	2	2	2	2	2	2	2	2	2	1	
F-F Test	1	1	3	2	2	2	2	2	2	2	2	2	2	2	2	2	2	
F-Haaland's	1	12	3	2	10	11	6	6	8	8	7	9	5	9	8	10	13	
F.-Min Press	1	20	6	7	11	12	8	7	9	17	10	18	6	17	17	11	20	
F.-UNSCR.	1	13	6	7	11	12	8	7	9	8	10	9	6	9	12	11	14	
R^2_{calib}	0.4922	0.9474	0.5821	0.5280	0.9377	0.9913	0.9039	0.8980	0.9546	0.9367	0.9656	0.9565	0.8551	0.9568	0.9308	0.9341	0.9983	
$Offset_{calib}$	11.5017	1.1892	9.4653	10.6910	1.4085	0.1973	2.1773	2.3094	1.0276	1.4339	0.7802	0.9849	3.2824	0.9785	1.5673	1.4925	0.0390	
RMSEC (%)	0.5597	0.1802	0.5077	0.5396	0.1959	0.0733	0.2435	0.2508	0.1673	0.1976	0.1458	0.1638	0.2990	0.1632	0.2066	0.2016	0.0326	
R^2_{cv}	0.3571	0.5503	0.3812	0.3713	0.5927	0.5892	0.5911	0.5764	0.7203	0.7533	0.2591	0.7499	0.6170	0.7378	0.7013	0.5617	0.7895	
$Slope_{cv}$	0.3639	0.7416	0.4358	0.4028	0.6881	0.5946	0.6375	0.6116	0.7264	0.7856	0.4158	0.7932	0.6347	0.7760	0.7539	0.6704	0.7756	
$Offset_{cv}$	14.4083	5.8324	12.8099	13.5388	7.1100	9.2291	8.2196	8.8343	6.2048	4.8861	13.2645	4.6966	8.2809	5.0873	5.5666	7.5162	5.1063	
RMSECV (%)	0.6522	0.5800	0.6399	0.6449	0.5191	0.5213	0.5202	0.5294	0.4302	0.4040	0.7002	0.4068	0.5034	0.4165	0.4446	0.5385	0.3732	
$BIAS_{cv}$	0.0004	-0.0203	0.0323	0.0135	0.0447	0.0466	0.0102	0.0378	0.0074	0.0312	0.0334	0.0132	0.0069	0.0141	-0.0069	0.0503	0.0232	
SECV (%)	0.6637	0.5899	0.6504	0.6562	0.5263	0.5284	0.5293	0.5374	0.4378	0.4100	0.7117	0.4137	0.5123	0.4237	0.4524	0.5456	0.3791	
RPD_{cv}	1.2042	1.3548	1.2290	1.2180	1.5185	1.5125	1.5101	1.4873	1.8258	1.9496	1.1230	1.9318	1.5603	1.8865	1.7667	1.4649	2.1084	
RER_{cv}	6.2767	7.0618	6.4059	6.3487	7.9153	7.8837	7.8715	7.7526	9.5166	10.1620	5.8534	10.0694	8.1327	9.8333	9.2087	7.6355	10.9896	
$RMSECV_{MP}$	0.6522	0.5250	0.5691	0.6135	0.5100	0.4912	0.4973	0.4683	0.4166	0.3577	0.6279	0.3825	0.4447	0.3794	0.4131	0.5104	0.3465	

Table C-33: PLSR statistics for models based on the WU scans and the ash, AIA, EIA, ASL, EXTR_PD, Carbon, Nitrogen, MAN_SRS, ARA_SRS, and moisture contents. Refer to Appendix A for descriptions of abbreviations.

Number	*1	*2	*3	*4	*5	*6	*7	*8	*9	*10	*11	*12	*13	*14	*15	*16	*17
Dataset	WU					WU	WU	WU	WU	WU	WU	WU	WU	WU			
Constituent	ASH					AIA	EIA	ASL	EXTR_PD	CARBON	N	MAN_SR	ARA_SRS	MOISTURE			
Pre. (1)	MSC	MSC	SG	SG	SG	SNVDT	SNVDT	SNVDT	SNVDT	SNVDT	EMSC	SG	SNVDT	SNV	EMSC	SNVDT	EMSC
Specific (1)	F,1.1-2.5	F,1.1-2.5	1,3,14,14	1,3,14,14	2,3,14,14	1.1-2.5,2	1.1-2.5,2	1.1-2.5,2	1.1-2.5,2	1.1-2.5,2	F,1.1-2.5	1,3,14,14	1.1-2.5,2	1.1-2.5	F,1.1-2.5	1.1-2.5,2	F,1.1-2.5
Pre. (2)	SG																
Specific (2)	1,3,14,14																
PLS-λ 10 ³ nm	0.4-2.5	1.1-2.5	0.4-2.5	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5	1.1-1.8	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5
Sam. Excl.						50019 50047	50019 50046	50019				50019	50019		50023 50040	50023 50040	
F-Wold's	1	12	4	10	2	2	3	1	1	1	2	4	1	6	9	6	6
F-Wold 0.95	1	9	4	9	2	2	3	1	1	1	1	4	1	6	5	3	5
F-Wold 0.9	1	1	4	3	2	1	1	1	1	1	1	1	1	2	4	3	3
F-F Test	1	9	4	9	2	2	3	1	1	1	2	4	1	6	9	6	5
F-Haaland's	14	11	4	8	8	1	1	1	1	5	1	3	6	4	6	4	3
F.-Min Press	17	18	20	15	10	2	3	1	1	6	2	16	7	6	11	6	9
F.-UNSCR.	5	9	4	8	8	2	3	1	1	6	2	4	7	6	7	6	5
R^2_{calib}	0.9830	0.9591	0.8936	0.9592	0.9729	0.7983	0.7904	0.6053	0.1109	0.8503	0.1499	0.5960	0.7387	0.8258	0.9254	0.9070	0.8035
$Offset_{calib}$	0.0905	0.2195	0.5669	0.2177	0.1447	0.6537	0.8634	0.8074	3.8350	6.7369	0.2084	0.0596	0.6019	7.6017	3.2451	4.0465	8.5711
RMSEC (%)	0.4514	0.7007	1.1294	0.6999	0.5704	0.9886	1.2321	0.0978	0.8126	0.6721	0.0192	0.0148	0.0589	1.5990	1.0576	1.1809	1.6979
R^2_{CV}	0.8128	0.8772	0.8261	0.9089	0.8871	0.7642	0.7133	0.5487	0.0583	0.7298	0.0990	0.2431	0.4620	0.7805	0.8717	0.8500	0.7729
$Slope_{CV}$	0.8911	0.9158	0.8400	0.9240	0.8733	0.7385	0.6932	0.5379	0.0455	0.7902	0.1071	0.3276	0.7015	0.7961	0.9161	0.8734	0.7816
$Offset_{CV}$	0.5913	0.4598	0.8942	0.4074	0.7135	0.8274	1.2059	0.9463	4.1072	9.4296	0.2190	0.0991	0.6898	8.8939	3.6322	5.4803	9.5490
RMSECV (%)	1.5310	1.2402	1.4758	1.0678	1.1888	1.1846	1.4943	0.1083	0.8652	0.9226	0.0202	0.0210	0.0876	1.8335	1.4489	1.5660	1.8654
$BIAS_{CV}$	0.0110	0.0108	0.0416	0.0022	0.0383	-0.0204	-0.0578	0.0009	-0.0100	-0.0107	0.0001	-0.0002	0.0025	-0.0023	-0.0201	-0.0281	0.0200
SECV (%)	1.5475	1.2536	1.4911	1.0793	1.2011	1.2061	1.5206	0.1102	0.8799	0.9325	0.0204	0.0214	0.0891	1.8534	1.4651	1.5835	1.8855
RPD_{CV}	2.2621	2.7925	2.3476	3.2433	2.9146	1.8587	1.8022	1.4373	0.9962	1.8830	1.0311	1.1099	1.3169	2.0892	2.6729	2.4732	2.0537
RER_{CV}	10.3715	12.8031	10.7635	14.8699	13.3628	8.4300	8.4827	6.8053	3.2691	8.3540	4.4095	3.9949	4.8514	9.4428	11.9448	11.0522	9.2819
$RMSECV_{MP}$	1.3883	1.1272	1.4258	1.0145	1.1020	1.1310	1.3918	0.1083	0.8652	0.8896	0.0198	0.0188	0.0865	1.7105	1.3219	1.4229	1.7126

Table C-34: Statistics for the best PLSR models, for the DS, DB DW, DU and WU datasets, for GLU_SRS, XYL_SRS, and TOT_SRS. Refer to Appendix A for descriptions of abbreviations.

Number	*1	*2	*3	*4	*5	*6	*7	*8	*9	*10	*11	*12	*13	*14	*15
Constituent	GLU_SRS					XYL_SRS					TOT_SRS				
Dataset	DS	DB	DW	DU	WU	DS	DB	DW	DU	WU	DS	DB	DW	DU	WU
Pre. (1)	SNVDT	SNVDT	MSC	SNVDT	EMSC	SG	SG	SG	SG	SG	SNVDT	SNVDT	NONE	EMSC	EMSC
Specific (1)	1.1-2.5,2	1.1-2.5,2	F,1.1-2.5	1.1-2.5,2	F,1.1-2.5	1,3,7,7	1,3,7,7	1,3,14,14	1,3,14,14	4,4,30,30	1.1-2.5,2	1.1-2.5,2		F,1.1-2.5	F,1.1-2.5
Pre. (2)						SNVDT	SNV	SNV	SNV						
Specific (2)						1.1-2.5,2	1.1-2.5	1.1-2.5	1.1-2.5						
PLS-λ 10 ³ nm	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.2	1.1-2.5	1.1-2.5	1.1-2.5	1.1-1.8	1.1-2.5	1.1-2.5	1.1-2.5	1.1-1.8	1.1-1.8
Sam. Excl.	50019	50019	50019	50019	50019	50019	50019 50007	50019	50019	50019	50019	50019	50019	50019	50019
F-Wold's	4	4	4	3	4	3	6	10	2	3	4	1	1	4	1
F-Wold 0.95	4	4	4	3	4	1	6	2	2	2	3	1	1	1	1
F-Wold 0.9	4	4	4	1	4	1	6	2	2	2	3	1	1	1	1
F-F Test	4	4	4	3	4	3	6	2	2	2	3	1	1	1	1
F-Haaland's	4	3	3	3	4	11	6	10	9	9	3	3	5	4	7
F.-Min Press	6	4	4	3	16	20	18	10	10	18	4	4	5	4	7
F.-UNSCR.	6	4	4	3	4	12	6	10	10	9	3	4	5	4	7
R^2_{calib}	0.9133	0.9208	0.9223	0.9384	0.9405	0.9873	0.9363	0.9473	0.9449	0.9565	0.8955	0.9101	0.9545	0.9038	0.9666
$Offset_{calib}$	3.5255	3.2002	3.1370	2.4897	2.4011	0.2886	1.4450	1.1941	1.2483	0.9849	6.9705	5.9613	3.0181	6.3794	2.2163
RMSEC (%)	0.4440	0.5336	0.5283	0.4707	0.4623	0.0661	0.1980	0.1803	0.1844	0.1638	0.6037	0.7604	0.5411	0.7867	0.4637
R^2_{cv}	0.8693	0.8844	0.8734	0.8654	0.9014	0.5929	0.7434	0.6803	0.3330	0.7499	0.8547	0.8444	0.9208	0.8086	0.9034
$Slope_{cv}$	0.8733	0.8755	0.8620	0.8934	0.9000	0.6782	0.6699	0.6847	0.4516	0.7932	0.8505	0.8182	0.8626	0.7738	0.9075
$Offset_{cv}$	5.1538	5.0234	5.5900	4.2821	4.0459	7.3109	7.4740	7.1476	12.4148	4.6966	9.9478	12.0811	9.0925	15.0323	6.2181
RMSECV (%)	0.5661	0.6676	0.6986	0.7204	0.6164	0.3889	0.4602	0.4599	0.6643	0.4068	0.7393	1.0365	0.7396	1.1493	0.8168
$BIAS_{cv}$	0.0028	-0.0058	0.0166	-0.0216	0.0083	-0.0049	-0.0126	0.0072	-0.0049	0.0132	-0.0218	0.0201	-0.0228	0.0297	0.0807
SECV (%)	0.5769	0.6793	0.7107	0.7329	0.6273	0.3962	0.4685	0.4680	0.6760	0.4137	0.7531	1.0546	0.7523	1.1693	0.8272
RPD_{cv}	2.6634	2.8400	2.7146	2.6326	3.0756	1.5094	1.7055	1.7078	1.1823	1.9318	2.5271	2.4479	3.4315	2.2079	3.1211
RER_{cv}	11.5609	12.5205	11.9679	11.6061	13.5594	6.0934	8.8933	8.9020	6.1624	10.0694	11.7234	12.0077	16.8321	10.8301	15.3095
$RMSECV_{MP}$	0.5161	0.6189	0.6343	0.7204	0.5633	0.3534	0.4150	0.4599	0.6368	0.3825	0.7345	0.9435	0.7396	1.1493	0.8168

Table C-35: Statistics for the best PLSR models, for the DS, DB DW, DU and WU datasets, for ARA_SRS, GAL_SRS, and MAN_SRS. Refer to Appendix A for descriptions of abbreviations.

Number	*1	*2	*3	*4	*5	*6	*7	*8	*9	*10	*11	*12	*13	*14	*15	
Constituent	ARA_SRS					GAL_SRS					MAN_SRS					
Dataset	DS	DB	DW	DU	WU	DS	DB	DW	DU	WU	DS	DB	DW	DU	WU	
Pre. (1)	NONE	NONE	NONE	SNVDT	SNVDT	SG	SG	NONE	N/A	SG	SG	SG	SG	SG	SG	
Specific (1)				1.1-2.5,2	1.1-2.5,2	1,3,14,14	4,4,30,30			1,3,14,14	1,3,14,14	1,3,14,14	1,3,14,14	1,3,14,14	1,3,14,14	1,3,14,14
Pre. (2)																
Specific (2)																
PLS- λ 10 ³ nm	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5	1.1-1.8	1.1-2.5	1.1-1.8	1.1-2.5			1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5
Sam. Excl.	50019	50019	50019	50019	50019	50019	50019	50019			50004 50019	50019 50013	50019	50019	50019	50019
F-Wold's	4	2	1	4	1	4	4	2			2	6	5	1	1	4
F-Wold 0.95	4	2	1	1	1	4	1	1			1	2	1	1	1	4
F-Wold 0.9	1	1	1	1	1	4	1	1			1	1	1	1	1	1
F-F Test	4	2	1	1	1	4	1	1			1	6	1	1	1	4
F-Haaland's	7	7	4	7	6	4	3	15			9	5	4	5	8	3
F.-Min Press	8	8	17	8	7	7	4	16			10	6	5	20	14	16
F.-UNSCR.	8	8	6	8	7	4	4	16			10	5	5	17	14	4
R^2_{calib}	0.8500	0.8454	0.4514	0.8565	0.7387	0.8824	0.5955	0.9803			0.9302	0.8587	0.7205	0.6854	0.8398	0.5960
Offset _{calib}	0.3453	0.3561	1.2633	0.3304	0.6019	0.1003	0.3441	0.0158			0.0592	0.0209	0.0413	0.0465	0.0237	0.0596
RMSEC (%)	0.0438	0.0453	0.0854	0.0437	0.0589	0.0311	0.0602	0.0133			0.0252	0.0081	0.0123	0.0131	0.0093	0.0148
R^2_{cv}	0.5489	0.4442	0.2107	0.4771	0.4620	0.7512	0.3728	0.7001			0.3304	0.5835	0.2703	0.0852	0.2950	0.2431
Slope _{cv}	0.7712	0.6270	0.3407	0.6680	0.7015	0.7933	0.4368	0.6768			0.5011	0.6094	0.3683	0.3406	0.5294	0.3276
Offset _{cv}	0.5132	0.8516	1.5242	0.7551	0.6898	0.1736	0.4774	0.2737			0.4208	0.0567	0.0929	0.0986	0.0710	0.0991
RMSECV (%)	0.0790	0.0890	0.1061	0.0863	0.0876	0.0470	0.0776	0.0537			0.0806	0.0150	0.0206	0.0231	0.0203	0.0210
BIAS _{cv}	-0.0135	-0.0072	0.0059	-0.0094	0.0025	-0.0027	-0.0017	-0.0013		-0.0025	-0.0009	-0.0003	0.0013	0.0016	-0.0002	
SECV (%)	0.0793	0.0903	0.1078	0.0873	0.0891	0.0478	0.0790	0.0546		0.0820	0.0152	0.0210	0.0235	0.0206	0.0214	
RPD _{cv}	1.4551	1.2994	1.0884	1.3433	1.3169	1.9338	1.2194	1.7635		1.1856	1.4499	1.1305	1.0110	1.1533	1.1099	
RER _{cv}	5.4518	4.7869	4.0097	4.9486	4.8514	6.3330	3.8317	5.5414		3.6907	5.3573	4.0689	3.6390	4.1512	3.9949	
RMSECV _{MP}	0.0758	0.0865	0.1006	0.0778	0.0865	0.0466	0.0769	0.0485		0.0791	0.0146	0.0193	0.0204	0.0185	0.0188	

Table C-36: Statistics for the best PLSR models, for the DS, DB DW, DU and WU datasets, for KL, ASL, and AIR. Refer to Appendix A for descriptions of abbreviations.

Number	*1	*2	*3	*4	*5	*6	*7	*8	*9	*10	*11	*12	*13	*14	*15
Constituent	KL					ASL					AIR				
Dataset	DS	DB	DW	DU	WU	DS	DB	DW	DU	WU	DS	DB	DW	DU	WU
Pre. (1)	SNVDT	SNVDT	SNVDT	SNVDT	SNVDT	SNVDT	SNVDT	SNVDT	SNVDT	SNVDT	SNVDT	SNVDT	SNVDT	SNVDT	SNVDT
Specific (1)	1.1-2.5,2	1.1-2.5,2	1.1-2.5,2	1.1-2.5,2	1.1-2.5,2	1.1-2.5,2	1.1-2.5,2	1.1-2.5,2	1.1-2.5,2	1.1-2.5,2	2,1.1-2.5	1.1-2.5,2	1.1-2.5,2	1.1-2.5,2	1.1-2.5,2
Pre. (2)															
Specific (2)															
PLS- λ 10 ³ nm	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5
Sam. Excl.	50019	50019	50019	50019 50042	50019	50019	50019	50019	50019	50019	50019 50037 50047	50019 50037 50047	50019 50037 50047	50019 50037 50047	50019 50037 50047
F-Wold's	6	7	1	1	5	1	3	1	3	1	4	4	3	3	3
F-Wold 0.95	6	6	1	1	1	1	1	1	3	1	1	3	3	3	3
F-Wold 0.9	6	1	1	1	1	1	1	1	3	1	1	3	3	3	3
F-F Test	6	6	1	1	2	1	1	1	3	1	1	3	3	3	3
F-Haaland's	5	10	10	9	12	1	3	3	2	1	4	2	3	5	3
F.-Min Press	6	13	12	9	13	4	3	3	3	1	11	4	3	5	5
F.-UNSCR.	6	11	12	9	13	3	3	3	3	1	4	3	3	5	5
R^2_{catib}	0.8958	0.9471	0.8419	0.9482	0.9565	0.3094	0.7294	0.7098	0.6336	0.6053	0.9281	0.8769	0.9060	0.9443	0.8981
$Offset_{catib}$	1.9173	0.9738	2.9083	0.9527	0.7980	1.4273	0.5537	0.6014	0.7496	0.8074	1.5289	2.6537	2.0273	1.1999	2.1971
RMSEC (%)	0.2126	0.1552	0.2682	0.1562	0.1407	0.1086	0.0810	0.0858	0.0942	0.0978	0.3506	0.6583	0.5754	0.4427	0.5990
R^2_{CV}	0.7470	0.7839	0.2327	0.7745	0.6578	0.1912	0.6301	0.5493	0.5721	0.5487	0.8743	0.8444	0.8311	0.8620	0.8387
$Slope_{CV}$	0.8439	0.8554	0.6492	0.8501	0.8100	0.2283	0.6291	0.5738	0.5890	0.5379	0.8927	0.7978	0.8251	0.8625	0.8271
$Offset_{CV}$	2.8566	2.6689	6.4915	2.7663	3.4759	1.5953	0.7620	0.8725	0.8376	0.9463	2.2851	4.3222	3.7990	3.0104	3.6966
RMSECV (%)	0.3440	0.3248	0.6120	0.3259	0.4087	0.1220	0.0980	0.1070	0.1054	0.1083	0.4437	0.8016	0.9014	0.7723	0.8078
$BIAS_{CV}$	-0.0159	0.0096	0.0416	0.0089	-0.0177	0.0006	0.0031	-0.0020	-0.0033	0.0009	0.0027	-0.0388	0.0278	0.0449	-0.0321
SECV (%)	0.3502	0.3304	0.6214	0.3317	0.4156	0.1243	0.0997	0.1091	0.1073	0.1102	0.4528	0.8159	0.9182	0.7857	0.8226
RPD_{CV}	1.9165	2.0779	1.1048	2.1071	1.6520	1.0708	1.5883	1.5214	1.4768	1.4373	2.9476	2.3438	2.0826	2.4337	2.3247
RER_{CV}	7.7334	8.1959	4.3577	8.1632	6.5159	5.6632	7.5201	6.8716	6.9921	6.8053	12.0095	11.0625	9.8299	11.4868	10.9723
$RMSECV_{MP}$	0.3050	0.3049	0.5528	0.3259	0.3812	0.1174	0.0980	0.1070	0.0974	0.1083	0.4062	0.7416	0.9014	0.7723	0.7736

Table C-37: Statistics for the best PLSR models, for the DS, DB DW, DU and WU datasets, for ASH, EIA, and AIA. Refer to Appendix A for descriptions of abbreviations.

Number	*1	*2	*3	*4	*5	*6	*7	*8	*9	*10	*11	*12	*13	*14	*15
Constituent	ASH					EIA					AIA				
Dataset	DS	DB	DW	DU	WU	DS	DB	DW	DU	WU	DS	DB	DW	DU	WU
Pre. (1)	MSC	MSC	MSC	MSC	SG	SNVDT	SNVDT	SNVDT	SNVDT	SNVDT	SNVDT	SNVDT	NONE	SNVDT	SNVDT
Specific (1)	F,1.1-2.5	F,1.1-2.5	F,1.1-2.5	F,1.1-2.5	1,3,14,14	1.1-2.5,2	1.1-2.5,2	1.1-2.5,2	1.1-2.5,2	1.1-2.5,2	2,1.1-2.5	1.1-2.5,2		1.1-2.5,2	1.1-2.5,2
Pre. (2)	SG	SG	SG	SG											
Specific (2)	1,3,14,14	1,3,14,14	1,3,14,14	1,3,14,14											
PLS- λ 10 ³ nm	0.4-2.5	0.4-2.5	0.4-2.5	0.4-2.5	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5
Sam. Excl.			50020			50019	50019 50046	50019 50046	50019 50046	50019 50046	50019 50047	50019 50047	50019 50047	50019 50037	50019 50037
F-Wold's	4	5	1	8	10	1	2	3	2	3	1	2	3	2	2
F-Wold 0.95	4	1	1	5	9	1	2	3	2	3	1	2	1	2	2
F-Wold 0.9	3	1	1	2	3	1	2	3	2	1	1	2	1	2	1
F-F Test	4	5	1	5	9	1	2	3	2	3	1	2	1	2	2
F-Haaland's	3	5	7	6	8	1	2	3	6	1	1	2	5	2	1
F.-Min Press	4	18	10	8	15	6	4	3	6	3	1	2	5	9	2
F.-UNSCR.	4	5	9	4	8	1	4	3	6	3	1	2	5	6	2
R^2_{calib}	0.8942	0.9347	0.9456	0.9555	0.9592	0.7759	0.8645	0.8418	0.9569	0.7904	0.8787	0.8728	0.9042	0.8161	0.7983
$Offset_{calib}$	0.4658	0.3481	0.2835	0.4636	0.2177	0.8402	0.5580	0.6516	0.1767	0.8634	0.3592	0.4125	0.2944	0.5525	0.6537
RMSEC (%)	0.6918	0.8851	0.7939	0.8096	0.6999	0.8773	0.9905	1.0704	0.5586	1.2321	0.5686	0.7853	0.6376	0.8815	0.9886
R^2_{cv}	0.8219	0.8593	0.8320	0.8697	0.9089	0.7481	0.7987	0.7310	0.8811	0.7133	0.8414	0.8598	0.7683	0.7624	0.7642
$Slope_{cv}$	0.8049	0.8778	0.8649	0.8613	0.9240	0.7408	0.7759	0.7460	0.8829	0.6932	0.8517	0.8094	0.7446	0.7250	0.7385
$Offset_{cv}$	0.8628	0.6149	0.6923	0.7949	0.4074	0.9578	0.8528	1.0859	0.3907	1.2059	0.4149	0.5916	0.7357	0.7553	0.8274
RMSECV (%)	0.9308	1.3275	1.4718	1.3399	1.0678	0.9674	1.2520	1.4473	0.9621	1.4943	0.6355	0.9105	1.0975	1.0945	1.1846
$BIAS_{cv}$	0.0043	-0.0364	-0.0117	-0.0760	0.0022	-0.0143	-0.0703	0.0397	-0.0916	-0.0578	-0.0242	-0.0263	-0.0494	-0.0708	-0.0204
SECV (%)	0.9478	1.3413	1.4880	1.3632	1.0793	0.9865	1.2729	1.4733	0.9753	1.5206	0.6477	0.9268	1.1173	1.1122	1.2061
RPD_{cv}	2.2847	2.6098	2.3126	3.0893	3.2433	1.9158	2.1528	1.8600	2.8098	1.8022	2.5702	2.4188	1.8787	1.8819	1.8587
RER_{cv}	8.4328	11.9656	10.7859	11.3624	14.8699	7.0471	10.1328	8.7547	13.2253	8.4827	10.3268	10.9705	9.1001	9.1414	8.4300
$RMSECV_{MP}$	0.8924	1.2547	1.3530	1.1757	1.0145	0.8665	1.1611	1.4473	0.9621	1.3918	0.6355	0.9105	1.0975	0.9769	1.1310

Table C-38: Statistics for the best PLSR models, for the DS, DB DW, DU and WU datasets, for the EXTR_PD and carbon contents. Refer to Appendix A for descriptions of abbreviations.

Number	*1	*2	*3	*4	*5	*6	*7	*8	*9
Constituent	EXTR_PD					CARBON			
Dataset	DS	DB	DW	DU	WU	DB	DW	DU	WU
Pre. (1)	SNVDT								
Specific (1)	3,1.1-2.5	1.1-2.5,3	1.1-2.5,2	1.1-2.5,2	1.1-2.5,2	1.1-2.5,2	1.1-2.5,2	1.1-2.5,2	1.1-2.5,2
Pre. (2)									
Specific (2)									
PLS- λ 10 ³ nm	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5	0.4-2.5	0.4-2.5	1.1-2.5	1.1-2.5
Sam. Excl.	50025								
F-Wold's	2	3	2	1	1	3	4	3	1
F-Wold 0.95	2	1	1	1	1	1	4	1	1
F-Wold 0.9	2	1	1	1	1	1	1	1	1
F-F Test	2	1	1	1	1	2	4	3	1
F-Haaland's	2	1	14	13	1	1	4	1	5
F.-Min Press	2	3	14	14	1	6	4	3	6
F.-UNSCR.	2	1	14	14	1	1	4	3	6
R^2_{calib}	0.8199	0.2346	0.9492	0.9217	0.1109	0.7104	0.8149	0.6874	0.8503
$Offset_{calib}$	0.7707	3.3017	0.2177	0.3395	3.8350	13.0338	8.3272	14.0660	6.7369
RMSEC (%)	0.3669	0.7540	0.1974	0.2412	0.8126	0.9349	0.7473	0.9712	0.6721
R^2_{cv}	0.7932	0.1920	0.5871	0.2113	0.0583	0.6771	0.7200	0.6725	0.7298
$Slope_{cv}$	0.7765	0.1816	0.8545	0.4635	0.0455	0.6697	0.7433	0.6696	0.7902
$Offset_{cv}$	0.9547	3.5245	0.6382	2.3764	4.1072	14.8803	11.5207	14.9119	9.4296
RMSEC_{cv} (%)	0.4083	0.8014	0.5832	0.7918	0.8652	1.0085	0.9391	1.0157	0.9226
$BIAS_{cv}$	-0.0017	-0.0057	0.0098	0.0620	-0.0100	0.0175	-0.0291	0.0461	-0.0107
SECV (%)	0.4160	0.8151	0.5934	0.8028	0.8799	1.0193	0.9488	1.0256	0.9325
RPD_{cv}	2.1173	1.0754	1.5027	1.0919	0.9962	1.7227	1.8506	1.7120	1.8830
RER_{cv}	6.9140	3.5291	4.8473	3.5831	3.2691	7.6427	8.2100	7.5955	8.3540
$RMSEC_{MP}$	0.4083	0.7959	0.5832	0.7497	0.8652	0.9863	0.9391	0.9294	0.8896

Table C-39: Summary constituent concentration statistics for the samples involved in the DS calibrations and the “other” (DB, DW, DU, WU) calibrations. All values expressed in % whole dry mass basis. Note that, in contrast to the statistics associated with the histograms in Table C-3 and Table C-4, sample 50017 is included in the calculation of the statistics for the “Other” category. This sample was excluded from the WM and AF statistics (but included in the EF statistics) for the lignocellulosic constituents (sugars, KL, ASL) in Table C-3 and Table C-4 because no EXTR_PD data were obtained. In this Table the EXTR_CV data is used for the EXTR_PD value for this sample.

Constituent	Dataset	N	AV	SD	MAX	MIN	RANGE	KURT	SKEW	SEL	SEL/AV
GLU_SRS	DS	27	40.65	1.54	42.55	35.89	6.67	2.27	-1.23	0.21	0.51
	Others	29	40.39	1.93	42.55	34.05	8.51	3.60	-1.67	0.20	0.50
XYL_SRS	DS	27	22.74	0.60	23.96	21.55	2.41	-0.53	-0.02	0.14	0.60
	Others	29	22.65	0.80	23.96	19.80	4.17	4.69	-1.56	0.14	0.64
GAL_SRS	DS	27	0.85	0.09	0.99	0.68	0.30	-0.95	-0.43	0.01	1.56
	Others	29	0.85	0.10	0.99	0.68	0.30	-1.07	-0.42	0.01	1.53
ARA_SRS	DS	27	2.30	0.12	2.50	2.07	0.43	-0.40	-0.01	0.03	1.20
	Others	29	2.30	0.12	2.50	2.07	0.43	-0.62	0.00	0.03	1.17
RHA_SRS	DS	27	0.11	0.01	0.13	0.09	0.04	0.04	-0.61	0.00	4.18
	Others	29	0.11	0.01	0.13	0.09	0.04	0.10	-0.71	0.00	4.12
MAN_SRS	DS	27	0.15	0.02	0.19	0.11	0.09	-0.54	0.38	0.02	13.30
	Others	29	0.15	0.02	0.19	0.11	0.09	-0.60	0.37	0.02	12.73
TOT_SRS	DS	27	66.69	1.90	69.56	60.73	8.83	2.34	-0.91	0.30	0.45
	Others	29	66.34	2.58	69.56	56.90	12.66	5.95	-1.99	0.29	0.44
KL	DS	27	18.40	0.67	19.65	16.94	2.71	0.13	-0.39	0.31	1.68
	Others	29	18.39	0.69	19.65	16.94	2.71	-0.20	-0.36	0.30	1.65
ASL	DS	27	2.07	0.13	2.32	1.62	0.70	3.99	-1.15	0.10	4.73
	Others	29	2.05	0.16	2.32	1.57	0.75	3.23	-1.38	0.09	4.60
AIR	DS	27	21.28	1.54	25.06	18.24	6.83	0.93	0.83	0.28	1.31
	Others	29	21.55	2.02	28.65	18.24	10.41	4.64	1.78	0.28	1.31
AIA	DS	27	2.89	1.67	7.76	1.07	6.69	1.88	1.46	0.40	13.76
	Others	29	3.17	2.24	11.23	1.07	10.17	5.37	2.15	0.41	12.93
EIA	DS	26	3.75	1.89	8.62	1.67	6.95	1.12	1.37	0.33	13.76
	Others	28	4.12	2.74	14.57	1.67	12.90	7.15	2.43	0.35	8.59%
ASH	DS	27	4.39	2.21	9.79	1.80	7.99	0.69	1.20	0.50	11.45
	Others	47	5.33	3.50	17.77	1.72	16.05	5.15	2.20	0.47	8.91
CARBON	Others	47	45.00	1.76	46.88	39.09	7.79	3.40	-1.89	N/A	N/A
NITROGEN	Others	47	0.25	0.02	0.29	0.20	0.09	0.16	0.01	N/A	N/A
MOISTURE	Others	47	43.63	3.87	48.51	31.01	17.50	1.65	-1.28	1.81	4.15

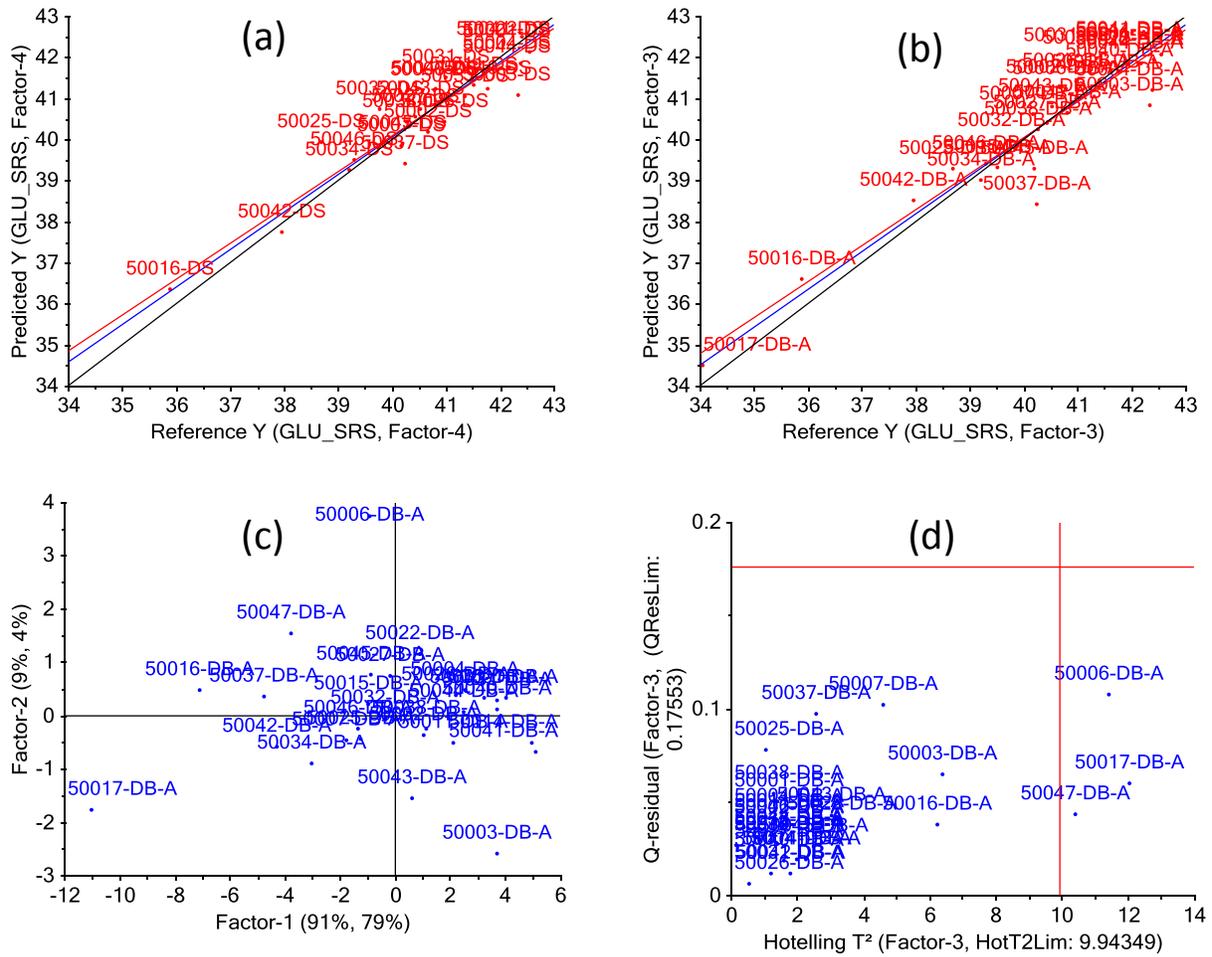


Figure C-32: Plots for GLU_SRS PLSR models. (a) Predicted y vs. reference y for the GLU_SRS constituent in a PLSR model involving the DS dataset; (b) the same as (a) but using the DB dataset; (c) a F1 vs. F2 scores plot for the DB GLU_SRS calibration; (d) an influence plot for this 3 Factor model.

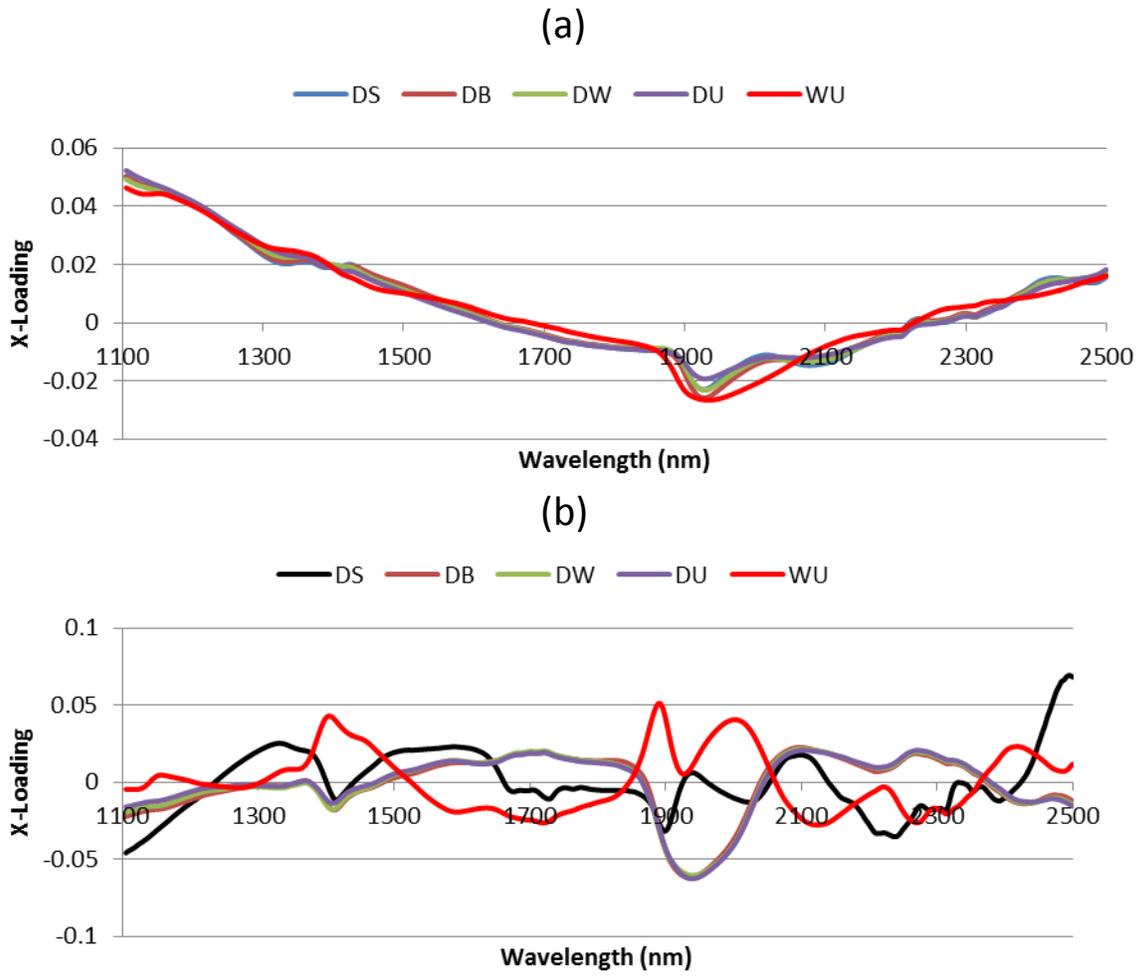


Figure C-33: X-loadings plots for each of the GLU_SRS models for the 5 datasets. (a) Factor 1, (b) factor 2

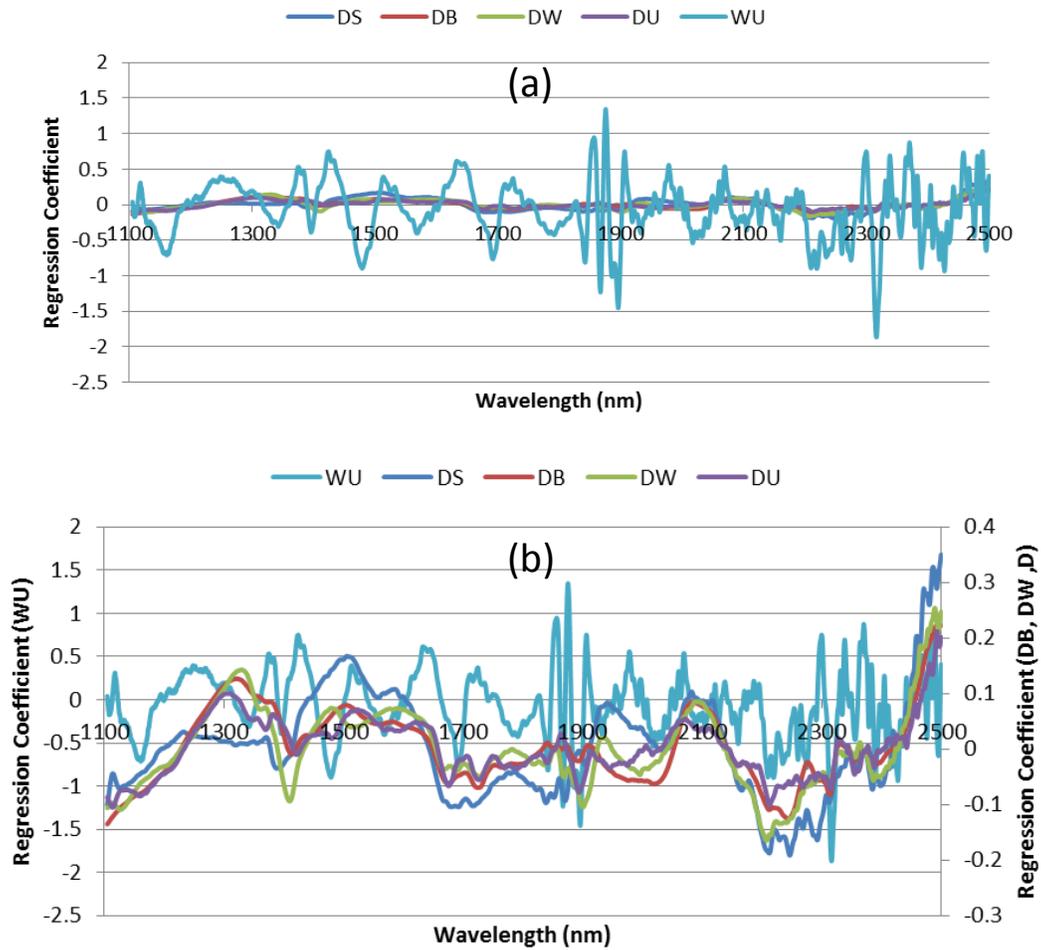


Figure C-34: Regression coefficients plots for the same models as in Figure C-33: (a) all 5 datasets using the same y-axis, (b) separate y-axes for the WU dataset and other datasets.

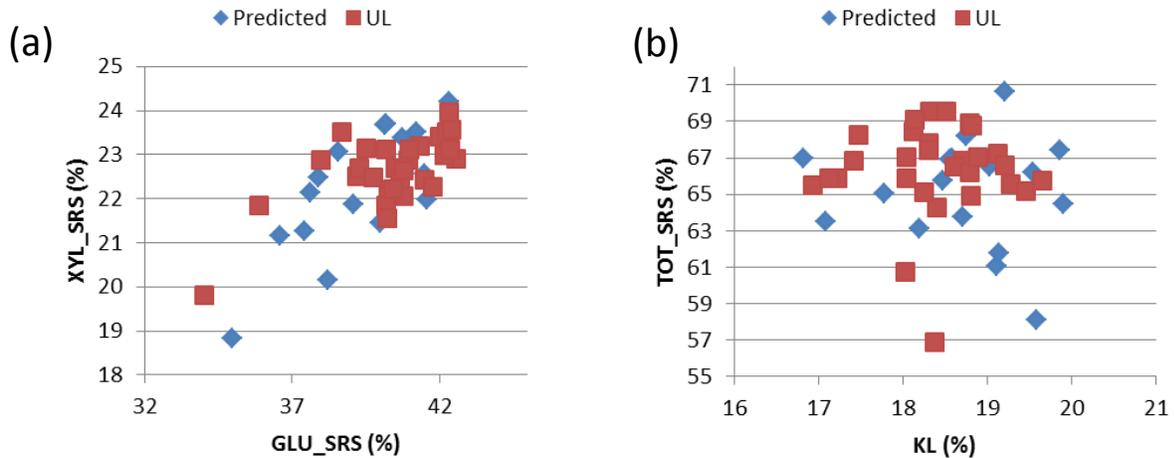


Figure C-35: Plots involving the samples analysed at UL (squares) and the predicted constituent concentrations of the 18 samples not analysed at UL (diamonds). (a) a XYL_SRS vs. GLU_SRS plot; (b) a plot involving the TOT_SRS and KL contents.

Appendix D Figures and Tables for Chapter 13: Peat

Table D-1: Summary of the peat samples collected by the Author and supplied by Bord na Mona.

NIR	ID	Humification	Date Obtained	Bog	Power Station
25001	PTHG4	High	19/1/09		Lanesborough
25002	PTHG5	High	19/1/09		Lanesborough
25003	PTHG3	High	19/1/09		Lanesborough
25004	PTLW3	Low	19/1/09		Lanesborough
25005	PTMM3	Medium	19/1/09		Lanesborough
25006	PTMM4	Medium	19/1/09	Bloomhill	Shannonbridge
25007	PTMH2	Med. to High	19/1/09		Shannonbridge
25008	PTLW4	Low	19/1/09	Roscommon	Shannonbridge
25009	PTMM5	Medium	19/1/09	Derrygadda	Shannonbridge
25010	PTLW1	Low	19/1/09	Clonad	Edenderry
25011	PTMM1	Medium	19/1/09	Ballydermot	Edenderry
25012	PTHG1	High	19/1/09	Cloncreen	Edenderry
25013	PTHG2	High	19/1/09	Cloncreen	Edenderry
25014	PTMM2	Medium	19/1/09	Ballydermot	Edenderry
25015	PTLW2	Low	19/1/09	Clonad	Edenderry
25016	PTLW5	Low	09/02/09	Prosperous	Horticultural peat
25017	PTLW6	Low	09/02/09	Kilberry	Horticultural peat (very good quality)
25018	PTLW7	Low	19/1/09	Bunahinly	Blackwater/Boora – West Offaly Power
25019	PTLW8	Low	09/02/09	Ballivor	Horticultural peat (good quality)
25020	PTLW9	Low	10/02/09	Coolnamona	Blackwater/Boora – West Offaly Power
25021	PTML1	Low to Med	09/02/09	Lanes	Littleton
25022	PTML2	Low to Med	09/02/09	Near Leigh	Littleton
25023	PTML3	Low to Med	09/02/09	Cul na Mona	Horticulture peat
25024	PTMM6	Medium	09/02/09	Prosperous	Horticultural peat (poor quality)
25025	PTMM7	Medium	09/02/09	Attymon	West Offaly Power
25026	PTXX1	Unknown	March 2010	Littleton	
25027	PTXX2	Unknown	March 2010	Littleton	
25028	PTXX3	Unknown	March 2010	Littleton	
25029	PTXX4	Unknown	March 2010	Littleton	
25030	PTHG6	High	March 2010		Edenderry
25031	PTLW10	Low	March 2010		Edenderry
25032	PTMM8	Medium	March 2010		Edenderry
25033	PTXX5	Unclassified	March 2010	Cual na Gun	Edenderry
25034	PTXX6	Unclassified	March 2010	Ballivor, Bog 1	Horticultural peat
25035	PTXX7	Unclassified	March 2010	Ballivor, Bog 2	
25036	PTXX8	Unclassified	March 2010	Ballivor, Bog 3	
25037	PTXX9	Unclassified	March 2010	Drumman	Lanesborough
25038	PTXX10	Unclassified	March 2010	Ballybeg	Lanesborough
25039	PTXX11	Unclassified	March 2010	Cavemount	Lanesborough
25040	PTXX12	Unclassified	March 2010	Esker	Lanesborough
25041	PTXX13	Unclassified	March 2010	Blackriver	Lanesborough
25042	PTXX14	Unclassified	March 2010	Lullymoor	Lanesborough
25043	PTHG7	High	March 2010	Cloncreen	Lanesborough
25044	PTMM9	Medium	March 2010	Ballycane	Lanesborough
25045	PTLW11	Low	March 2010	Clonad	Lanesborough
25046	PTXX15	Unclassified	March 2010		Obtained from BNM, Newbridge
25047	PTXX16	Unclassified	March 2010		Obtained from BNM, Newbridge
25048	PTXX17	Unclassified	March 2010		Obtained from BNM, Newbridge
25049	PTXX18	Unclassified	March 2010	Attymon (Galway)	
25050	PTLW12	Low	March 2010		Obtained from BNM, Newbridge
25051	PTLW13	Low	March 2010		Obtained from BNM, Newbridge
25052	PTMM10	Medium	March 2010		Obtained from BNM, Newbridge
25053	PTHG8	High	March 2010		Obtained from BNM, Newbridge

Table D-2: Extractives-free data (% DM, extractives-free basis) including standard deviation of the duplicates (SD) for peat samples 25001-25037.

NIR #	UNIV. CODE	EIA_EF		AIR_EF		AIA_EF		KL_EF		ASL_EF		ARA_EF_SRS		GAL_EF_SRS		RHA_EF_SRS		GLU_EF_SRS		XYL_EF_SRS		MAN_EF_SRS		TOT_EF_SRS	
		AV	SD	AV	SD	AV	SD	AV	SD	AV	SD	AV	SD	AV	SD	AV	SD	AV	SD	AV	SD	AV	SD	AV	SD
25001	PTHG4	5.95	(0.29)	67.37	(0.29)	3.02	(0.06)	64.35	(0.23)	1.79	(0.06)	0.47	(0.00)	1.71	(0.00)	0.68	(0.01)	10.18	(0.08)	2.47	(0.00)	1.39	(0.01)	16.22	(0.10)
25002	PTHG5	8.20	(0.77)	58.29	(0.36)	4.78	(0.06)	53.50	(0.29)	2.53	(0.05)	0.39	(0.00)	2.13	(0.01)	1.09	(0.01)	14.05	(0.01)	2.58	(0.02)	1.93	(0.03)	21.08	(0.04)
25003	PTHG3	2.38	(0.00)	63.50	(0.30)	0.09	(0.08)	63.40	(0.38)	2.09	(0.05)	0.20	(0.00)	1.81	(0.02)	0.91	(0.03)	14.20	(0.07)	2.61	(0.03)	1.90	(0.01)	20.71	(0.12)
25004	PTLW3	2.62	(0.03)	49.84	(0.34)	1.02	(0.35)	48.82	(0.01)	3.37	(0.07)	0.39	(0.01)	2.73	(0.04)	1.45	(0.04)	16.68	(0.22)	3.15	(0.03)	2.30	(0.05)	25.24	(0.33)
25005	PTMM3	2.17		58.86	(0.24)	0.32	(0.04)	58.54	(0.28)	2.40	(0.03)	0.31	(0.00)	2.39	(0.03)	1.12	(0.03)	15.84	(0.21)	2.82	(0.06)	2.05	(0.05)	23.41	(0.34)
25006	PTMM4	2.85	(0.07)	70.46	(0.16)	0.16	(0.03)	70.30	(0.13)	1.21	(0.01)	0.27	(0.00)	1.65	(0.02)	0.52	(0.01)	11.70	(0.24)	1.85	(0.02)	1.37	(0.01)	16.85	(0.29)
25007	PTMH2	4.62	(0.09)	68.65	(0.34)	0.42	(0.06)	68.23	(0.28)	1.73	(0.10)	0.33	(0.00)	1.49	(0.02)	0.58	(0.03)	10.20	(0.30)	1.96	(0.02)	1.28	(0.05)	15.26	(0.39)
25008	PTLW4	1.69	(0.01)	60.53	(0.60)	0.37	(0.07)	60.16	(0.67)	2.27	(0.13)	0.27	(0.00)	2.20	(0.05)	1.01	(0.04)	14.29	(0.36)	2.51	(0.08)	2.09	(0.06)	21.34	(0.54)
25009	PTMM5	3.00	(0.01)	60.70	(0.45)	0.83	(0.05)	59.87	(0.40)	2.41	(0.08)	0.33	(0.01)	2.03	(0.04)	0.94	(0.05)	14.10	(0.29)	2.45	(0.04)	1.88	(0.05)	20.78	(0.41)
25010	PTLW1	3.03	(0.30)	50.13	(1.01)	1.11	(0.18)	49.02	(0.83)	3.21	(0.12)	0.27	(0.00)	2.63	(0.04)	1.61	(0.09)	18.71	(0.44)	3.14	(0.04)	2.67	(0.07)	27.42	(0.60)
25011	PTMM1	3.27	(0.11)	60.23	(0.13)	1.35	(0.05)	58.88	(0.09)	2.29	(0.00)	0.26	(0.00)	1.96	(0.00)	0.97	(0.02)	16.01	(0.26)	2.45	(0.04)	1.93	(0.00)	22.61	(0.29)
25012	PTHG1	12.30	(0.63)	73.87	(0.12)	3.38	(0.18)	70.49	(0.06)	1.74	(0.02)	0.40	(0.01)	0.87	(0.00)	0.38	(0.00)	4.42	(0.04)	1.52	(0.01)	0.71	(0.01)	7.93	(0.03)
25013	PTHG2	11.84	(1.25)	73.76	(0.24)	4.13	(0.05)	69.63	(0.29)	1.59	(0.03)	0.39	(0.01)	0.84	(0.01)	0.36	(0.00)	4.55	(0.11)	1.59	(0.04)	0.71	(0.01)	8.08	(0.17)
25014	PTMM2	3.17	(0.38)	60.12	(0.01)	1.15	(0.23)	58.97	(0.22)	2.31	(0.12)	0.27	(0.00)	1.91	(0.01)	0.85	(0.02)	15.35	(0.28)	2.29	(0.06)	1.82	(0.02)	21.64	(0.37)
25015	PTLW2	2.29	(0.49)	49.43	(0.32)	0.80	(0.16)	48.63	(0.16)	3.44	(0.01)	0.28	(0.01)	2.58	(0.03)	1.48	(0.05)	17.86	(0.42)	3.01	(0.07)	2.50	(0.06)	26.22	(0.56)
25016	PTLW5	6.27	(0.99)	51.04	(0.90)	4.73	(1.15)	46.31	(0.25)	2.84	(0.02)	0.40	(0.01)	2.72	(0.02)	1.39	(0.00)	17.95	(0.05)	2.96	(0.01)	2.23	(0.01)	26.26	(0.09)
25017	PTLW6	3.60	(0.02)	52.44	(0.29)	1.36	(0.36)	51.07	(0.65)	2.98	(0.05)	0.36	(0.02)	2.51	(0.07)	1.46	(0.06)	16.58	(0.14)	2.98	(0.07)	2.16	(0.00)	24.59	(0.30)
25018	PTLW7	2.66	(0.02)	54.62	(0.32)	1.78	(0.23)	52.84	(0.08)	3.15	(0.04)	0.61	(0.02)	2.73	(0.06)	1.14	(0.03)	15.24	(0.42)	2.96	(0.07)	2.22	(0.07)	23.76	(0.64)
25019	PTLW8	1.79		58.98	(0.48)	0.42	(0.06)	58.56	(0.55)	2.02	(0.05)	0.32	(0.00)	2.38	(0.02)	0.98	(0.03)	17.23	(0.34)	3.07	(0.07)	2.01	(0.03)	25.01	(0.46)
25020	PTLW9	2.35	(0.04)	68.79	(0.56)	0.75	(0.16)	68.05	(0.40)	1.73	(0.01)	0.28	(0.00)	1.64	(0.00)	0.70	(0.02)	11.96	(0.07)	2.30	(0.01)	1.26	(0.01)	17.44	(0.07)
25021	PTLM1	15.17	(0.20)	73.75	(0.66)	7.89	(1.66)	65.86	(1.00)	1.56	(0.01)	0.98	(0.01)	1.14	(0.02)	0.50	(0.00)	5.36	(0.17)	1.64	(0.07)	0.67	(0.01)	9.79	(0.27)
25022	PTLM2	7.82	(0.34)	70.09	(0.18)	3.50	(0.06)	66.59	(0.24)	1.47	(0.02)	0.38	(0.01)	1.54	(0.03)	0.65	(0.01)	10.68	(0.10)	1.85	(0.09)	1.26	(0.02)	15.71	(0.25)
25023	PTLM3	2.00	(0.05)	63.52	(0.27)	0.32	(0.09)	63.20	(0.36)	1.78	(0.03)	0.23	(0.02)	2.15	(0.12)	0.85	(0.22)	14.09	(0.03)	2.27	(0.04)	1.95	(0.03)	20.69	(0.10)
25024	PTMM6	3.85	(0.59)	68.26	(0.13)	0.86	(0.11)	67.40	(0.23)	1.71	(0.03)	0.37	(0.01)	1.47	(0.01)	0.59	(0.04)	10.83	(0.46)	2.47	(0.17)	1.33	(0.03)	16.46	(0.61)
25025	PTMM7	2.46	(0.01)	70.16	(0.47)	0.45	(0.12)	69.71	(0.35)	1.13	(0.03)	0.27	(0.00)	1.61	(0.02)	0.49	(0.02)	12.05	(0.04)	1.58	(0.04)	1.40	(0.01)	16.91	(0.04)
25026	PTXX1	7.98	(0.01)	64.51	(0.33)	1.80	(0.26)	62.71	(0.07)	2.20	(0.03)	1.39	(0.02)	1.77	(0.00)	0.94	(0.02)	6.44	(0.06)	2.42	(0.02)	1.53	(0.01)	13.56	(0.07)
25027	PTXX2	2.16	(0.11)	57.83	(0.58)	0.56	(0.21)	57.26	(0.37)	2.46	(0.04)	0.27	(0.01)	2.08	(0.05)	1.05	(0.04)	15.49	(0.15)	2.76	(0.00)	1.96	(0.03)	22.56	(0.24)
25028	PTXX3	1.53	(0.04)	59.71	(0.28)	0.47	(0.02)	59.24	(0.30)	2.44	(0.02)	0.26	(0.00)	2.25	(0.01)	0.97	(0.01)	15.38	(0.36)	2.67	(0.05)	2.08	(0.06)	22.65	(0.47)
25029	PTXX4	1.94	(0.12)	65.04	(1.24)	0.44	(0.16)	64.60	(1.09)	1.67	(0.01)	0.25	(0.00)	1.81	(0.04)	0.79	(0.03)	14.71	(0.47)	2.49	(0.06)	1.74	(0.05)	21.00	(0.62)
25030	PTHG6	5.34	(0.42)	70.75	(0.20)	1.14	(0.08)	69.60	(0.13)	1.68	(0.04)	0.35	(0.01)	1.43	(0.01)	0.59	(0.01)	10.51	(0.12)	1.94	(0.06)	1.24	(0.04)	15.47	(0.24)
25031	PTLW10	5.30	(0.21)	68.01	(0.77)	0.62	(0.09)	67.38	(0.86)	1.79	(0.00)	0.70	(0.01)	1.51	(0.02)	0.71	(0.04)	9.13	(0.25)	2.61	(0.03)	1.29	(0.06)	15.25	(0.37)
25032	PTMM8	6.13	(0.25)	69.92	(0.39)	1.52	(0.05)	68.40	(0.35)	1.71	(0.01)	0.48	(0.00)	1.49	(0.03)	0.69	(0.01)	8.56	(0.14)	1.84	(0.01)	1.35	(0.02)	13.72	(0.21)
25033	PTXX5	4.35	(0.21)	65.54	(0.06)	0.78	(0.16)	64.75	(0.11)	1.97	(0.02)	0.34	(0.01)	1.85	(0.01)	0.82	(0.01)	12.00	(0.02)	2.16	(0.02)	1.61	(0.01)	17.96	(0.03)
25034	PTXX6	1.33	(0.07)	46.11	(0.25)	0.50	(0.05)	45.61	(0.20)	3.48	(0.05)	0.26		2.89		1.80		20.92		3.37		2.85		30.29	
25035	PTXX7	3.07	(0.42)	51.75	(0.39)	1.66	(0.63)	50.08	(0.24)	2.87	(0.01)	0.30	(0.00)	2.62	(0.03)	1.42	(0.01)	19.17	(0.00)	3.29	(0.11)	2.50	(0.01)	27.88	(0.08)
25036	PTXX8	1.47	(0.02)	58.11	(0.52)	0.32	(0.08)	57.79	(0.44)	2.30	(0.02)	0.20	(0.00)	2.25	(0.02)	1.17	(0.01)	17.31	(0.33)	2.62	(0.17)	2.30	(0.03)	24.68	(0.56)
25037	PTXX9	4.57	(0.11)	60.47	(0.49)	1.20	(0.31)	59.27	(0.80)	2.29	(0.07)	0.46	(0.01)	2.04	(0.02)	1.06	(0.00)	13.53	(0.16)	2.61	(0.00)	1.93	(0.03)	20.57	(0.17)

Table D-3: Extractives-free data (% DM, extractives-free basis) including standard deviation of the duplicates (SD) for peat samples 25038-250053.

NIR #	UNIV. CODE	EIA_EF		AIR_EF		AIA_EF		KL_EF		ASL_EF		ARA_EF_SRS		GAL_EF_SRS		RHA_EF_SRS		GLU_EF_SRS		XYL_EF_SRS		MAN_EF_SRS		TOT_EF_SRS	
		AV	SD	AV	SD	AV	SD	AV	SD	AV	SD	AV	SD	AV	SD	AV	SD	AV	SD	AV	SD	AV	SD	AV	SD
25038	PTXX10	5.21	(0.08)	69.50	(0.17)	1.38	(0.28)	68.12	(0.45)	2.12	(0.03)	0.27	(0.00)	1.25	(0.01)	0.62	(0.00)	9.15	(0.12)	1.74	(0.01)	1.19	(0.02)	13.61	(0.11)
25039	PTXX11	7.93	(0.10)	74.60	(0.13)	1.48	(0.01)	73.12	(0.14)	1.53	(0.01)	0.61	(0.02)	0.97	(0.02)	0.39	(0.01)	4.73	(0.17)	1.79	(0.08)	0.70	(0.00)	8.81	(0.30)
25040	PTXX12	4.76	(0.40)	69.44	(0.24)	2.39	(0.05)	67.05	(0.19)	1.55	(0.03)	0.24	(0.00)	1.47	(0.00)	0.63	(0.01)	11.00	(0.05)	1.81	(0.05)	1.36	(0.01)	15.87	(0.09)
25041	PTXX13	8.63	(0.22)	69.11	(0.32)	3.71	(0.54)	65.39	(0.22)	1.77	(0.08)	0.50	(0.00)	1.55	(0.02)	0.64	(0.01)	10.62	(0.18)	1.86	(0.04)	1.45	(0.06)	15.99	(0.30)
25042	PTXX14	7.42	(0.06)	66.40	(0.19)	2.32	(0.03)	64.08	(0.23)	1.57	(0.00)	0.57	(0.01)	1.41	(0.03)	0.67	(0.02)	9.67	(0.37)	2.15	(0.04)	1.22	(0.03)	15.02	(0.46)
25043	PTHG7	13.94	(0.41)	72.50	(0.52)	6.29	(0.24)	66.20	(0.28)	1.62	(0.01)	0.53	(0.00)	0.82	(0.03)	0.40	(0.01)	3.83	(0.19)	1.39	(0.03)	0.65	(0.02)	7.23	(0.27)
25044	PTMM9	7.93	(0.17)	70.46	(0.02)	3.80	(0.71)	66.66	(0.72)	1.66	(0.02)	0.37	(0.01)	1.36	(0.02)	0.69	(0.02)	9.63	(0.25)	1.94	(0.02)	1.26	(0.02)	14.56	(0.27)
25045	PTLW11	5.16	(0.38)	57.56	(0.06)	2.26	(0.12)	55.29	(0.18)	2.86	(0.05)	0.29		2.19		1.06		14.79		2.80		2.11		22.16	
25046	PTXX15	23.06	(0.94)	61.65	(0.65)	12.65	(0.89)	49.00	(0.24)	2.91	(0.07)	1.15	(0.02)	0.52	(0.01)	0.22	(0.01)	11.93	(0.19)	4.76	(0.04)	0.52	(0.00)	18.88	(0.26)
25047	PTXX16	20.94	(0.63)	60.60	(0.82)	12.06	(0.63)	48.54	(0.20)	2.75	(0.02)	1.17	(0.02)	0.53	(0.00)	0.22	(0.00)	12.40	(0.60)	5.09	(0.26)	0.54	(0.00)	19.73	(0.85)
25048	PTXX17	18.42	(0.16)	60.17	(0.99)	11.32	(1.29)	48.85	(0.30)	2.82	(0.13)	1.25	(0.04)	0.57	(0.02)	0.23	(0.01)	12.95	(0.55)	5.31	(0.23)	0.55	(0.01)	20.64	(0.82)
25049	PTXX18	5.99	(0.04)	76.59	(0.05)	0.45	(0.05)	76.14	(0.01)	1.38	(0.07)	0.41	(0.00)	0.97	(0.01)	0.32	(0.01)	4.96	(0.04)	1.27	(0.00)	0.73	(0.01)	8.34	(0.04)
25050	PTLW12	2.38	(0.16)	56.83	(0.21)	0.80	(0.02)	56.03	(0.19)	2.79	(0.04)	0.57	(0.01)	2.46	(0.03)	1.13	(0.03)	14.74	(0.06)	2.88	(0.00)	2.03	(0.04)	22.68	(0.07)
25051	PTLW13	2.02	(0.02)	52.46	(0.45)	0.34	(0.00)	52.12	(0.45)	3.05	(0.00)	0.42	(0.00)	2.73	(0.02)	1.44	(0.01)	17.28	(0.14)	3.26	(0.02)	2.45	(0.02)	26.13	(0.15)
25052	PTMM10	4.58	(0.01)	63.40	(0.30)	0.81	(0.57)	63.00	(0.27)	2.08		0.39	(0.01)	1.98	(0.02)	1.00	(0.00)	12.61	(0.21)	2.37	(0.06)	1.80	(0.04)	19.15	(0.32)
25053	PTHG8	9.91	(0.12)	75.01	(0.29)	1.17	(0.02)	73.84	(0.31)	1.62	(0.03)	0.56	(0.01)	0.90	(0.01)	0.39	(0.00)	4.46	(0.07)	1.64	(0.03)	0.74	(0.06)	8.30	(0.18)

Table D-4: Whole Mass data (% DM), including standard deviation of the duplicates (SD) for some statistics, for peat samples 25001-25017.

NIR #	UNIV. CODE	EXTR_P	EXTR_P	EXTR_C	EXTR_C	ASH_	ASH	UA	KL	ASL	AIR	AIA	ARA_	GAL_	RHA_	GLU_	XYL	MAN	TOT_	MC	MC
		D_AV	D_SD	V_AV	V_SD	AV	SD						SRS	SRS	SRS	SRS	SRS	SRS	SRS	AV	SD
25001	PTHG4	5.72	(0.05)	6.27	(0.33)	6.02	(0.35)	0.74	60.67	1.68	63.52	2.85	0.44	1.61	0.65	9.60	2.33	1.31	15.29	60.85	(1.88)
25002	PTHG5	6.02	(0.48)	5.33	(0.36)	7.76	(0.00)	1.19	50.29	2.38	54.78	4.49	0.36	2.00	1.02	13.21	2.43	1.81	19.81	66.44	(3.13)
25003	PTHG3	7.50	(0.16)	7.80	(0.42)	2.13	(0.09)	1.08	58.65	1.93	58.73	0.09	0.18	1.67	0.84	13.13	2.42	1.76	19.16	71.60	(0.03)
25004	PTLW3	5.53				1.89	(0.06)	0.25	46.12	3.18	47.08	0.96	0.37	2.58	1.37	15.75	2.98	2.17	23.84	45.19	(1.94)
25005	PTMM3	6.18		6.97	(0.30)	1.99	(0.17)	0.43	54.92	2.25	55.22	0.30	0.29	2.24	1.05	14.86	2.64	1.92	21.96	59.93	(0.25)
25006	PTMM4	6.62	(0.21)	7.13	(0.50)	2.60	(0.04)		65.64	1.13	65.79	0.15	0.25	1.54	0.48	10.93	1.73	1.28	15.74	51.91	(0.01)
25007	PTMH2	6.60		7.00		4.24	(0.03)	0.77	63.72	1.62	64.11	0.39	0.30	1.39	0.54	9.52	1.83	1.20	14.25	64.85	(0.78)
25008	PTLW4	7.52	(0.52)	8.23	(0.60)	1.56	(0.04)	1.47	55.64	2.10	55.98	0.35	0.25	2.03	0.93	13.21	2.32	1.93	19.74	68.45	(1.43)
25009	PTMM5	5.06	(0.13)			2.99	(0.01)	0.59	56.84	2.28	57.63	0.79	0.31	1.93	0.89	13.39	2.32	1.78	19.73	56.39	(0.98)
25010	PTLW1	6.19		6.66		2.53		0.18	45.98	3.01	47.03	1.05	0.25	2.47	1.51	17.55	2.94	2.51	25.72	70.93	(0.93)
25011	PTMM1	7.38		8.09		1.71	(0.03)	0.18	54.54	2.12	55.79	1.25	0.24	1.82	0.90	14.83	2.27	1.78	20.94	60.49	(0.98)
25012	PTHG1	4.62	(0.06)	5.06	(0.00)	9.99	(0.29)	0.55	67.23	1.66	70.46	3.23	0.39	0.83	0.36	4.22	1.45	0.68	7.57	38.73	(0.94)
25013	PTHG2	4.84	(0.12)	5.39	(0.26)	10.23	(0.39)	1.17	66.26	1.51	70.19	3.93	0.37	0.80	0.34	4.33	1.51	0.68	7.69	46.10	(0.68)
25014	PTMM2	8.03		7.93		2.81	(0.23)	1.09	54.23	2.12	55.29	1.06	0.25	1.76	0.79	14.11	2.10	1.68	19.90	57.24	(1.63)
25015	PTLW2	6.51		6.21		2.47	(0.04)	0.99	45.46	3.22	46.21	0.75	0.26	2.41	1.38	16.69	2.81	2.33	24.51	68.52	(0.25)
25016	PTLW5	6.00		6.07		5.71	(0.25)	0.74	43.53	2.67	47.98	4.45	0.37	2.55	1.30	16.87	2.78	2.10	24.69	87.31	(1.59)
25017	PTLW6	6.79				3.89	(0.10)	1.19	47.61	2.78	48.88	1.27	0.33	2.34	1.36	15.45	2.78	2.01	22.92	65.50	(1.69)

Table D-5: Whole Mass data (% DM), including standard deviation of the duplicates (SD) for some statistics, for peat samples 25018-25053.

NIR #	UNIV. CODE	EXTR_P D_AV	EXTR_P D_SD	EXTR_C V_AV	EXTR_C V_SD	ASH_ AV	ASH SD	UA	KL	ASL	AIR	AIA	ARA_ SRS	GAL_ SRS	RHA_ SRS	GLU_ SRS	XYL _SRS	MAN _SRS	TOT_ SRS	MC AV	MC SD
25018	PTLW7	5.96	(0.08)	6.70	(0.13)	2.34	(0.06)	0.73	49.69	2.96	51.37	1.68	0.57	2.56	1.07	14.33	2.78	2.09	22.34	38.21	(1.56)
25019	PTLW8	6.32		7.52		1.72		0.94	54.86	1.89	55.25	0.40	0.30	2.23	0.92	16.14	2.88	1.88	23.42	60.97	(0.52)
25020	PTLW9	7.75	(0.07)	8.26	(0.32)	2.14	(0.22)	0.33	62.77	1.60	63.46	0.69	0.26	1.52	0.65	11.04	2.12	1.16	16.09	38.16	(1.36)
25021	PTLM1	4.88		4.88		14.25		0.48	62.65	1.49	70.15	7.50	0.93	1.09	0.48	5.10	1.56	0.63	9.31	62.20	(2.16)
25022	PTLM2	5.87	(0.27)	6.27	(0.02)	7.56	(0.39)	0.29	62.68	1.38	65.97	3.29	0.36	1.45	0.61	10.05	1.74	1.19	14.79	62.71	(2.01)
25023	PTLM3	7.74	(0.22)	8.05	(0.18)	3.05	(0.12)	0.64	58.31	1.64	58.60	0.29	0.21	1.98	0.78	13.00	2.09	1.80	19.09	72.55	(0.68)
25024	PTMM6	6.84						0.46	62.79	1.59	63.59	0.80	0.34	1.37	0.55	10.09	2.30	1.23	15.34	57.94	(1.22)
25025	PTMM7	6.77						0.12	64.99	1.06	65.41	0.42	0.26	1.50	0.45	11.23	1.47	1.31	15.77	79.78	(0.27)
25026	PTXX1	4.20	(0.05)	4.29	(0.08)	10.39	(0.04)	0.63	60.08	2.11	61.80	1.73	1.33	1.69	0.90	6.17	2.32	1.47	12.99	55.49	(0.10)
25027	PTXX2	7.86	(0.01)	7.09	(0.14)	1.96	(0.28)	0.93	52.77	2.27	53.28	0.52	0.25	1.91	0.97	14.27	2.54	1.81	20.78	70.74	(0.44)
25028	PTXX3	7.62		8.84	(0.31)	1.43	(0.08)	0.80	54.73	2.26	55.16	0.43	0.24	2.08	0.90	14.21	2.47	1.92	20.92	62.12	(1.05)
25029	PTXX4	8.19	(0.04)	7.66	(0.12)	1.79	(0.11)		59.31	1.54	59.71	0.40	0.23	1.66	0.72	13.51	2.28	1.60	19.28	49.16	(2.10)
25030	PTHG6	6.43	(0.18)	6.31	(0.24)	4.76	(0.05)		65.13	1.57	66.20	1.07	0.32	1.34	0.55	9.84	1.82	1.16	14.48	53.67	(0.42)
25031	PTLW10	5.47	(0.09)	5.78	(0.11)	4.81	(0.23)		63.70	1.70	64.28	0.59	0.67	1.43	0.67	8.63	2.47	1.22	14.42	49.48	(0.06)
25032	PTMM8	5.91	(0.26)	5.96	(0.11)	5.99	(0.36)		64.35	1.61	65.78	1.43	0.45	1.40	0.65	8.05	1.74	1.27	12.91	50.83	(0.06)
25033	PTXX5	6.39	(0.03)	6.69	(0.13)	4.00	(0.13)	0.61	60.61	1.85	61.35	0.73	0.32	1.73	0.76	11.24	2.02	1.51	16.81	57.53	(0.41)
25034	PTXX6	6.73	(0.22)	5.54	(0.14)	1.40	(0.05)	1.55	42.54	3.25	43.00	0.47	0.24	2.70	1.68	19.51	3.14	2.66	28.25	79.07	(0.49)
25035	PTXX7	6.06	(0.21)	6.60	(0.27)	2.70	(0.05)		47.05	2.70	48.61	1.56	0.28	2.46	1.33	18.01	3.09	2.35	26.19	63.87	(0.34)
25036	PTXX8	7.01	(0.19)	6.96	(0.12)	1.38	(0.00)		53.74	2.14	54.04	0.30	0.19	2.09	1.09	16.10	2.44	2.14	22.95	61.00	(2.23)
25037	PTXX9	7.57	(0.12)	6.58	(0.14)	4.62	(0.24)	0.30	54.78	2.12	55.89	1.11	0.42	1.89	0.98	12.50	2.41	1.78	19.01	58.60	(0.55)
25038	PTXX10	5.61	(0.05)	6.15	(0.07)	4.90	(0.00)	0.46	64.30	2.00	65.60	1.30	0.26	1.18	0.59	8.64	1.64	1.13	12.85	58.27	(0.18)
25039	PTXX11	4.62	(0.07)	4.67	(0.34)	7.08	(0.19)		69.74	1.46	71.16	1.42	0.58	0.93	0.37	4.51	1.71	0.67	8.40	46.49	(0.10)
25040	PTXX12	7.08	(0.19)	6.70	(1.19)	5.37	(0.24)		62.30	1.44	64.52	2.22	0.22	1.36	0.58	10.22	1.68	1.27	14.75	58.21	(3.81)
25041	PTXX13	6.12	(0.11)	6.52	(0.02)	8.01	(0.07)		61.39	1.66	64.88	3.48	0.47	1.46	0.60	9.97	1.75	1.36	15.01	51.69	(0.03)
25042	PTXX14	4.97	(0.11)	5.05	(0.06)	7.68	(0.11)		60.89	1.49	63.10	2.21	0.55	1.34	0.63	9.19	2.04	1.15	14.27	59.10	(0.43)
25043	PTHG7	4.53	(0.16)	4.00	(0.17)	13.85	(0.16)	0.13	63.21	1.54	69.22	6.01	0.51	0.79	0.38	3.66	1.33	0.62	6.90	41.50	(0.11)
25044	PTMM9	5.77	(0.35)	5.84	(0.09)	7.45	(0.32)		62.81	1.57	66.39	3.58	0.35	1.28	0.65	9.07	1.83	1.19	13.72	53.73	(0.35)
25045	PTLW11	6.27	(0.08)	6.01	(0.41)	5.30	(0.17)	1.19	51.83	2.68	53.95	2.12	0.27	2.05	0.99	13.86	2.62	1.98	20.77	52.15	(1.94)
25046	PTXX15	2.31	(0.63)	1.90	(0.03)	18.99	(0.20)		47.87	2.84	60.22	12.36	1.12	0.51	0.21	11.65	4.65	0.51	18.44	68.05	(0.81)
25047	PTXX16	3.73	(0.50)	2.83	(0.33)	17.98	(0.19)		46.73	2.65	58.34	11.61	1.13	0.51	0.21	11.93	4.90	0.52	18.99	69.78	(0.24)
25048	PTXX17	3.22	(0.03)	2.81	(0.37)	18.89	(0.18)		47.27	2.73	58.23	10.96	1.21	0.55	0.22	12.53	5.14	0.54	19.97	69.66	(0.02)
25049	PTXX18	5.16	(0.82)	4.88	(0.26)	5.73	(0.10)		72.21	1.30	72.64	0.43	0.39	0.92	0.30	4.70	1.20	0.69	7.91	78.15	(0.75)
25050	PTLW12	7.88	(0.17)	6.42	(0.16)	1.30	(1.64)		51.62	2.57	52.36	0.74	0.53	2.26	1.04	13.58	2.65	1.87	20.90	54.16	(0.54)
25051	PTLW13	6.95	(0.33)	6.62	(0.30)	1.87	(0.06)		48.50	2.84	48.82	0.31	0.39	2.54	1.34	16.08	3.03	2.28	24.31	64.16	(0.61)
25052	PTMM10	6.02	(0.19)	6.19	(0.48)	4.67	(0.00)	0.95	59.20	1.96	59.58	0.76	0.37	1.86	0.94	11.85	2.23	1.69	18.00	59.55	(0.00)
25053	PTHG8	4.28	(0.01)			9.74	(0.11)	0.94	70.68	1.55	71.81	1.12	0.54	0.86	0.38	4.27	1.57	0.71	7.95	50.07	(0.18)

Table D-6: Histograms (using % whole dry mass data) for selected chemical properties of the 53 peat samples, with associated statistics.

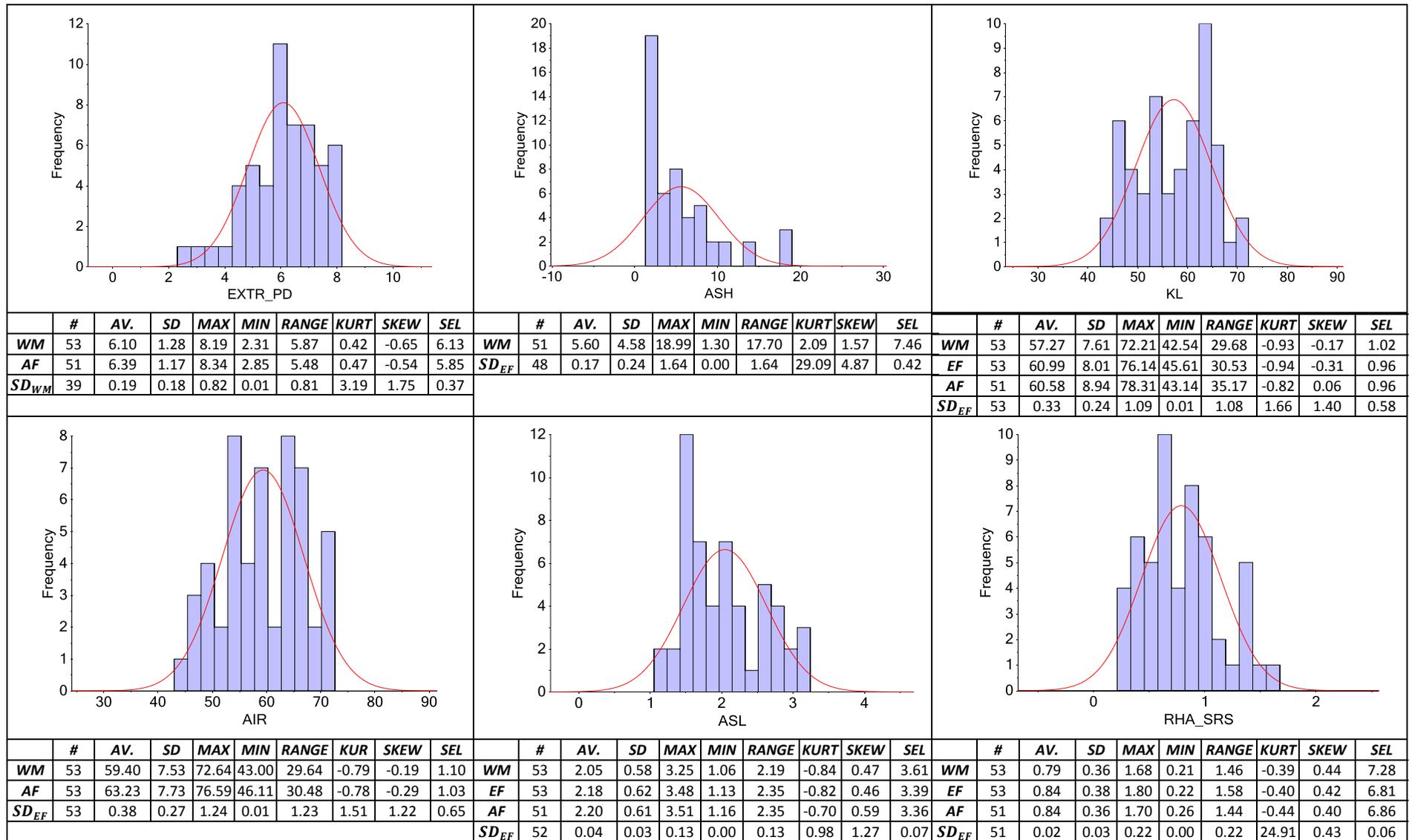


Table D-7: Histograms (using % whole dry mass data) for selected chemical properties of the 53 peat samples, with associated statistics.

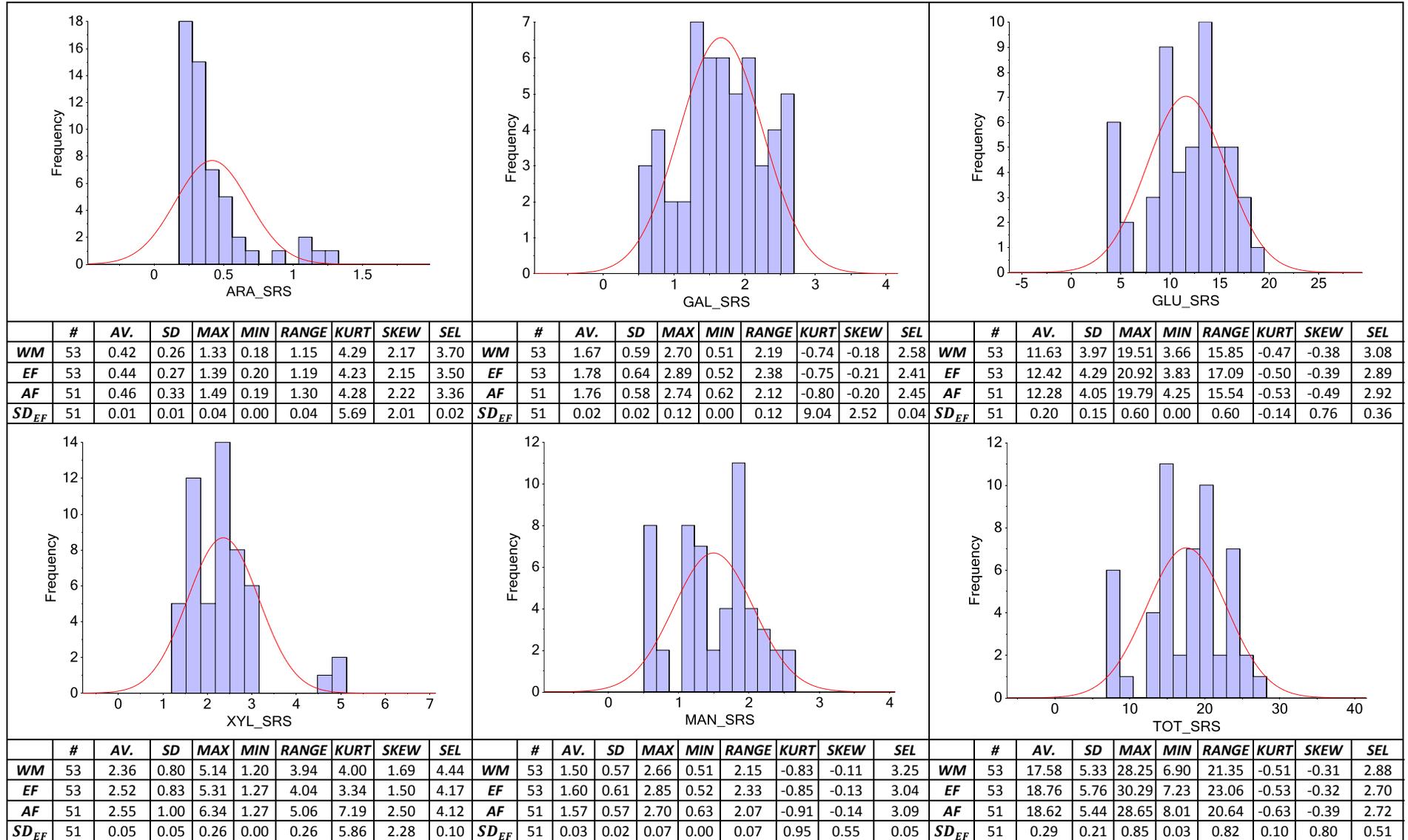


Table D-8: Correlation table for selected constituents of the 53 peat samples. Absolute values greater than 0.5 are highlighted in bold.

	EXTR_CV	ASH	UA	ASA	KL	ASL	AIR	AIA	EIA	ARA_SRS	GAL_SRS	RHA_SRS	GLU_SRS	XYL_SRS	MAN_SRS	TOT_SRS
EXTR_PD	0.851	-0.837	0.092	-0.833	-0.101	-0.037	-0.368	-0.714	-0.833	-0.749	0.600	0.476	0.485	-0.269	0.616	0.416
EXTR_CV		-0.825	0.030	-0.772	0.095	-0.185	-0.198	-0.744	-0.811	-0.742	0.527	0.350	0.355	-0.385	0.531	0.283
ASH			-0.477	0.900	0.118	-0.107	0.461	0.913	0.985	0.806	-0.777	-0.667	-0.540	0.285	-0.770	-0.488
UA				-0.464	-0.784	0.825	-0.803	-0.300	-0.438	-0.149	0.773	0.856	0.715	0.824	0.799	0.758
ASA					0.383	-0.267	0.647	0.698	0.907	0.693	-0.858	-0.736	-0.745	0.035	-0.853	-0.701
KL						-0.923	0.931	-0.211	0.067	-0.070	-0.619	-0.703	-0.848	-0.781	-0.656	-0.891
ASL							-0.872	0.159	-0.039	0.125	0.572	0.694	0.718	0.735	0.607	0.779
AIR								0.161	0.417	0.204	-0.839	-0.884	-0.936	-0.590	-0.872	-0.961
AIA									0.934	0.736	-0.573	-0.463	-0.214	0.533	-0.560	-0.166
EIA										0.777	-0.763	-0.638	-0.496	0.330	-0.753	-0.446
ARA_SRS											-0.489	-0.402	-0.348	0.473	-0.520	-0.248
GAL_SRS												0.950	0.808	0.126	0.975	0.811
RHA_SRS													0.799	0.214	0.957	0.814
GLU_SRS														0.557	0.857	0.992
XYL_SRS															0.178	0.621
MAN_SRS																0.853

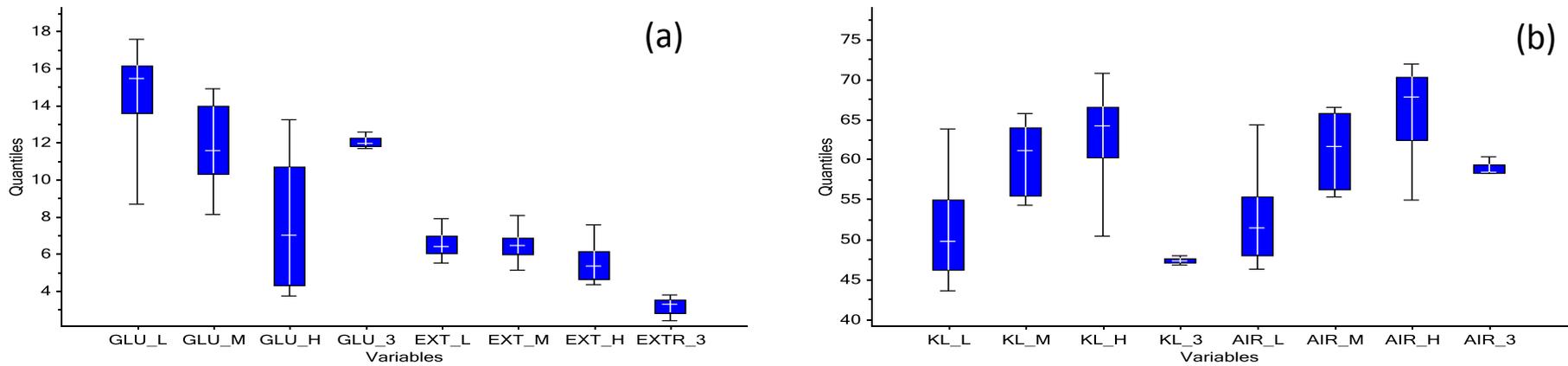


Figure D-1: Quantile plots for the Low (L), Medium (M), High (H), and "3" (3) peat classes for (a) GLU_SRS (GLU) and EXTR_PD (EXT), (b) KL and AIR.

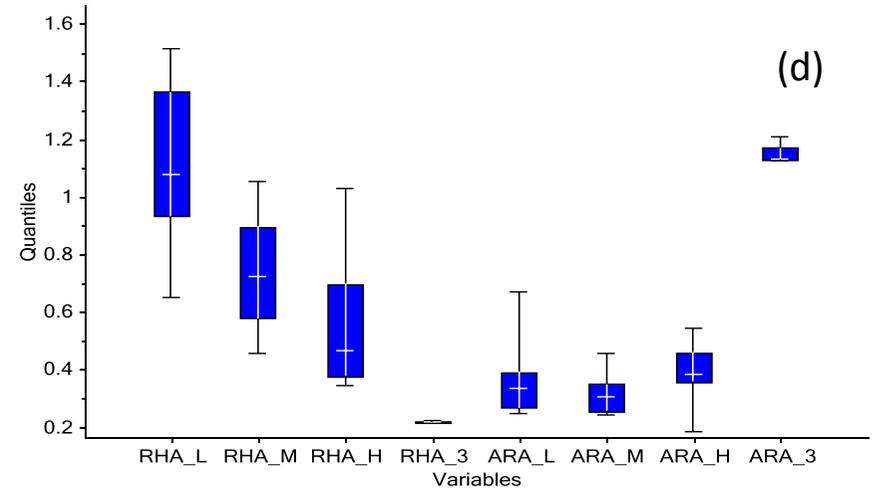
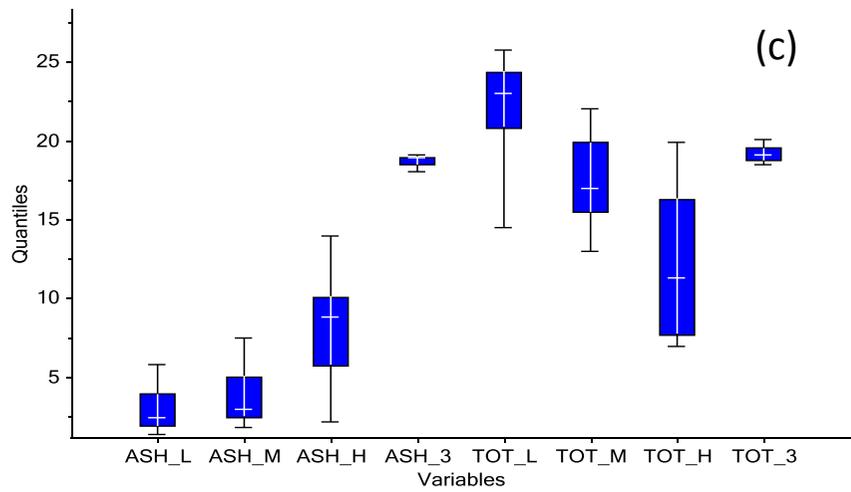
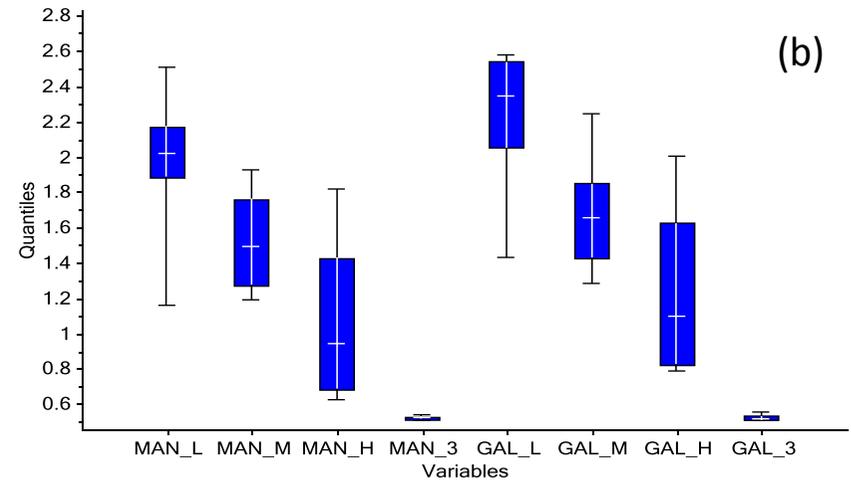
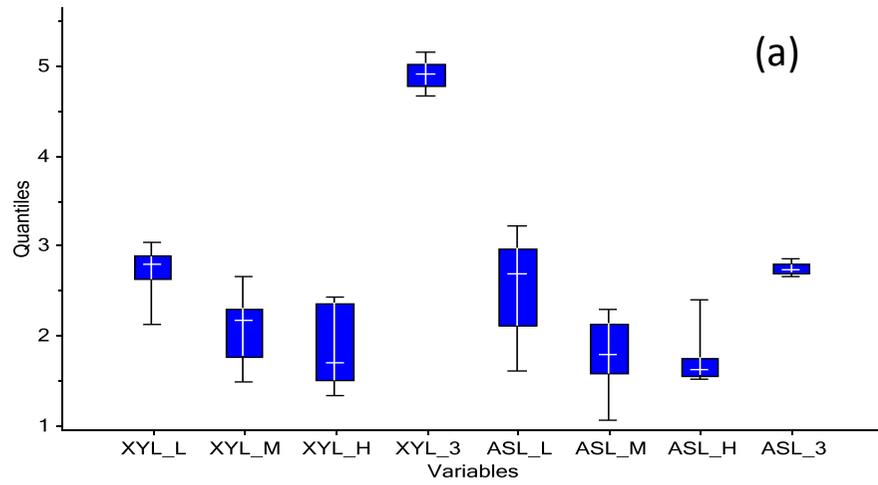


Figure D-2: Quantile plots for the Low (L), Medium (M), High (H), and "3" (3) peat classes for (a) XYL_SRS (XYL) and ASL, (b) MAN_SRS (MAN) and GAL_SRS (GAL), (c) Ash and TOT_SRS (TOT), (d) RHA_SRS (RHA) and ARA_SRS (ARA).

Table D-9: Histograms, with statistics, for the moisture content (% wet basis), and uronic acids (UA) content (% whole dry whole mass).

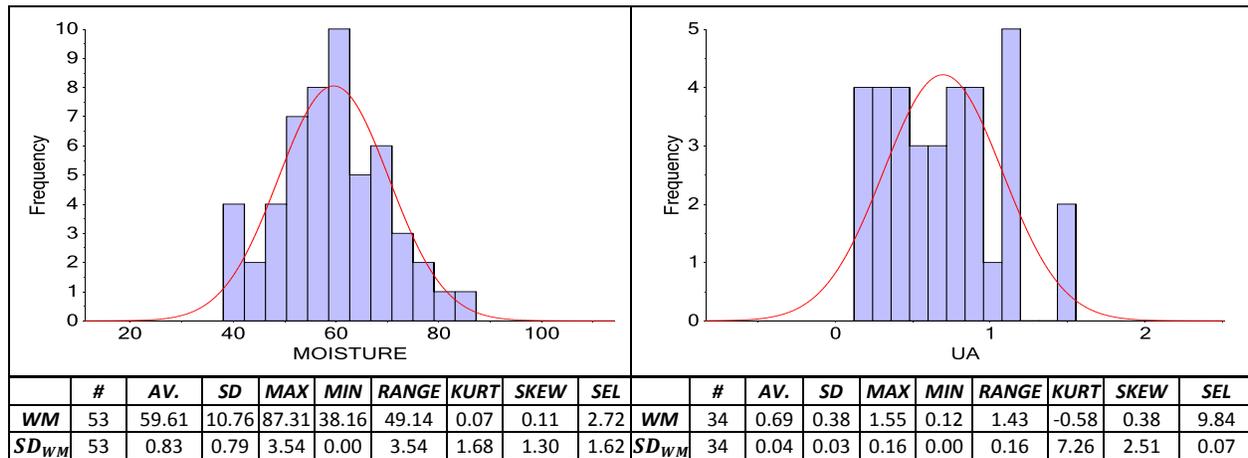


Table D-10: Summary data for the sugar recoveries of the 13 hydrolysis batches.

	Average Sugar Recoveries over all Batches (%)						Standard Deviation within a Batch (%)					
	Ara	Gal	Rha	Glu	Xyl	Man	Ara	Gal	Rha	Glu	Xyl	Man
Av	92.60	94.04	94.98	95.35	86.72	93.64	0.45	0.29	0.61	0.18	0.24	1.65
Max	93.66	94.67	96.83	96.02	87.66	94.99	1.81	0.68	1.30	0.57	0.61	7.28
Min	91.06	93.38	93.66	94.51	85.65	91.87	0.07	0.04	0.11	0.01	0.05	0.28
Range	2.60	1.29	3.18	1.50	2.01	3.12	1.74	0.64	1.19	0.55	0.56	7.00
SD	0.81	0.39	1.14	0.46	0.70	1.02	0.48	0.21	0.45	0.17	0.18	1.95

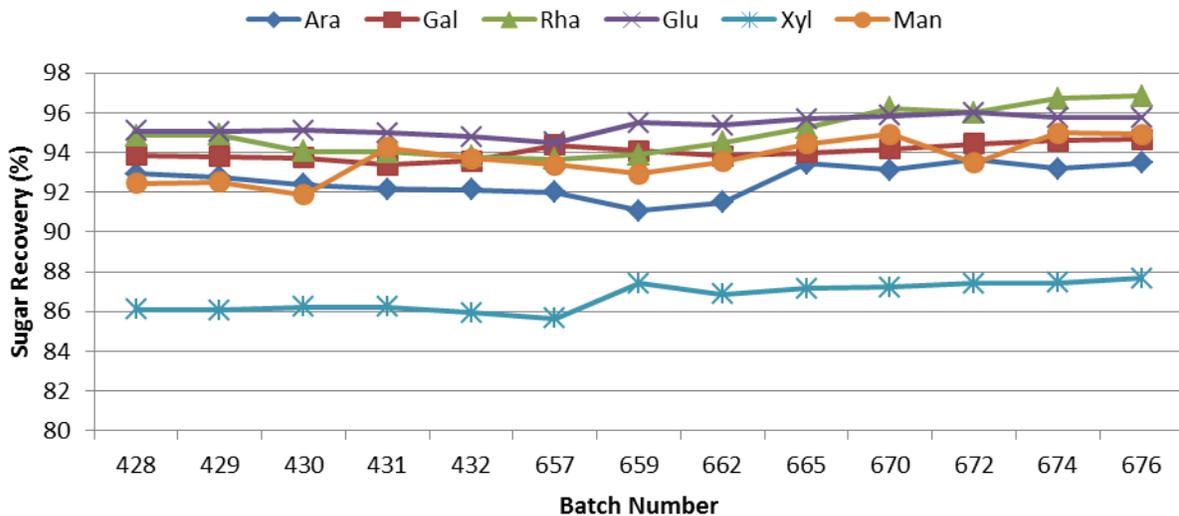


Figure D-3: Chart plotting the sugar recovery rates for each sugar over the 13 batches. Batch numbers 659 onwards represent a change from 13 to 11 tubes per hydrolysis.

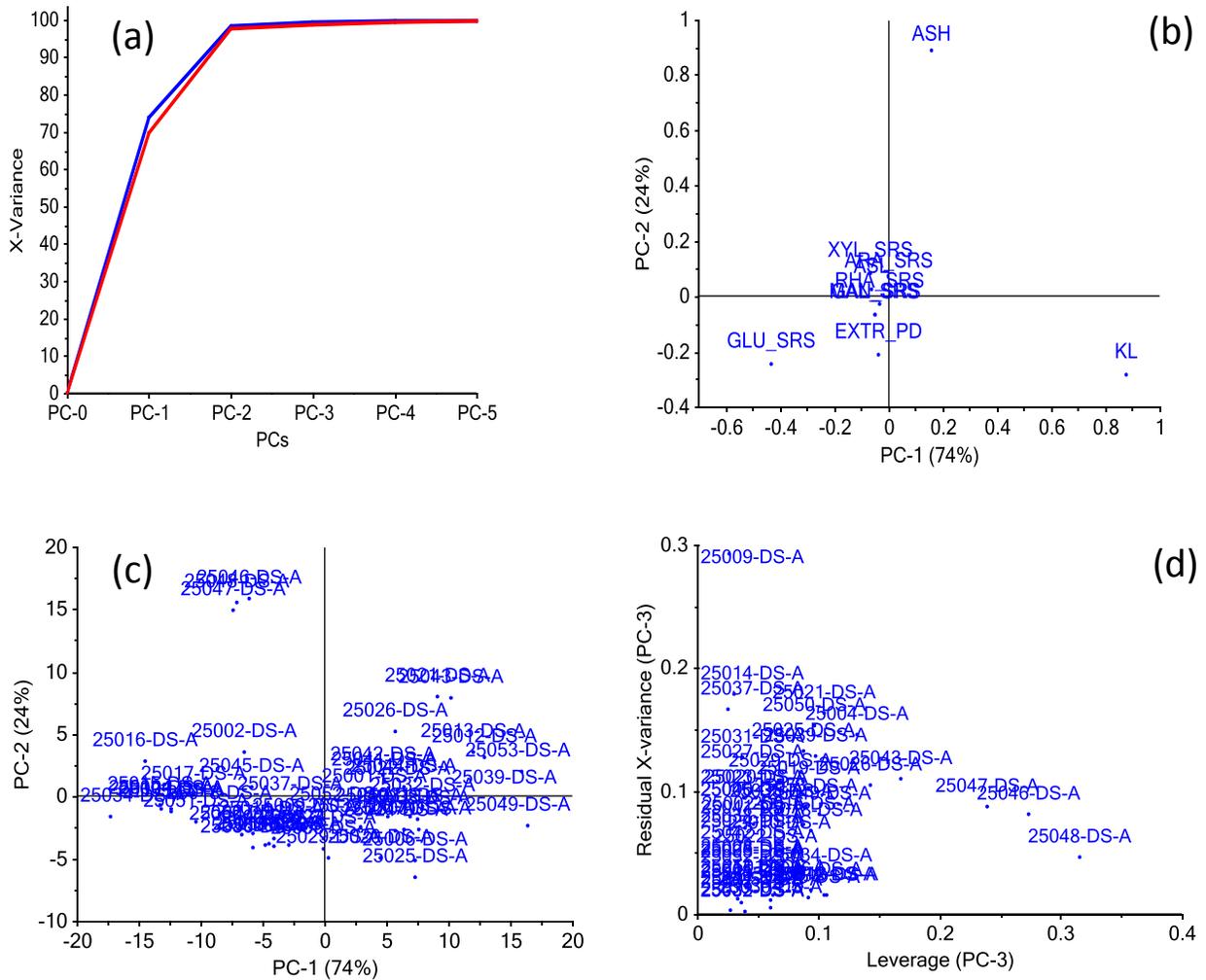


Figure D-4: Data correspond to the PCA involving 10 variables (chemical data) and 53 peat samples; (a) An explained variance plot with up to 5 PCs; (b) a PC1 vs. PC2 loadings plot; (c) a PC1 vs. PC2 scores plot; (d) an influence plot using a 3 PC model.

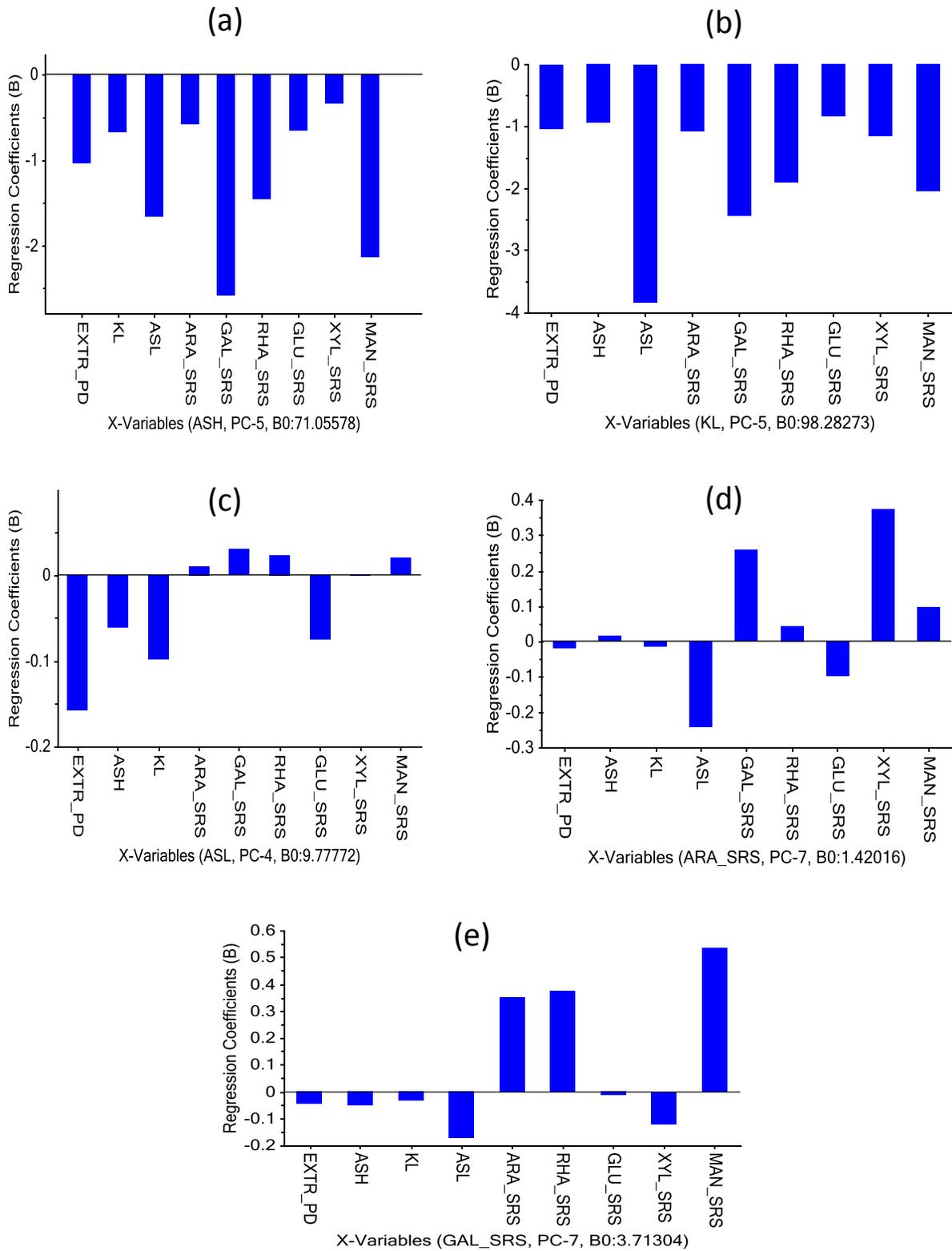


Figure D-5: Regression coefficients plots for PCR models using chemical data. (a) ash content model, using 5 PCs; (b) KL content model, using 5 PCs; (c) ASL content model, using 4 PCs; (d) ARA_SRS content model, using 7 PCs; (e) GAL_SRS content model, using 7 PCs.

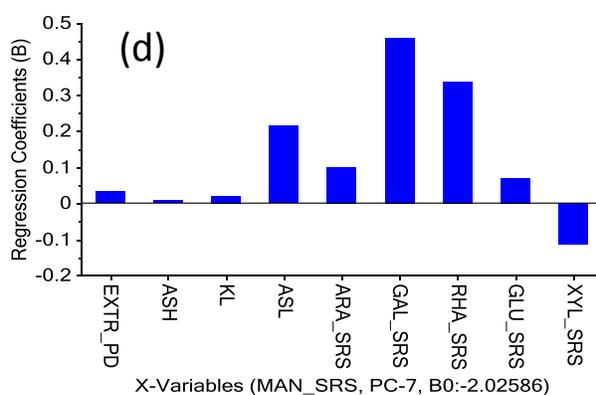
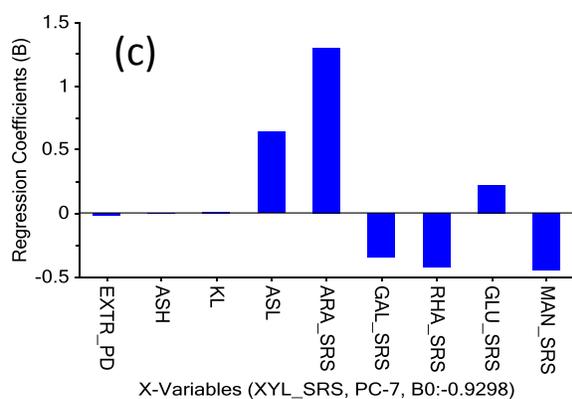
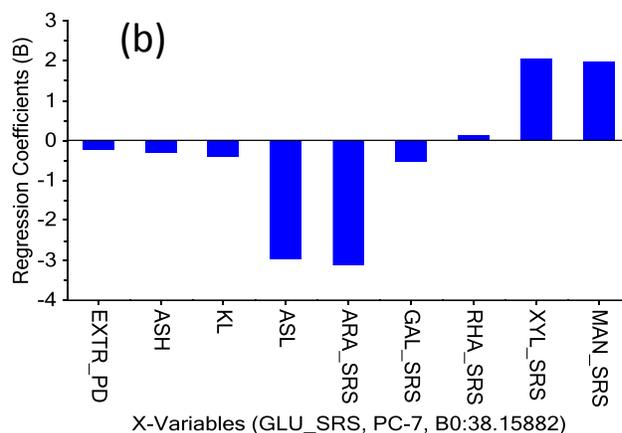
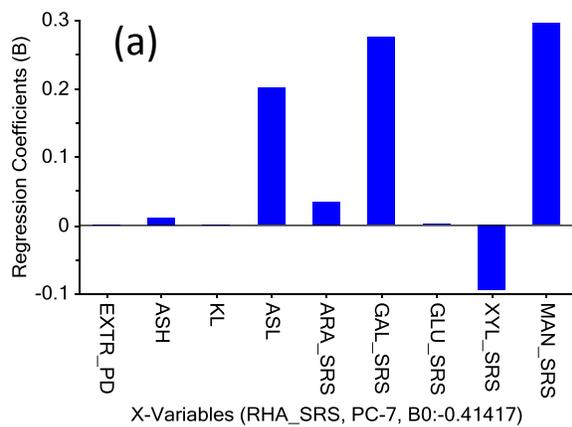


Figure D-6: Further regression coefficients plots for PCR models using chemical data. (a) RHA_SRS content model, using 7 PCs; (b) GLU_SRS content model, using 7 PCs; (c) XYL_SRS content model, using 7 PCs; (d) MAN_SRS content model, using 7 PCs.

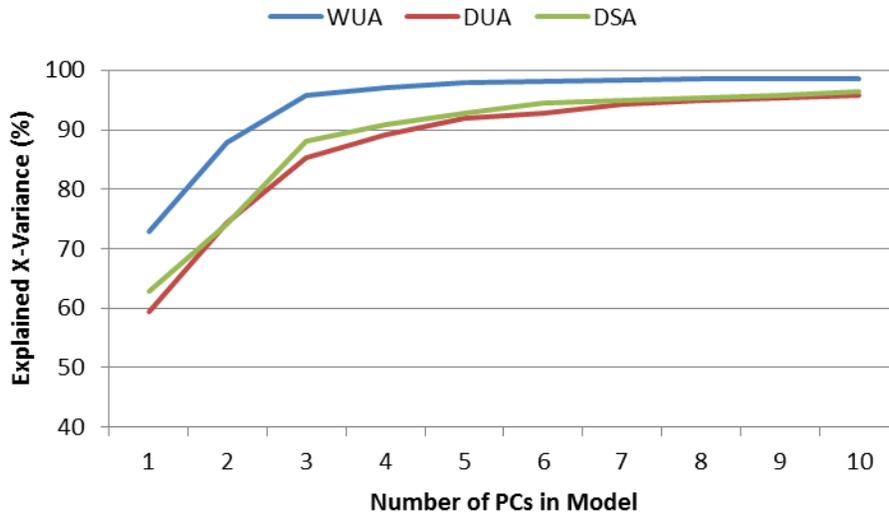


Figure D-7: An explained X-variance plot, under full cross validation and with increasing numbers of PCs, for PCA models based on the 400-2500 nm spectra for the WU, DU, and DS datasets.

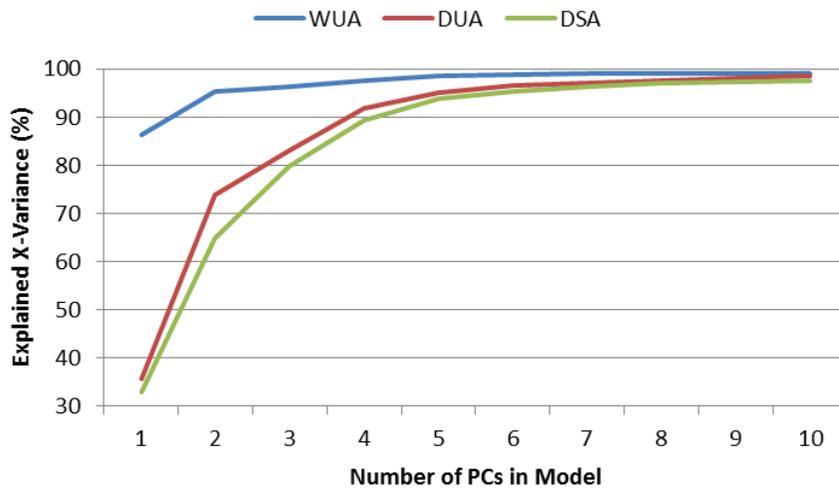


Figure D-8: An explained X-variance plot, under full cross validation and with increasing numbers of PCs, for PCA models based on the 1100-2500 nm spectra for the WU, DU, and DS datasets.

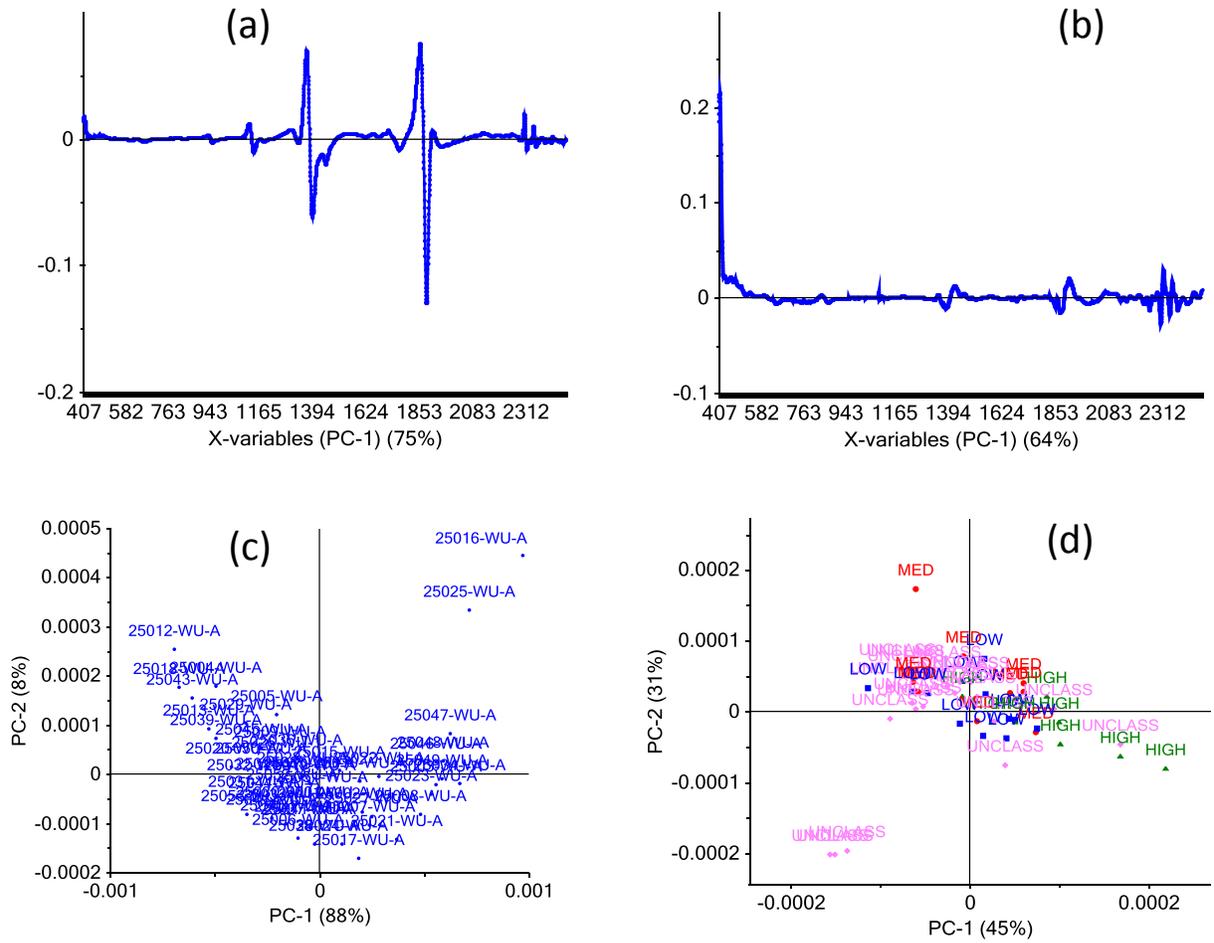


Figure D-9: Plots for various PCA models based on the peat spectra. (a) A PC1 loadings plot for the WU 400-2500 nm model, (b) a PC1 loadings plot for the DS 400-2500 nm model, (c) a PC1 vs. PC2 scores plot for the WU 1100-2500 nm model. (d) a PC1 vs. PC2 scores plot for the DU 1100-2500 nm model, with samples labelled according to whether they were classified with high (HIGH), medium (MED), or low (LOW) degrees of humification or whether they were unclassified (UNCLASS).

Table D-11: Regression statistics for PLSR models for the glucose, total sugars, and ash contents of peat samples, all on a whole mass basis (%).

Dataset	GLUCOSE						TOTAL SUGARS						ASH		
	DS	DS	DU	DU	WU	WU	DS	DS	DU	DU	WU	WU	DS	DU	WU
Pre. (1)	SG	SNV	SNV	SNV											
Specific (1)	1,1,10,10	1,1,10,10	2,2,25,25	2,2,25,25	2,2,25,25	2,2,25,25	1,1,10,10	1,1,10,10	2,2,25,25	2,2,25,25	1,1,10,10	1,1,10,10	1.1-2.5	1.1-2.5	1.1-2.5
PLS- λ 10 ³ nm	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5
Sam. Excl.		25038		25038		25038		25038		25038		25038			
Calib:Valid	40:13	39:13	38:13	37:13	40:13	39:13	40:13	39:13	38:13	37:13	40:13	39:13	38:13	36:13	38:13
F-Haaland's	6	11	7	8	9	9	4	9	7	7	8	10	5	6	8
R^2_{calib}	0.960	0.992	0.971	0.976	0.969	0.983	0.943	0.990	0.968	0.973	0.961	0.980	0.939	0.953	0.930
$Offset_{calib}$	0.476	0.095	0.347	0.280	0.367	0.203	1.010	0.186	0.562	0.483	0.692	0.353	0.320	0.254	0.373
RMSEC (%)	0.813	0.363	0.704	0.633	0.717	0.535	1.307	0.561	0.994	0.920	1.078	0.767	1.078	0.967	1.151
R^2_{CV}	0.926	0.939	0.923	0.942	0.879	0.915	0.916	0.942	0.928	0.932	0.851	0.913	0.898	0.917	0.799
RMSECV (%)	1.102	1.007	1.148	1.001	1.419	1.200	1.593	1.328	1.504	1.461	2.117	1.626	1.402	1.283	2.009
$BIAS_{CV}$	-0.019	-0.053	-0.042	-0.039	0.026	-0.043	-0.123	-0.058	0.075	-0.021	-0.015	-0.179	0.004	-0.017	-0.105
SECV (%)	1.115	1.019	1.162	1.014	1.437	1.215	1.609	1.344	1.522	1.481	2.144	1.637	1.420	1.301	2.033
R^2_{pred}	0.973	0.952	0.950	0.949	0.922	0.903	0.954	0.948	0.943	0.935	0.940	0.930	0.906	0.873	0.888
$Slope_{pred}$	0.910	0.885	1.090	1.109	0.991	1.007	0.894	0.889	1.085	1.088	1.020	1.018	0.876	0.837	0.942
$Offset_{pred}$	0.793	1.208	-1.307	-1.519	-0.239	-0.338	1.705	1.594	-2.020	-2.048	-0.554	-0.454	0.262	0.327	-0.731
RMSEP (%)	0.662	0.823	0.988	1.034	1.035	1.151	1.042	1.132	1.412	1.484	1.178	1.264	1.584	1.881	2.003
$Bias_{pred}$	-0.215	-0.084	-0.301	-0.294	-0.334	-0.256	-0.119	-0.314	-0.560	-0.536	-0.215	-0.160	-0.540	-0.734	-1.099
SEP (%)	0.652	0.852	0.979	1.032	1.019	1.168	1.077	1.132	1.349	1.440	1.206	1.305	1.550	1.803	1.743
RPD_{pred}	5.677	4.342	3.777	3.586	3.470	3.029	4.463	4.246	3.564	3.339	3.870	3.575	3.247	2.793	2.948
RER_{pred}	19.106	14.616	12.713	12.070	11.603	10.129	15.237	14.496	12.168	11.397	12.444	11.495	11.177	9.611	9.854
ALL SAMPLES IN CALIBRATION SET															
F-Haaland's	5	6	6	6	8	8	5	5	6	6	9	8	6	11	9
R^2_{calib}	0.956	0.980	0.961	0.964	0.960	0.970	0.956	0.978	0.961	0.964	0.963	0.969	0.944	0.981	0.933
RMSEC (%)	0.827	0.563	0.785	0.760	0.786	0.686	1.109	0.784	1.059	1.020	1.014	0.926	1.072	0.627	1.173
R^2_{CV}	0.941	0.967	0.935	0.938	0.891	0.914	0.945	0.972	0.937	0.938	0.901	0.909	0.912	0.937	0.862
RMSECV (%)	0.974	0.726	1.035	1.024	1.325	1.214	1.266	0.931	1.373	1.369	1.694	1.613	1.290	1.175	1.716
RPD_{CV}	4.040	5.440	3.844	3.906	2.968	3.253	4.171	5.681	3.904	3.923	3.119	3.282	3.519	3.908	2.641
RER_{CV}	16.134	21.635	15.169	15.349	11.851	12.938	16.705	22.706	15.398	15.441	12.491	13.116	13.609	14.919	10.212

Table D-12: Regression statistics for PLSR models for the xylose, rhamnose, and extractives contents of peat samples, all on a whole mass basis (%).

Dataset	XYLOSE						RHAMNOSE						EXTR_PD		
	DS	DS	DU	DU	WU	WU	DS	DS	DU	DU	WU	WU	DS	DU	WU
Pre. (1)	SG	SG	SG	SG											
Specific (1)	1,1,10,10	1,1,10,10	1,1,10,10	1,1,10,10	1,1,10,10	1,1,10,10	1,1,10,10	1,1,10,10	1,1,10,10	1,1,10,10	1,1,10,10	1,1,10,10	2,2,14,14	2,3,40,40	2,2,14,14
PLS- λ 10 ³ nm	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5
Sam. Excl.		25038		25038		25038		25038		25038		25038	25024	25024	25024
													25025	25025	25025
													25028	25028	25028
Calib:Valid	40:13	39:13	38:13	37:13	40:13	39:13	40:13	39:13	38:13	37:13	40:13	39:13	38:12	37:11	37:12
F-Haaland's	5	5	5	5	9	8	6	6	6	6	4	4	6	2	3
R^2_{calib}	0.942	0.952	0.956	0.956	0.969	0.971	0.954	0.966	0.939	0.939	0.900	0.903	0.905	0.749	0.621
$Offset_{calib}$	0.137	0.114	0.104	0.105	0.073	0.068	0.037	0.027	0.049	0.050	0.080	0.078	0.572	1.520	2.315
RMSEC (%)	0.190	0.172	0.169	0.170	0.140	0.135	0.077	0.066	0.091	0.091	0.111	0.110	0.402	0.638	0.757
R^2_{CV}	0.892	0.905	0.918	0.917	0.910	0.923	0.923	0.936	0.875	0.876	0.849	0.838	0.733	0.673	0.401
RMSECV (%)	0.259	0.243	0.231	0.233	0.239	0.223	0.099	0.091	0.131	0.132	0.137	0.143	0.675	0.729	0.995
$BIAS_{CV}$	0.006	0.007	0.009	0.010	0.008	0.005	-0.004	-0.003	-0.014	-0.012	-0.005	-0.011	0.001	0.012	0.042
SECV (%)	0.262	0.247	0.234	0.236	0.241	0.226	0.101	0.092	0.132	0.133	0.138	0.145	0.684	0.739	1.008
R^2_{pred}	0.948	0.950	0.944	0.944	0.975	0.966	0.887	0.912	0.906	0.905	0.740	0.747	0.769	0.852	0.650
$Slope_{pred}$	1.084	1.076	1.074	1.071	1.103	1.070	0.935	0.950	0.856	0.860	0.854	0.851	0.976	0.879	0.601
$Offset_{pred}$	-0.309	-0.272	-0.277	-0.265	-0.260	-0.151	0.052	0.051	0.084	0.080	0.176	0.180	-0.133	0.560	2.309
RMSEP (%)	0.243	0.229	0.242	0.240	0.161	0.168	0.115	0.102	0.109	0.109	0.199	0.197	0.690	0.501	0.850
$Bias_{pred}$	-0.112	-0.092	-0.103	-0.098	-0.014	0.017	0.002	0.012	-0.026	-0.027	0.063	0.066	-0.281	-0.181	-0.010
SEP (%)	0.224	0.218	0.228	0.227	0.166	0.174	0.120	0.106	0.110	0.110	0.197	0.193	0.658	0.490	0.888
RPD_{pred}	3.757	3.861	3.692	3.703	4.889	4.673	2.938	3.341	3.213	3.200	1.898	1.935	1.867	2.596	1.681
RER_{pred}	14.258	14.651	14.009	14.051	19.081	18.237	10.762	12.236	11.769	11.722	6.599	6.728	7.042	9.454	6.266
ALL SAMPLES IN CALIBRATION SET															
F-Haaland's	5	9	5	6	9	9	7	7	6	6	9	9	6	2	3
R^2_{calib}	0.941	0.978	0.952	0.963	0.968	0.974	0.951	0.962	0.939	0.939	0.950	0.953	0.875	0.767	0.632
RMSEC (%)	0.193	0.118	0.176	0.156	0.141	0.129	0.078	0.069	0.089	0.090	0.079	0.077	0.450	0.609	0.774
R^2_{CV}	0.913	0.936	0.929	0.939	0.934	0.941	0.923	0.938	0.895	0.895	0.838	0.859	0.802	0.726	0.479
RMSECV (%)	0.238	0.204	0.220	0.212	0.208	0.194	0.100	0.090	0.119	0.122	0.145	0.136	0.594	0.674	0.940
RPD_{CV}	3.332	3.887	3.677	3.818	3.807	4.089	3.532	3.935	3.025	2.968	2.448	2.623	2.152	1.871	1.358
RER_{CV}	16.398	19.097	17.750	18.392	18.734	20.091	14.466	16.014	12.169	11.859	10.026	10.673	9.809	8.386	6.191

Table D-13: Regression statistics for PLSR models for the galactose, arabinose, and acid insoluble ash (AIA) contents of peat samples, all on a whole mass basis (%).

Dataset	GALACTOSE						ARABINOSE						AIA		
	DS	DS	DU	DU	WU	WU	DS	DS	DU	DU	WU	WU	DS	DU	WU
Pre. (1)	SG	SG	SG	SG											
Specific (1)	2,2,25,25	2,2,25,25	2,2,25,25	2,2,25,25	2,2,25,25	2,2,25,25	1,2,30,30	1,2,30,30	2,2,25,25	2,2,25,25	2,2,25,25	2,2,25,25	2,2,25,25	2,2,25,25	2,2,25,25
PLS- λ 10 ³ nm	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5
Sam. Excl.		25038		25038		25038		25038		25038		25038			
Calib:Valid	40:13	39:13	38:13	37:13	40:13	40:12	40:13	39:13	38:13	37:13	40:13	39:13	40:13	38:13	40:13
F-Haaland's	9	10	10	5	8	8	8	8	7	7	9	9	2	10	6
R^2_{calib}	0.970	0.988	0.984	0.940	0.972	0.972	0.956	0.954	0.922	0.922	0.958	0.958	0.822	0.981	0.848
$Offset_{calib}$	0.052	0.020	0.027	0.105	0.047	0.047	0.018	0.019	0.031	0.031	0.017	0.017	0.352	0.038	0.303
RMSEC (%)	0.101	0.063	0.075	0.146	0.101	0.101	0.055	0.056	0.073	0.074	0.052	0.053	1.149	0.379	1.048
R^2_{CV}	0.903	0.925	0.890	0.878	0.927	0.927	0.779	0.771	0.728	0.713	0.747	0.719	0.767	0.815	0.645
RMSECV (%)	0.182	0.161	0.200	0.209	0.164	0.164	0.124	0.127	0.137	0.142	0.129	0.138	1.315	1.200	1.641
$BIAS_{CV}$	0.003	-0.003	-0.004	-0.005	-0.010	-0.010	-0.003	-0.009	-0.005	-0.003	0.001	0.003	0.011	-0.036	-0.088
SECV (%)	0.184	0.163	0.202	0.211	0.165	0.165	0.126	0.128	0.139	0.144	0.131	0.140	1.332	1.216	1.659
R^2_{pred}	0.962	0.975	0.930	0.899	0.855	0.877	0.697	0.693	0.919	0.913	0.888	0.889	0.760	0.775	0.771
$Slope_{pred}$	0.968	0.992	0.966	1.013	0.794	0.832	0.862	0.855	0.927	0.919	0.929	0.931	0.819	0.867	0.844
$Offset_{pred}$	-0.014	-0.008	-0.020	-0.088	0.368	0.282	0.022	0.031	0.008	0.013	0.095	0.094	0.172	0.239	0.240
RMSEP (%)	0.128	0.093	0.168	0.205	0.204	0.185	0.156	0.155	0.078	0.080	0.111	0.111	1.478	1.425	1.473
$Bias_{pred}$	-0.063	-0.019	-0.073	-0.067	0.020	-0.008	-0.044	-0.038	-0.027	-0.026	0.061	0.061	-0.304	-0.112	-0.162
SEP (%)	0.115	0.095	0.158	0.201	0.211	0.193	0.156	0.157	0.076	0.079	0.096	0.096	1.506	1.479	1.524
RPD_{pred}	5.133	6.249	3.756	2.948	2.584	2.824	1.708	1.703	3.507	3.383	2.968	2.974	2.020	2.057	2.059
RER_{pred}	17.160	20.890	12.556	9.856	8.429	9.187	5.707	5.689	11.718	11.303	10.142	10.163	7.084	7.214	7.409
ALL SAMPLES IN CALIBRATION SET															
F-Haaland's	11	11	11	11	8	9	12	12	8	10	9	9	3	10	12
R^2_{calib}	0.982	0.989	0.985	0.985	0.957	0.967	0.973	0.973	0.935	0.970	0.953	0.953	0.827	0.966	0.971
RMSEC (%)	0.078	0.061	0.073	0.072	0.121	0.106	0.043	0.043	0.067	0.045	0.056	0.057	1.160	0.517	0.470
R^2_{CV}	0.948	0.963	0.938	0.935	0.888	0.909	0.891	0.888	0.821	0.842	0.860	0.832	0.794	0.820	0.788
RMSECV (%)	0.136	0.114	0.151	0.153	0.199	0.173	0.088	0.090	0.113	0.110	0.099	0.108	1.288	1.220	1.307
RPD_{CV}	4.303	5.107	3.939	3.896	2.937	3.375	2.970	2.930	2.318	2.398	2.625	2.430	2.164	2.312	2.132
RER_{CV}	16.036	18.975	14.417	14.215	10.943	12.540	12.950	12.699	10.044	10.324	11.446	10.532	9.436	9.954	9.298

Table D-14: Regression statistics for PLSR models for the mannose, acid soluble lignin (ASL), and uronic acids (UA) contents of peat samples, all on a whole mass basis (%).

Dataset	MANNOSE						ACID SOLUBLE LIGNIN (ASL)						URONIC ACIDS		
	DS	DS	DU	DU	WU	WU	DS	DS	DU	DU	WU	WU	DS	DU	WU
Pre. (1)	SG	SG	SG	SG	SG	SG	SG	SG	SG						
Specific (1)	1,1,10,10	1,1,10,10	1,1,10,10	1,1,10,10	1,1,10,10	1,1,10,10	1,1,10,10	1,1,10,10	1,1,10,10	1,1,10,10	1,1,10,10	1,1,10,10	2,2,25,25	2,2,14,14	2,2,14,14
PLS- λ 10 ³ nm	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5
Sam. Excl.		25038		25038		25038		25038		25038		25038	25053	25053	25053
Calib:Valid	40:13	40:12	38:13	38:12	40:13	39:13	40:13	39:13	38:13	37:13	40:13	39:13	25:8	25:8	25:8
F-Haaland's	5	5	4	4	8	8	5	5	3	4	9	8	2	2	3
R^2_{calib}	0.974	0.974	0.958	0.958	0.964	0.972	0.923	0.911	0.889	0.911	0.950	0.942	0.836	0.861	0.832
$Offset_{calib}$	0.039	0.039	0.065	0.065	0.053	0.041	0.158	0.181	0.229	0.184	0.104	0.119	0.113	0.095	0.118
RMSEC (%)	0.093	0.093	0.122	0.122	0.107	0.095	0.162	0.176	0.196	0.178	0.128	0.139	0.160	0.147	0.162
R^2_{CV}	0.962	0.962	0.930	0.930	0.882	0.906	0.854	0.942	0.853	0.871	0.877	0.855	0.763	0.793	0.703
RMSECV (%)	0.114	0.114	0.157	0.157	0.195	0.175	0.223	0.226	0.226	0.215	0.203	0.221	0.192	0.180	0.217
$BIAS_{CV}$	-0.002	-0.002	0.008	0.008	-0.009	-0.006	0.006	0.007	-0.001	-0.001	-0.009	0.002	-0.003	0.009	0.020
SECV (%)	0.115	0.115	0.159	0.159	0.198	0.177	0.226	0.229	0.229	0.218	0.205	0.224	0.196	0.183	0.221
R^2_{pred}	0.778	0.873	0.911	0.918	0.940	0.925	0.873	0.834	0.889	0.913	0.935	0.943	0.896	0.917	0.847
$Slope_{pred}$	0.873	0.926	0.977	0.992	1.003	1.002	0.948	0.902	0.884	0.896	0.848	0.789	0.781	0.785	0.825
$Offset_{pred}$	0.248	0.128	0.089	0.055	-0.042	-0.039	0.110	0.209	0.236	0.214	0.420	0.564	0.084	0.125	0.114
RMSEP (%)	0.250	0.185	0.162	0.157	0.145	0.161	0.202	0.230	0.185	0.164	0.197	0.219	0.134	0.110	0.127
$Bias_{pred}$	0.069	0.022	0.056	0.043	-0.037	-0.036	0.001	0.005	-0.006	-0.002	0.113	0.138	-0.066	-0.023	0.002
SEP (%)	0.250	0.192	0.158	0.157	0.146	0.164	0.211	0.239	0.192	0.171	0.167	0.177	0.125	0.115	0.135
RPD_{pred}	2.071	2.772	3.275	3.382	3.933	3.497	2.741	2.413	3.000	3.384	3.689	3.493	2.901	3.134	2.554
RER_{pred}	7.204	9.369	11.394	11.430	13.673	12.160	8.562	7.536	9.368	10.568	12.271	11.619	8.080	8.727	7.836
ALL SAMPLES IN CALIBRATION SET															
F-Haaland's	5	5	5	5	9	9	8	8	3	3	8	8	2	2	3
R^2_{calib}	0.944	0.963	0.953	0.955	0.969	0.974	0.946	0.946	0.888	0.895	0.939	0.944	0.840	0.867	0.844
RMSEC (%)	0.134	0.109	0.124	0.123	0.099	0.091	0.134	0.135	0.194	0.190	0.142	0.138	0.152	0.139	0.150
R^2_{CV}	0.926	0.949	0.933	0.939	0.921	0.936	0.882	0.881	0.870	0.879	0.878	0.878	0.804	0.810	0.792
RMSECV (%)	0.156	0.133	0.151	0.148	0.162	0.149	0.202	0.206	0.214	0.208	0.205	0.199	0.174	0.163	0.187
RPD_{CV}	3.613	4.261	3.796	3.885	3.487	3.801	2.855	2.821	2.717	2.814	2.810	2.930	2.188	2.337	2.037
RER_{CV}	13.683	16.046	14.135	14.379	13.207	14.315	10.748	10.520	10.148	10.406	10.580	10.925	8.091	8.640	7.531

Table D-15: Regression statistics for PLSR models for the Klason lignin (KL), and acid insoluble residue (AIR) contents of peat samples, all on a whole mass basis (%), and for the and moisture content (MC) on a wet basis (%).

Dataset	KLASON LIGNIN (KL)						ACID INSOLUBLE RESIDUE (AIR)						MC
	DS	DS	DU	DU	WU	WU	DS	DS	DU	DU	WU	WU	WU
Pre. (1)	SG	SG	SG	SG	SG	SG	SG	SG	SG	SG	SG	SG	SNVDT
Specific (1)	1,1,10,10	1,1,10,10	2,2,25,25	2,2,25,25	2,2,25,25	2,2,25,25	1,1,10,10		1,1,10,10	1,1,10,10	2,2,25,25	2,2,25,25	2,1.1-2.5
PLS- λ 10 ³ nm	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5
Sam. Excl.		25038		25038		25038		25038		25038		25038	
Calib:Valid	40:13	39:13	38:13	37:13	40:13	39:13	40:13	39:13	38:13	37:13	40:13	40:12	40:13
F-Haaland's	5	7	5	5	7	7	5	5	5	5	9	9	3
R^2_{calib}	0.936	0.978	0.951	0.953	0.945	0.950	0.923	0.976	0.968	0.969	0.979	0.979	0.966
$Offset_{calib}$	3.692	1.279	2.774	2.695	3.139	2.871	0.158	1.457	1.914	1.866	1.267	1.267	2.021
RMSEC (%)	1.949	1.151	1.720	1.701	1.782	1.711	0.162	1.230	1.433	1.425	1.132	1.132	2.015
R^2_{CV}	0.900	0.940	0.924	0.928	0.837	0.839	0.854	0.985	0.947	0.945	0.919	0.919	0.956
RMSECV (%)	2.430	1.892	2.154	2.098	3.083	3.071	0.223	1.526	1.848	1.880	2.207	2.207	2.304
$BIAS_{CV}$	-0.044	-0.135	0.016	0.037	-0.007	0.063	0.006	-0.011	-0.042	-0.076	0.066	0.066	0.062
SECV (%)	2.461	1.912	2.183	2.126	3.123	3.110	0.226	1.545	1.873	1.904	2.234	2.234	2.333
R^2_{pred}	0.939	0.929	0.862	0.857	0.837	0.838	0.873	0.935	0.885	0.886	0.878	0.898	0.985
$Slope_{pred}$	1.080	1.022	1.051	1.044	0.978	0.964	0.948	0.896	0.908	0.913	0.907	0.943	1.018
$Offset_{pred}$	-4.644	-1.985	-3.505	-3.141	2.413	3.062	0.110	6.609	5.594	5.218	6.059	4.154	-1.657
RMSEP (%)	2.025	2.133	3.053	3.098	3.335	3.241	0.202	1.659	2.089	2.079	2.337	2.258	1.370
$Bias_{pred}$	-0.103	-0.719	-0.586	-0.649	1.186	1.050	0.001	0.506	0.190	0.142	0.469	0.769	-0.577
SEP (%)	2.105	2.090	3.119	3.153	3.245	3.192	0.211	1.644	2.165	2.159	2.383	2.218	1.294
RPD_{pred}	3.494	3.518	2.358	2.333	2.311	2.350	2.741	3.882	2.948	2.956	2.852	3.098	7.866
RER_{pred}	11.224	11.301	7.574	7.492	7.280	7.401	8.562	13.101	9.948	9.978	8.945	9.610	29.590
ALL SAMPLES IN CALIBRATION SET													
F-Haaland's	8	7	5	5	8	10	8	6	5	5	9	9	3
R^2_{calib}	0.965	0.971	0.931	0.931	0.947	0.973	0.946	0.979	0.958	0.959	0.966	0.970	0.969
RMSEC (%)	1.400	1.275	1.997	2.003	1.728	1.245	0.134	1.089	1.550	1.545	1.375	1.298	1.870
R^2_{CV}	0.929	0.941	0.905	0.902	0.849	0.875	0.882	0.968	0.942	0.935	0.916	0.929	0.963
RMSECV (%)	2.044	1.873	2.396	2.434	2.988	2.676	0.202	1.403	1.855	1.874	2.210	2.055	2.094
RPD_{CV}	3.691	4.027	3.182	3.134	2.522	2.829	2.855	5.337	4.085	4.055	3.378	3.646	5.091
RER_{CV}	14.399	15.689	12.266	12.068	9.839	11.022	10.748	20.933	15.823	15.657	13.288	14.298	23.263

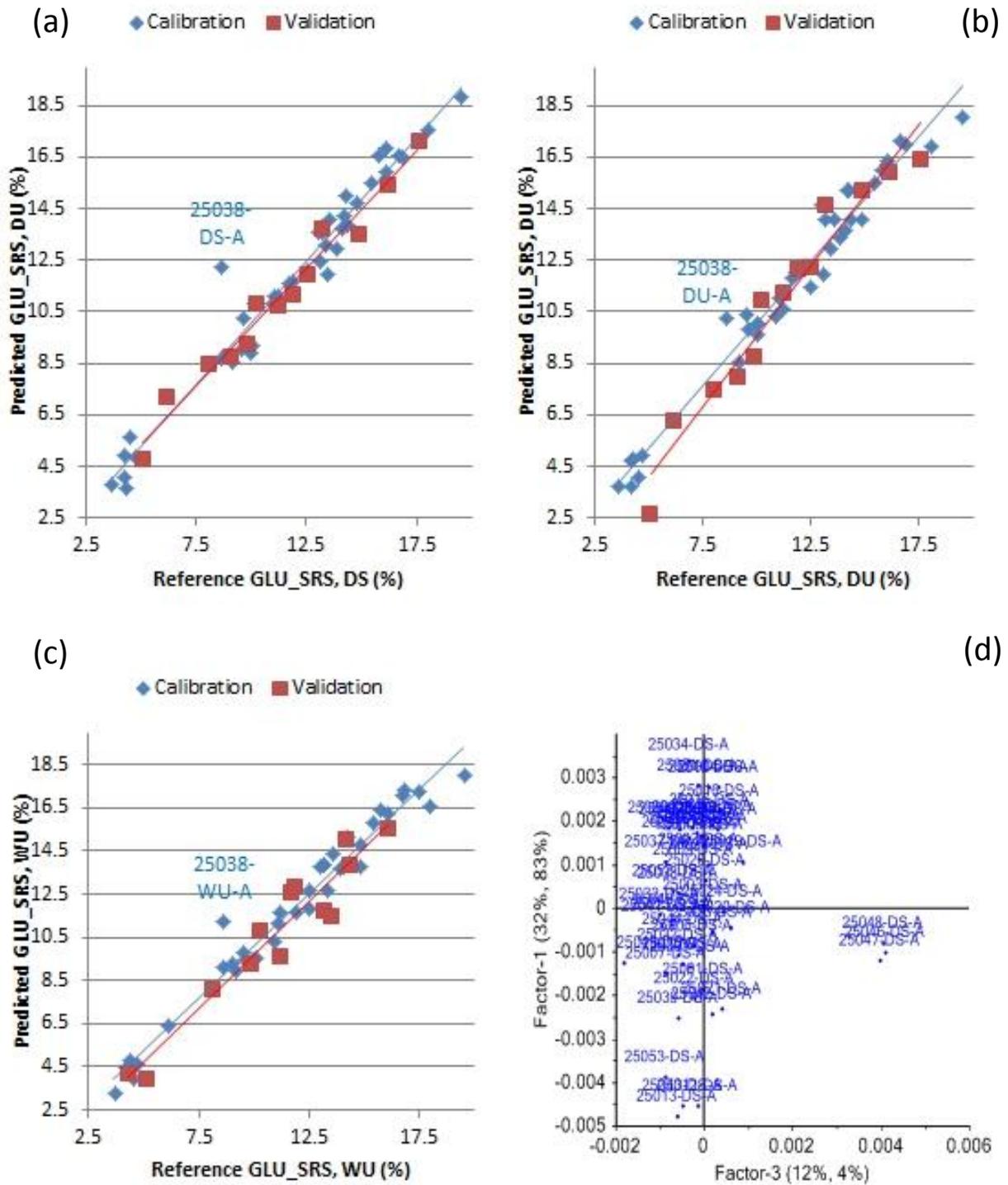


Figure D-10: Plots for GLU_SRS models. (a) Predicted glucose vs. reference glucose for the DS model, including sample 25038; (b) the same plot but for the DU model; (c) same plot but for the WU model; (d) a F1 vs. F3 scores plot for the DS GLU_SRS model with all 53 samples in the calibration set.

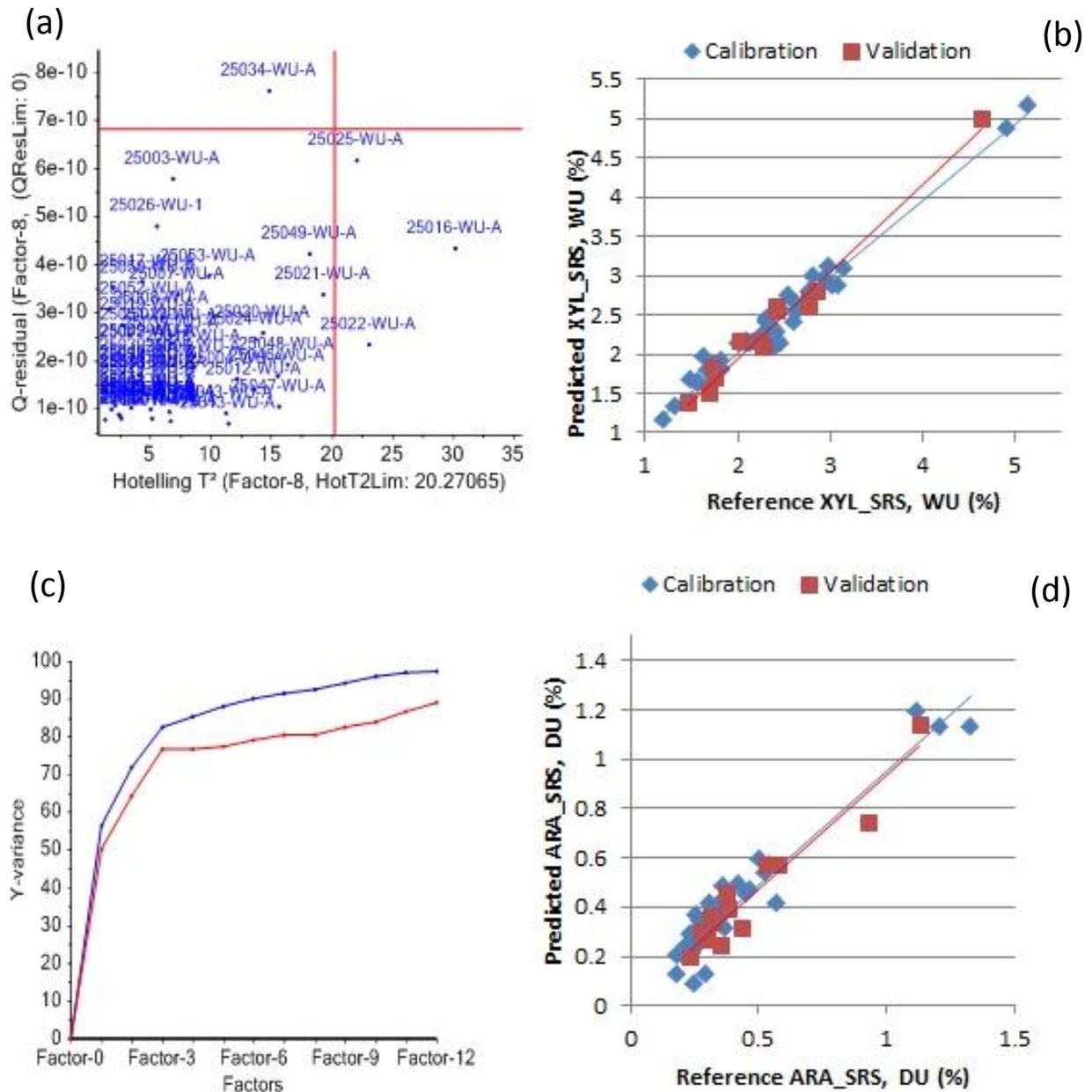


Figure D-11: Plots for glucose, xylose, and arabinose models. (a) An influence plot for the WU glucose model comprising all samples in the calibration set; (b) predicted xylose vs. reference xylose content for the WU model (sample 25038 included); (c) explained variance plot, with the number of factors in the model, for the DS arabinose calibration where all samples were in the validation set; (d) predicted arabinose vs. reference arabinose content for the DU model, sample 25038 included.

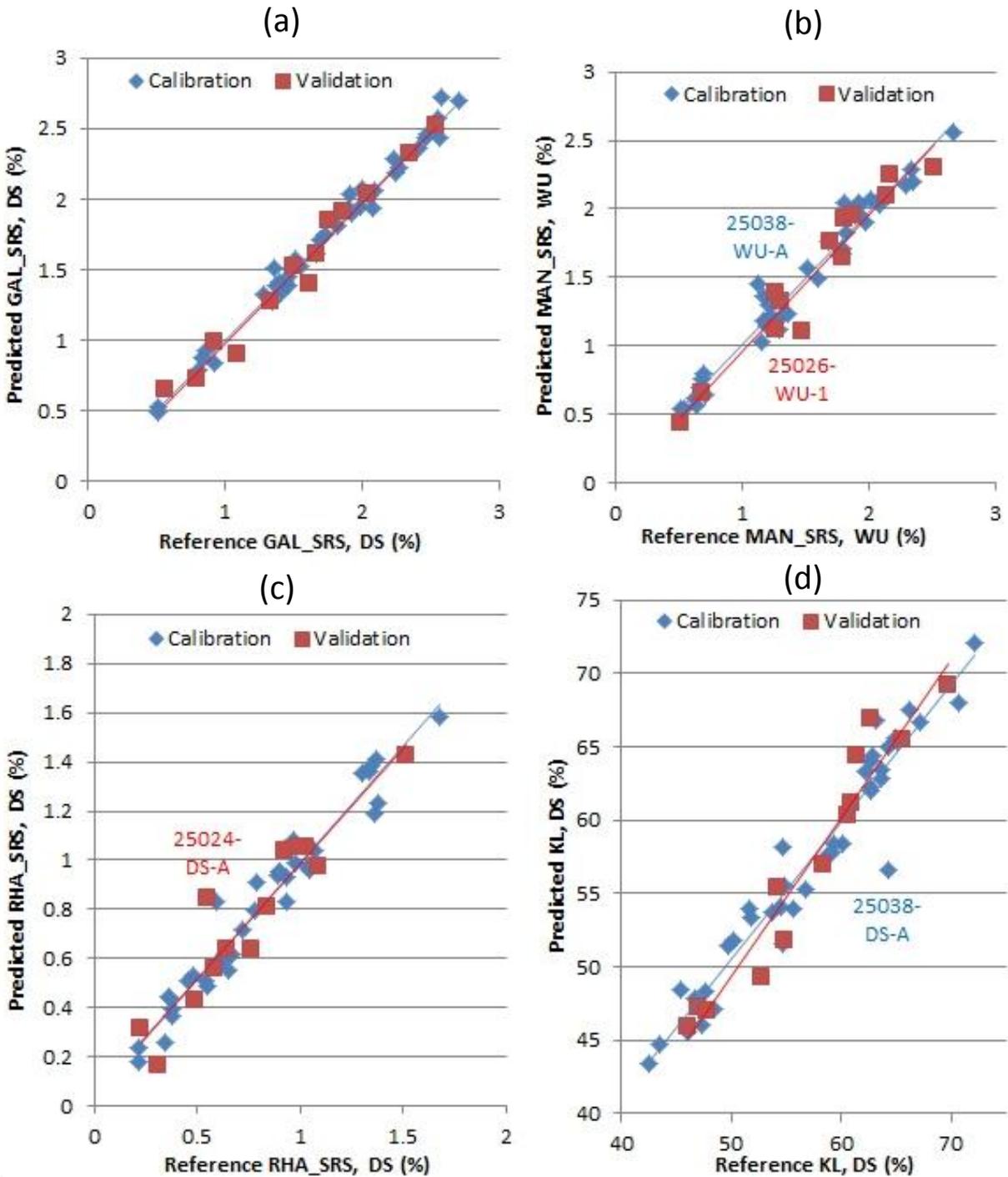


Figure D-12: Plots for galactose, mannose, rhamnose, and KL models. (a) Predicted galactose vs. reference galactose content for the DS model, sample 25038 excluded; (b) predicted mannose vs. reference mannose content for the WU model, sample 25038 included.; (c) predicted rhamnose vs. reference rhamnose content for the DS model, sample 25038 included; (d) predicted KL vs. reference KL for the DS model, sample 25038 included.

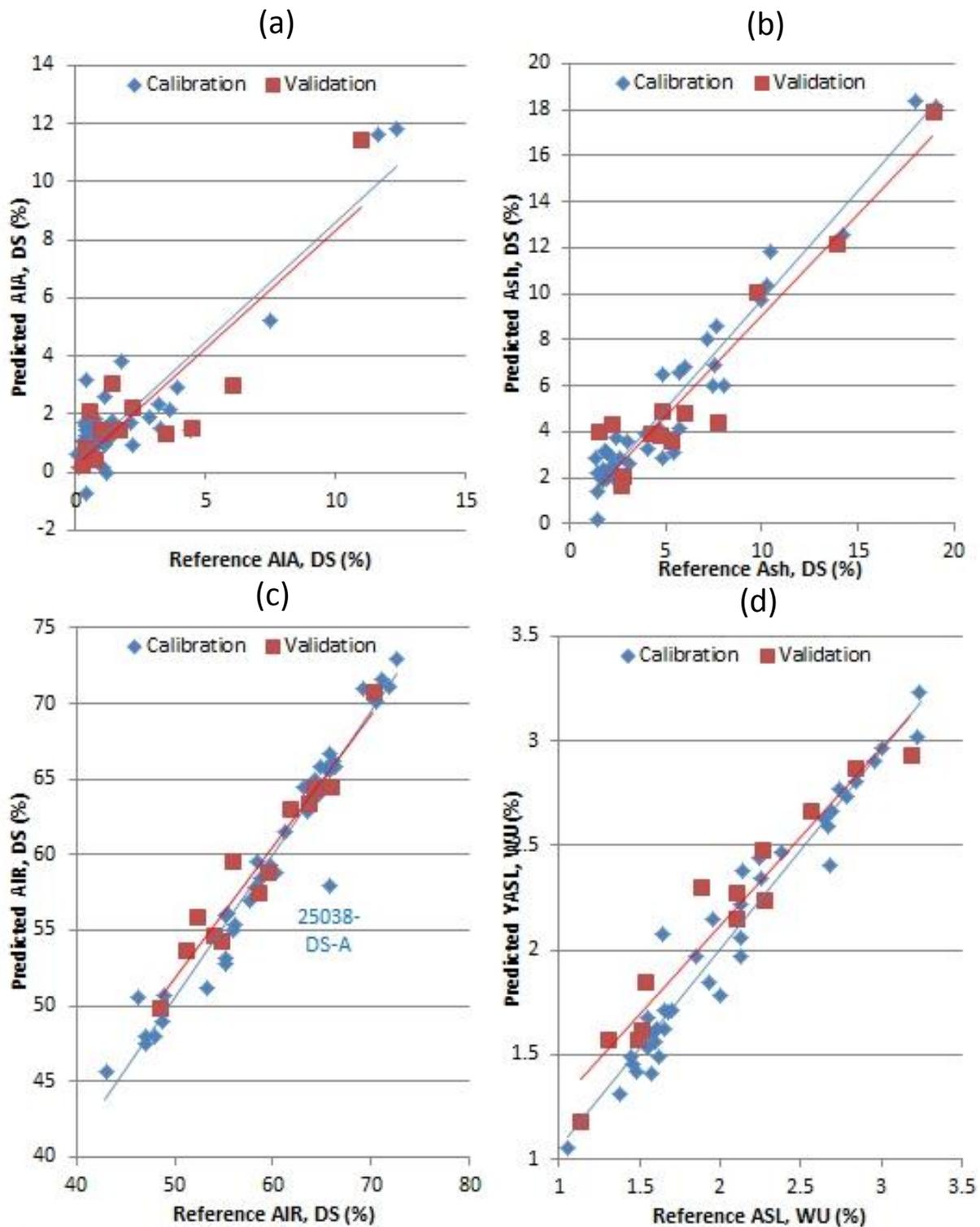


Figure D-13: Plots for AIA, ash, AIR, and ASL models. (a) Predicted AIA vs. reference AIA for the DS dataset; (b) predicted Ash vs. reference Ash for the DS dataset; (c) predicted AIR vs. reference AIR for the DS dataset; (d) predicted ASL vs. reference ASL for the WU dataset.

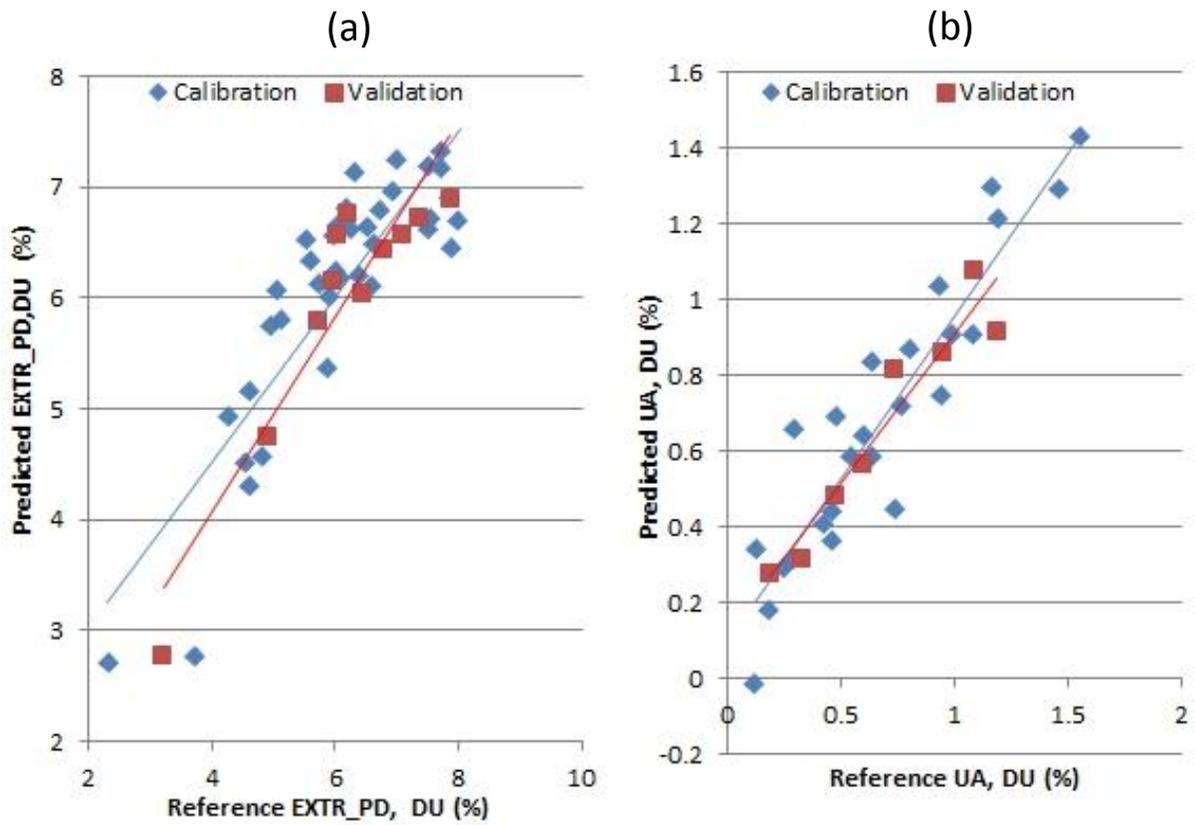


Figure D-14: Predicted y vs. reference y for: (a) EXTR_PD and the DU dataset; (b) uronic acids (UA) and the DU dataset.

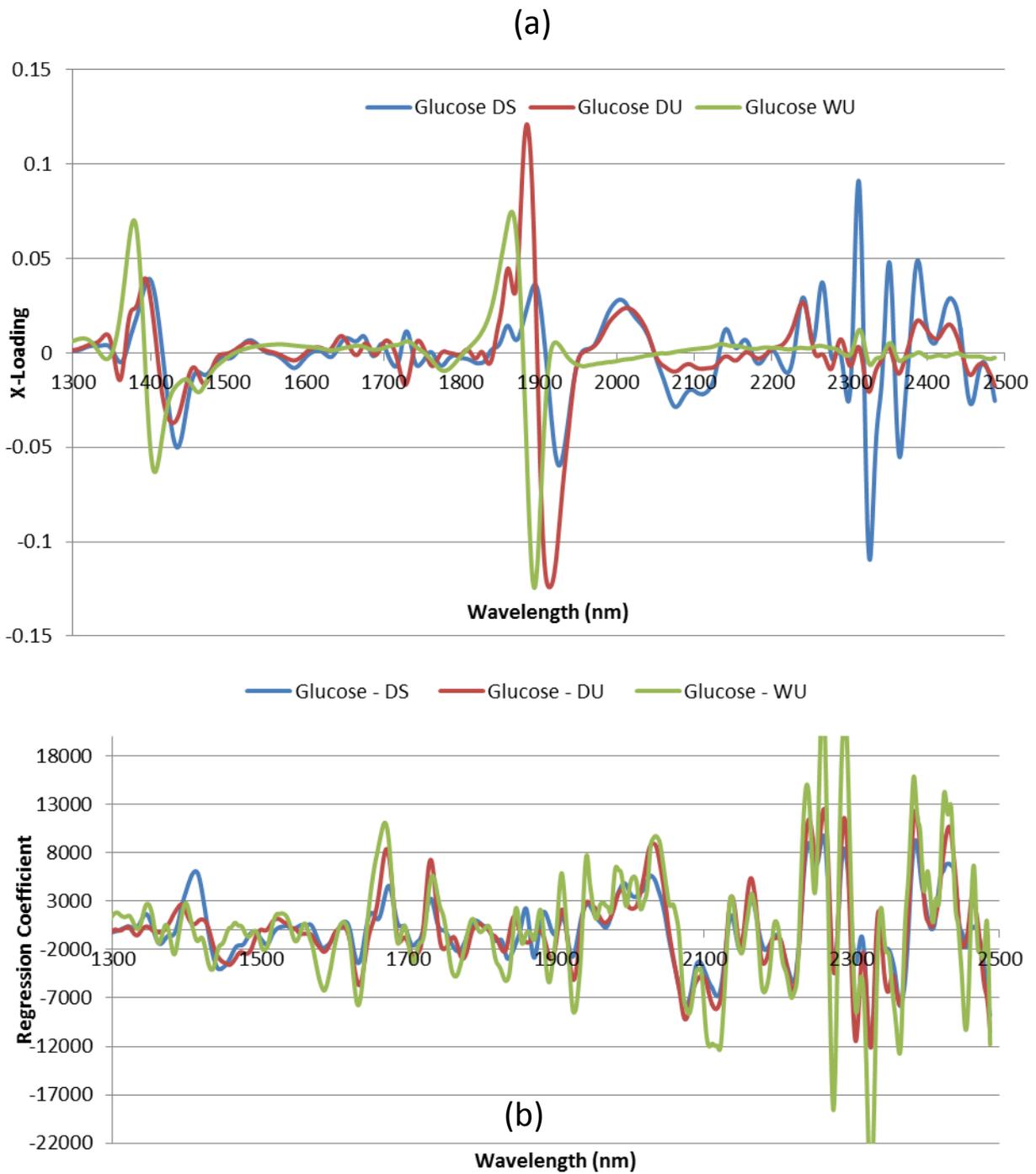


Figure D-15: (a) X-loadings for Factor 1 of the glucose PLSR models for the DS, DU, and WU datasets (b) Regression coefficients plots for these models. All spectra were transformed with SG2,2,25,25 prior to PLSR.

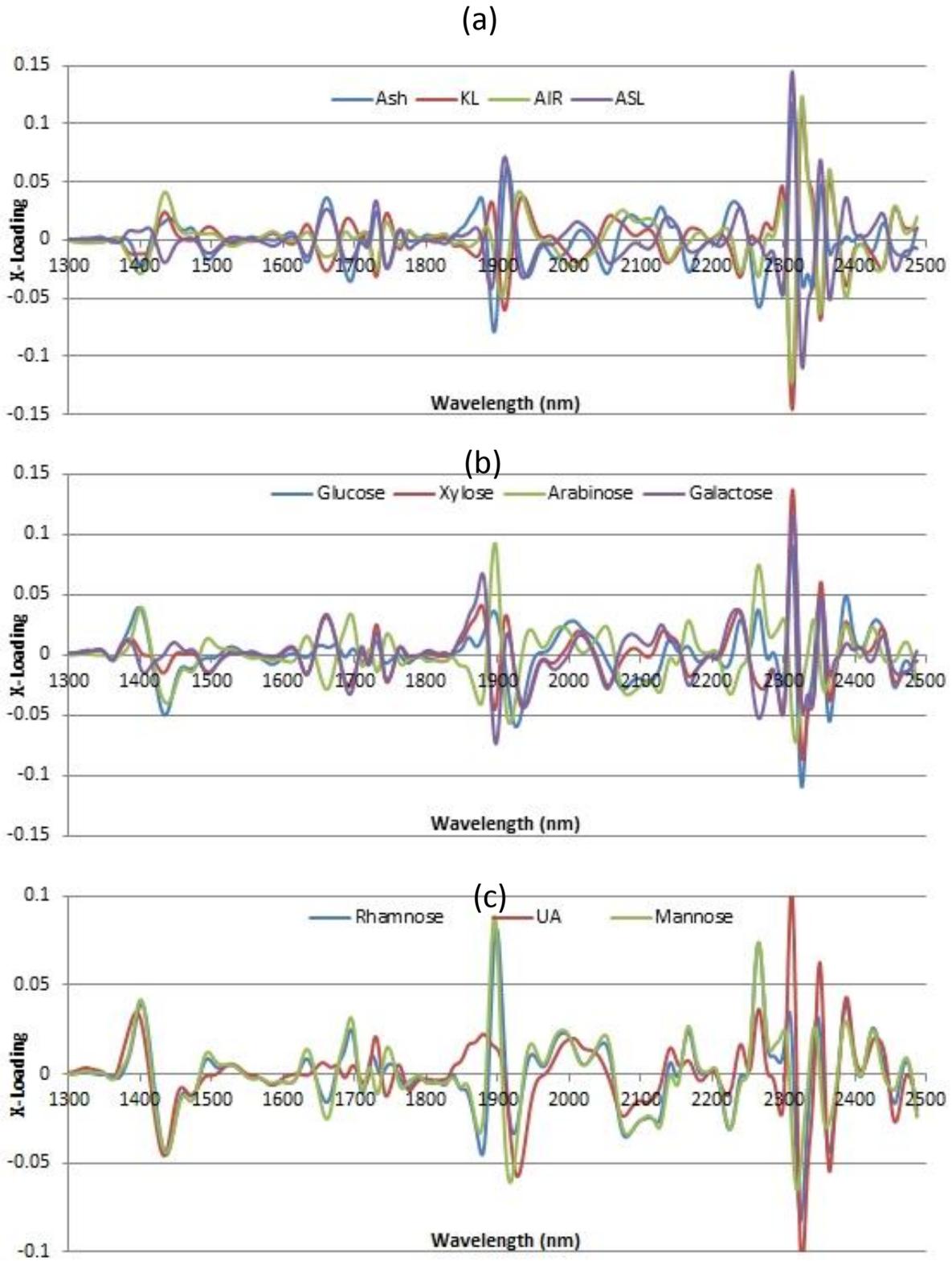


Figure D-16: X-loadings plots for F1 in PLS1 DS models for: (a) Ash, KL, AIR, and ASL; (b) Glucose, Xylose, Arabinose, Galactose; (c) Rhamnose, UA, and Mannose. All spectra were transformed by SG2,2,25,25.

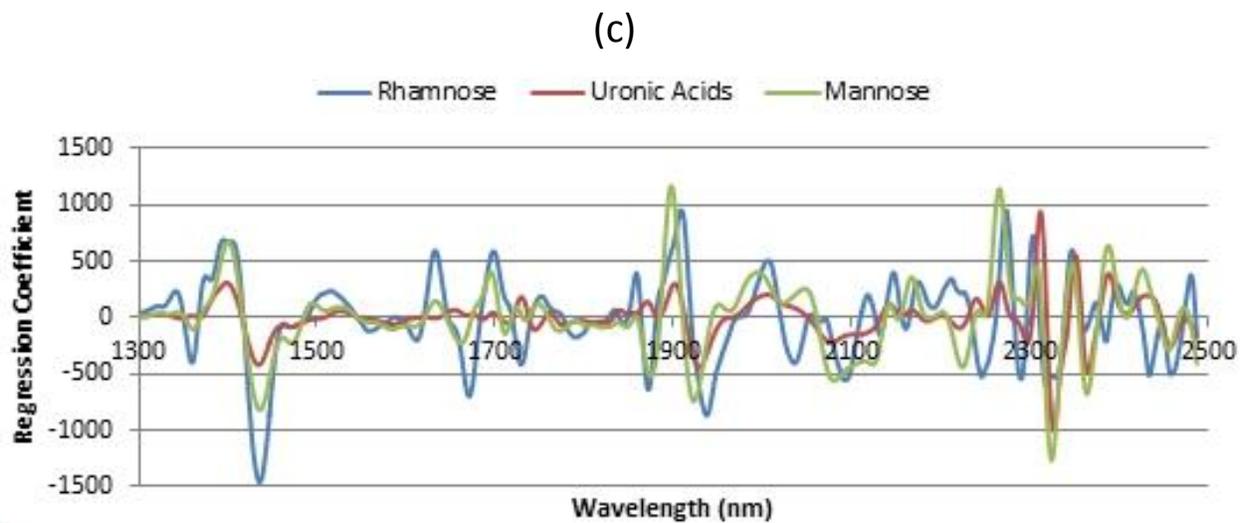
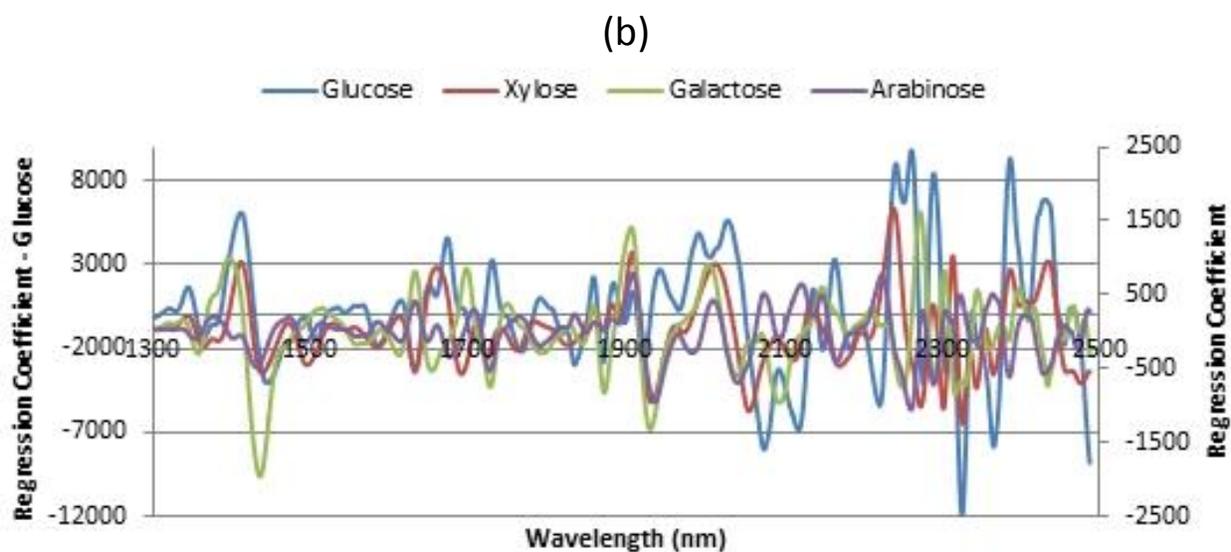
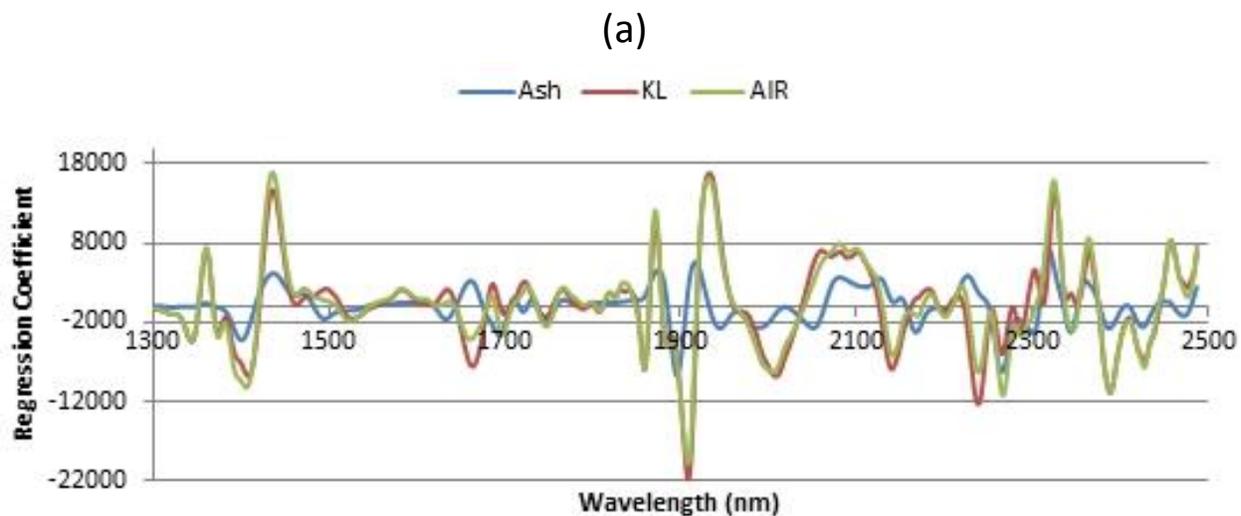


Figure D-17: Regression coefficient plots for DS PLS1 models for: (a) Ash, KL, AIR; (b) Glucose, Xylose, Arabinose, Galactose; (c) Rhamnose, Uronic Acids (UA) and Mannose. Spectra treated by SG2,2,25,25.

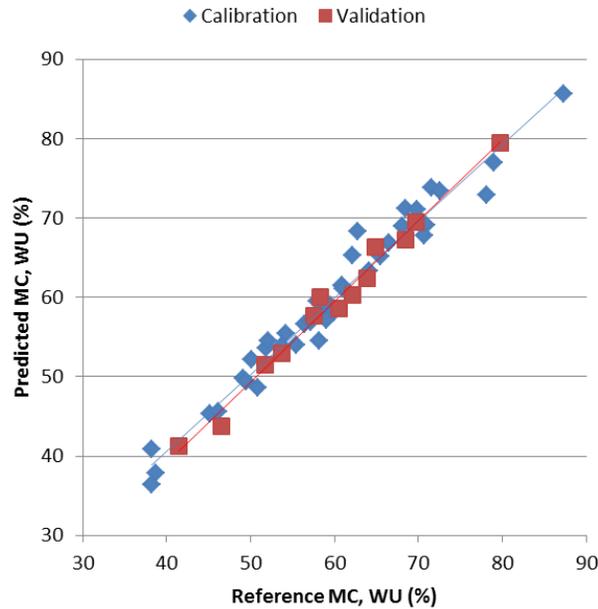


Figure D-18: Predicted moisture content (MC, % wet basis) vs. reference MC for the WU dataset.

Table D-16: Summary statistics for the moisture contents (% wet basis) for the different sample sets.

Samples	#	Av (%)	SD (%)	Max (%)	Min (%)	Range (%)
DS-E	46	11.43	1.14	14.05	9.13	4.92
DS-E Dishes	37	13.32	1.27	16.01	11.25	4.76
E-H	91	12.01	1.41	14.87	8.66	6.21

Table D-17: Regression statistics for the chosen calibrations for moisture content (% wet basis) in the extractives/hydrolysis analysis.

Samples	Pre.	λ	Calib:Valid	F	R^2_{cal}	RMSEC	RMSECV	R^2_{pred}	RMSEP	RPD_{pred}	RER_{pred}
DS-E	MSC	1100-2500	34:12	4	0.930	0.307	0.504	0.952	0.287	4.563	15.826
DS-E Dishes	MSC	1100-2500	28:9	4	0.966	0.275	0.329	0.966	0.243	5.497	14.691
E-H	MSC	1100-2500	69:22	5	0.927	0.398	0.444	0.922	0.397	3.454	12.505

Table D-18: Regression statistics for the moisture content (% wet basis) prediction of different data sets using various PLS models.

Model	DS-E	DS-E	DS-E Dishes	DS-E Dishes	E-H	E-H
Predicting	DS-E Dishes	E-H	DS-E	E-H	DS-E	DS-E Dishes
R^2_{pred}	0.942	0.811	0.811	0.757	0.915	0.915
Slope	0.918	0.838	0.795	0.768	0.903	0.831
Intercept	0.239	1.002	2.776	3.013	0.906	2.213
RMSEP (%)	0.900	1.123	0.657	0.731	0.387	0.383
Bias (%)	-0.847	-0.941	0.435	0.231	-0.201	-0.043
SEP (%)	0.307	0.616	0.498	0.697	0.334	0.385
RPD	1.407	1.259	1.739	1.934	2.953	3.311
RER	5.289	5.531	7.487	8.493	12.710	12.442

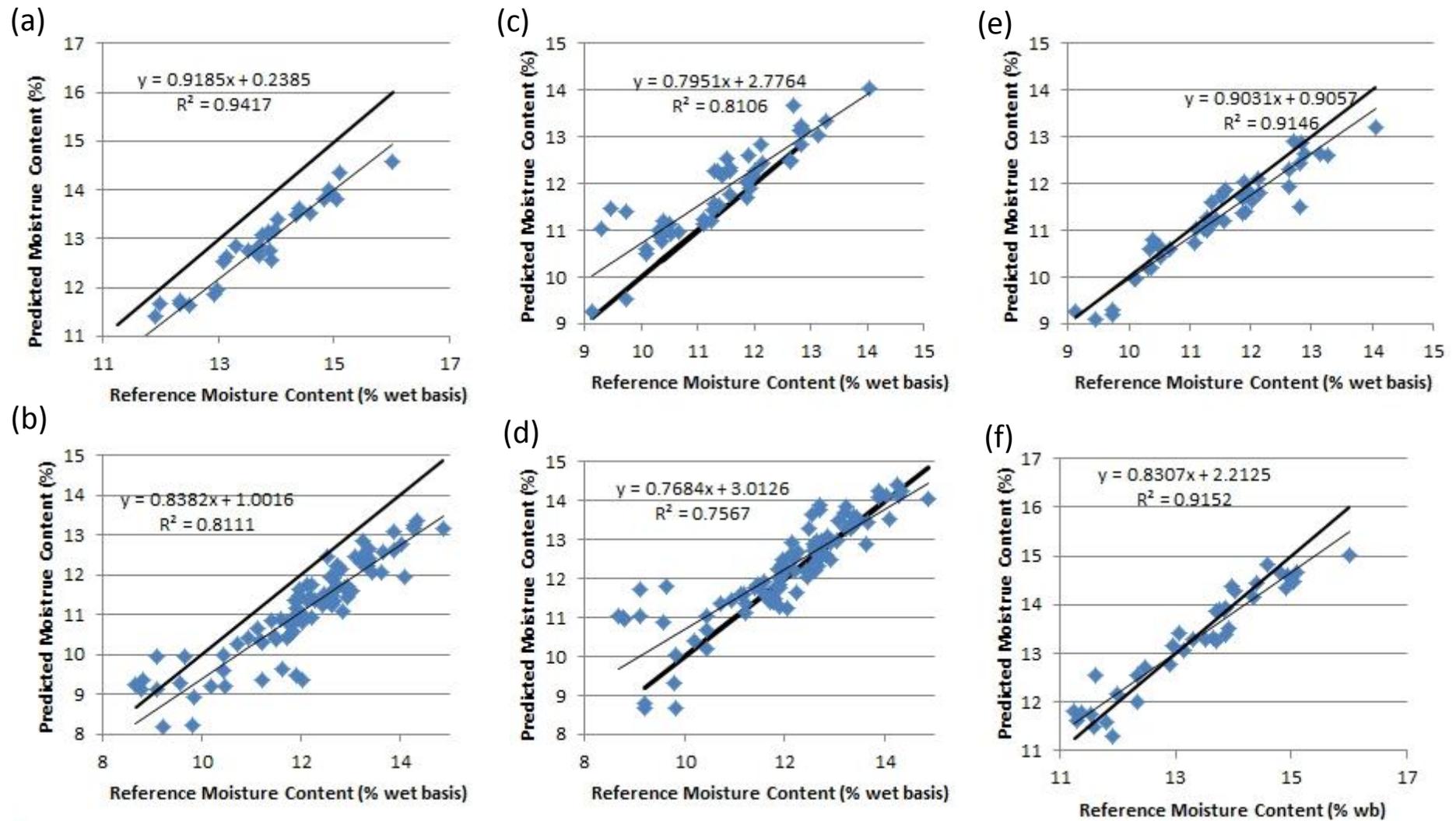


Figure D-19: Models for the moisture contents of samples prior to reference analysis are tested by predicting the moisture contents of the samples in the other models. (a) DS-E predicting DS-E Dishes; (b) DS-E predicting E-H; (c) DS-E Dishes predicting DS-E; (d) DS-E Dishes predicting E-H; (e) E-H predicting DS-E; (f) E-H predicting DS-E Dishes.

Table D-19: Differences in predicted values (% whole dry mass basis) for glucose, rhamnose and KL for replicate scans within the WU, DU, and DS datasets. See Section 13.3.4 for description of terms.

Constituent	Dataset	Bias	Av. Diff	Abs. SD	Max	Min	Av (%)	Max (%)	Min (%)
GLU_SRS	WU	-0.048	0.426	0.447	2.442	0.013	3.922%	14.389%	0.068%
GLU_SRS	DU	-0.187	0.479	0.440	1.819	0.015	5.202%	20.822%	0.240%
GLU_SRS	DS	0.133	0.329	0.447	2.631	0.002	2.937%	28.545%	0.014%
RHA_SRS	WU	0.001	0.025	0.023	0.084	0.000	4.711%	36.326%	0.011%
RHA_SRS	DU	-0.002	0.030	0.029	0.099	0.001	5.606%	44.604%	0.129%
RHA_SRS	DS	0.024	0.042	0.045	0.229	0.003	6.085%	39.294%	0.435%
KL	WU	0.081	0.822	0.816	3.225	0.008	1.491%	7.011%	0.015%
KL	DU	-0.013	0.663	0.522	2.234	0.024	1.213%	4.762%	0.040%
KL	DS	-0.326	0.678	0.717	2.935	0.018	1.234%	5.843%	0.033%

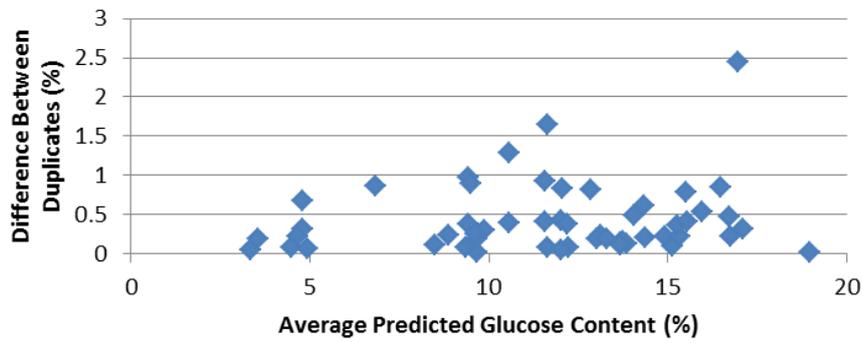


Figure D-20: Distribution of the absolute differences in predicted glucose contents (% WM DM) for the replicate WU scans, plotted according to the average predicted glucose content for that sample.

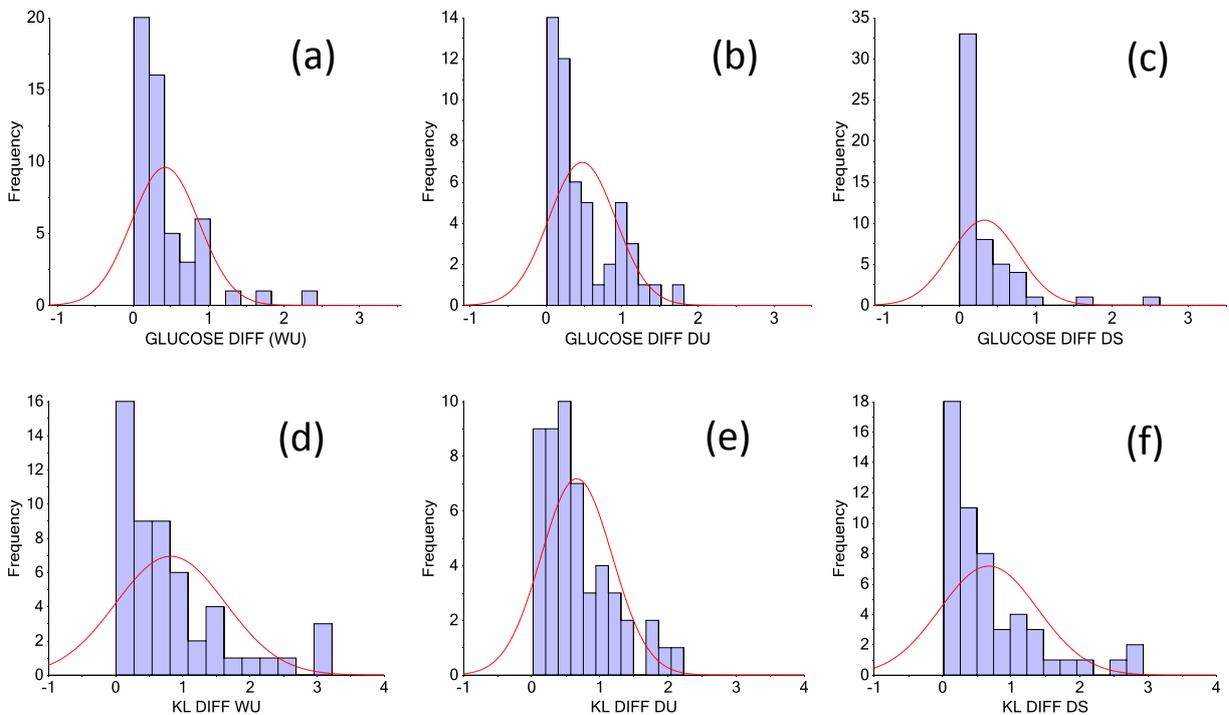


Figure D-21: Histograms for the absolute difference in predicted glucose (%) values for the (a) WU, (b) DU; and (c) DS scans; also for the difference in KL (%) values for (d) WU, (e) DU, and (f) DS scans.

Appendix E Figures and Tables for Chapter 14: Qualitative Analysis of Miscanthus

Table E-1: Summary of the different plant sections collected, according to Miscanthus variety. See Appendix A for a description of the abbreviations used.

Variety	Leaf Sections					Whole Stem Sections				Internode Sections				Node Sections				Whole Plant Sections				HP	TOTAL
	F	H	K	M	FL	X1	X2	X3	X4	X1T	X2T	X3T	X4T	X1N	X2N	X3N	X4N	WP	WP1	WP2	WP3		
<i>Giganteus</i>	72	84	30	7	3	6	5	2	0	77	61	38	2	73	65	39	2	27	15	15	7	0	630
<i>Sinensis</i>	2	2	2	2	2	2	2	0	0	0	0	0	0	0	0	0	0	1	1	1	0	0	17
Other	6	6	5	5	4	7	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	16	53
TOTAL	80	92	37	14	9	15	11	2	0	77	61	38	2	73	65	39	2	28	16	16	7	16	700

Table E-2: Summary of the total average wet spectra collected (WC and WU) according to the different plant sections and Miscanthus variety. See Appendix A for a description of the abbreviations used.

Variety	Leaf Sections					Whole Stem Sections				Internode Sections				Node Sections				Whole Plant Sections				HP	TOTAL
	F	H	K	M	FL	X1	X2	X3	X4	X1T	X2T	X3T	X4T	X1N	X2N	X3N	X4N	WP	WP1	WP2	WP3		
<i>Giganteus</i>	64	79	18	7	1	5	4	2	0	65	54	34	2	71	65	38	2	27	15	15	7	0	575
<i>Sinensis</i>	2	2	2	2	2	2	2	0	0	0	0	0	0	0	0	0	0	1	1	1	0	0	17
Other	6	6	5	4	4	7	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	36
TOTAL	72	87	25	13	7	14	10	2	0	65	54	34	2	71	65	38	2	28	16	16	7	0	628

Table E-3: Summary of the WU spectra collected according to the different plant sections and Miscanthus variety.

Variety	Leaf Sections					Whole Stem Sections				Internode Sections				Node Sections				Whole Plant Sections				HP	TOTAL
	F	H	K	M	FL	X1	X2	X3	X4	X1T	X2T	X3T	X4T	X1N	X2N	X3N	X4N	WP	WP1	WP2	WP3		
<i>Giganteus</i>	19	25	8	5	1	5	4	2	0	25	22	21	2	19	13	11	1	27	15	15	7	0	247
<i>Sinensis</i>	2	2	2	2	2	2	2	0	0	0	0	0	0	0	0	0	0	1	1	1	0	0	17
Other	6	6	5	4	4	7	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	36
TOTAL	27	33	15	11	7	14	10	2	0	25	22	21	2	19	13	11	1	28	16	16	7	0	300

Table E-4: Summary of the DU spectra collected according to the different plant sections and Miscanthus variety.

Variety	Leaf Sections					Whole Stem Sections				Internode Sections				Node Sections				Whole Plant Sections				HP	TOTAL
	F	H	K	M	FL	X1	X2	X3	X4	X1T	X2T	X3T	X4T	X1N	X2N	X3N	X4N	WP	WP1	WP2	WP3		
<i>Giganteus</i>	17	18	7	5	1	5	4	1	0	23	20	19	2	17	12	8	1	11	6	6	1	0	184
<i>Sinensis</i>	2	2	2	2	2	2	2	0	0	0	0	0	0	0	0	0	0	1	1	1	0	0	17
Other	6	6	5	5	3	7	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	16	52
TOTAL	25	26	14	12	6	14	10	1	0	23	20	19	2	17	12	8	1	12	7	7	1	16	253

Table E-5: Summary of the DV spectra collected according to the different plant sections and Miscanthus variety.

Variety	Leaf Sections					Whole Stem Sections				Internode Sections				Node Sections				Whole Plant Sections				HP	TOTAL
	F	H	K	M	FL	X1	X2	X3	X4	X1T	X2T	X3T	X4T	X1N	X2N	X3N	X4N	WP	WP1	WP2	WP3		
<i>Giganteus</i>	1	7	0	0	0	0	0	0	0	1	2	1	0	2	2	2	0	16	9	9	4	0	56
<i>Sinensis</i>	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Other	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
TOTAL	1	7	0	0	0	0	0	0	0	1	2	1	0	2	2	2	0	16	9	9	4	0	56

Table E-6: Summary of the DG spectra collected according to the different plant sections and Miscanthus variety.

Variety	Leaf Sections					Whole Stem Sections				Internode Sections				Node Sections				Whole Plant Sections				HP	TOTAL
	F	H	K	M	FL	X1	X2	X3	X4	X1T	X2T	X3T	X4T	X1N	X2N	X3N	X4N	WP	WP1	WP2	WP3		
<i>Giganteus</i>	16	17	8	5	2	6	4	2	0	22	20	19	2	16	9	7	1	9	6	6	3	0	180
<i>Sinensis</i>	2	1	2	1	1	1	2	0	0	0	0	0	0	0	0	0	0	1	1	1	0	0	13
Other	6	6	5	5	4	7	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	16	53
TOTAL	24	24	15	11	7	14	10	2	0	22	20	19	2	16	9	7	1	10	7	7	3	16	246

Table E-7: Summary of the DH spectra collected according to the different plant sections and Miscanthus variety.

Variety	Leaf Sections					Whole Stem Sections				Internode Sections				Node Sections				Whole Plant Sections				HP	TOTAL
	F	H	K	M	FL	X1	X2	X3	X4	X1T	X2T	X3T	X4T	X1N	X2N	X3N	X4N	WP	WP1	WP2	WP3		
<i>Giganteus</i>	2	8	0	0	0	0	0	0	0	3	2	2	0	4	5	4	0	8	3	3	1	0	45
<i>Sinensis</i>	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Other	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	16	16
TOTAL	2	8	0	0	0	0	0	0	0	3	2	2	0	4	5	4	0	8	3	3	1	16	61

Table E-8: Summary of the DS spectra collected according to the different plant sections and Miscanthus variety.

Variety	Leaf Sections					Whole Stem Sections				Internode Sections				Node Sections				Whole Plant Sections				HP	TOTAL
	F	H	K	M	FL	X1	X2	X3	X4	X1T	X2T	X3T	X4T	X1N	X2N	X3N	X4N	WP	WP1	WP2	WP3		
<i>Giganteus</i>	19	24	8	5	2	5	4	2	0	24	22	20	2	18	12	8	1	18	8	8	4	0	214
<i>Sinensis</i>	2	1	2	2	1	1	2	0	0	0	0	0	0	0	0	0	0	1	1	1	0	0	14
Other	6	6	5	5	4	7	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	16	53
TOTAL	27	31	15	12	7	13	10	2	0	24	22	20	2	18	12	8	1	19	9	9	4	16	281

Table E-9: Summary of the DT spectra collected according to the different plant sections and Miscanthus variety.

Variety	Leaf Sections					Whole Stem Sections				Internode Sections				Node Sections				Whole Plant Sections				HP	TOTAL
	F	H	K	M	FL	X1	X2	X3	X4	X1T	X2T	X3T	X4T	X1N	X2N	X3N	X4N	WP	WP1	WP2	WP3		
<i>Giganteus</i>	12	24	7	2	0	5	3	2	0	24	22	17	2	17	13	4	0	17	9	10	4	0	194
<i>Sinensis</i>	2	1	2	2	1	1	2	0	0	0	0	0	0	0	0	0	0	1	1	1	0	0	14
Other	5	6	3	3	4	7	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	16	47
TOTAL	19	31	12	7	5	13	8	2	0	24	22	17	2	17	13	4	0	18	10	11	4	16	255

Table E-10: Summary of the DF spectra collected according to the different plant sections and Miscanthus variety.

Variety	Leaf Sections					Whole Stem Sections				Internode Sections				Node Sections				Whole Plant Sections				HP	TOTAL
	F	H	K	M	FL	X1	X2	X3	X4	X1T	X2T	X3T	X4T	X1N	X2N	X3N	X4N	WP	WP1	WP2	WP3		
<i>Giganteus</i>	15	24	5	4	1	5	3	2	0	23	19	13	1	18	13	5	0	11	5	5	2	0	174
<i>Sinensis</i>	2	1	2	2	1	1	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	11
Other	5	3	2	2	3	7	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	16	41
TOTAL	22	28	9	8	5	13	8	2	0	23	19	13	1	18	13	5	0	11	5	5	2	16	226

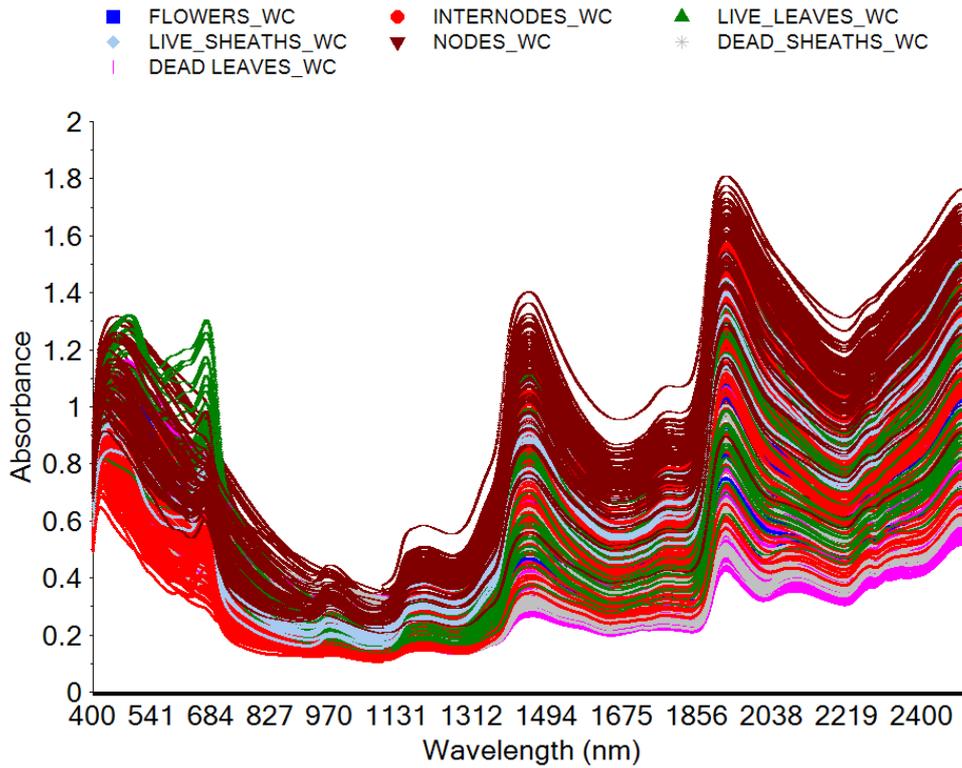


Figure E-1: The raw WU/WC spectra for separate plant fractions. The spectra of whole stem and WP samples are not included in this plot.

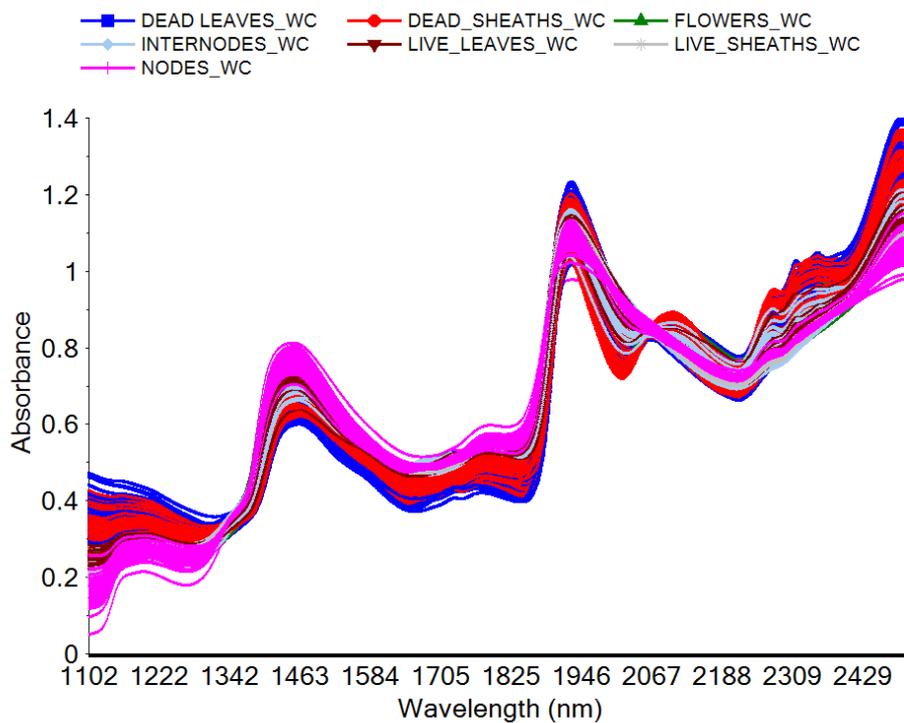


Figure E-2: The WU/WC scans, from Figure E-1, over the wavelength region 1100-2500 nm, after the MSC transform. The spectra of whole stem and WP samples are not included in this plot.

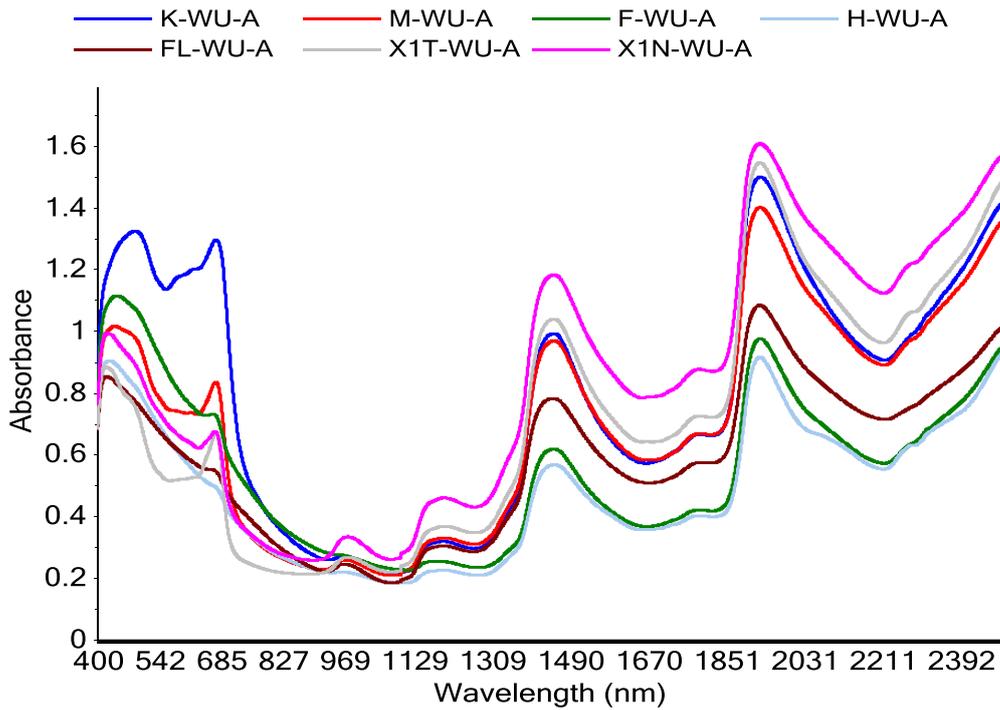


Figure E-3: The WU scans for separate fractions of a plant collected from Shanagolden on 15/11/07.

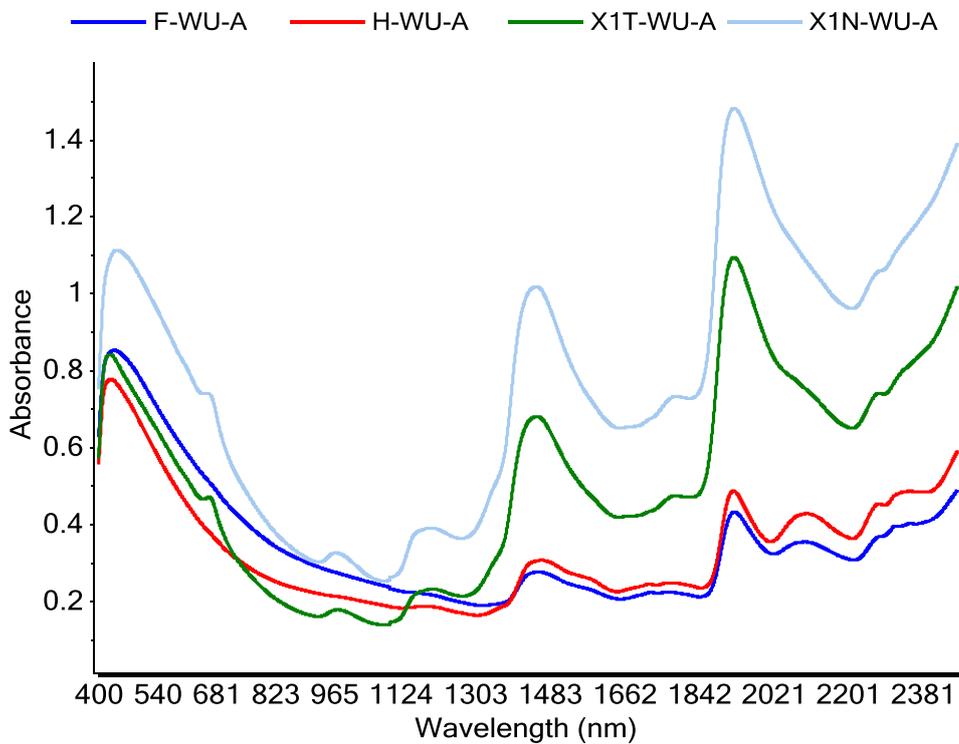


Figure E-4: The WU scans for separate fractions of a plant collected from Shanagolden on 19/3/08.

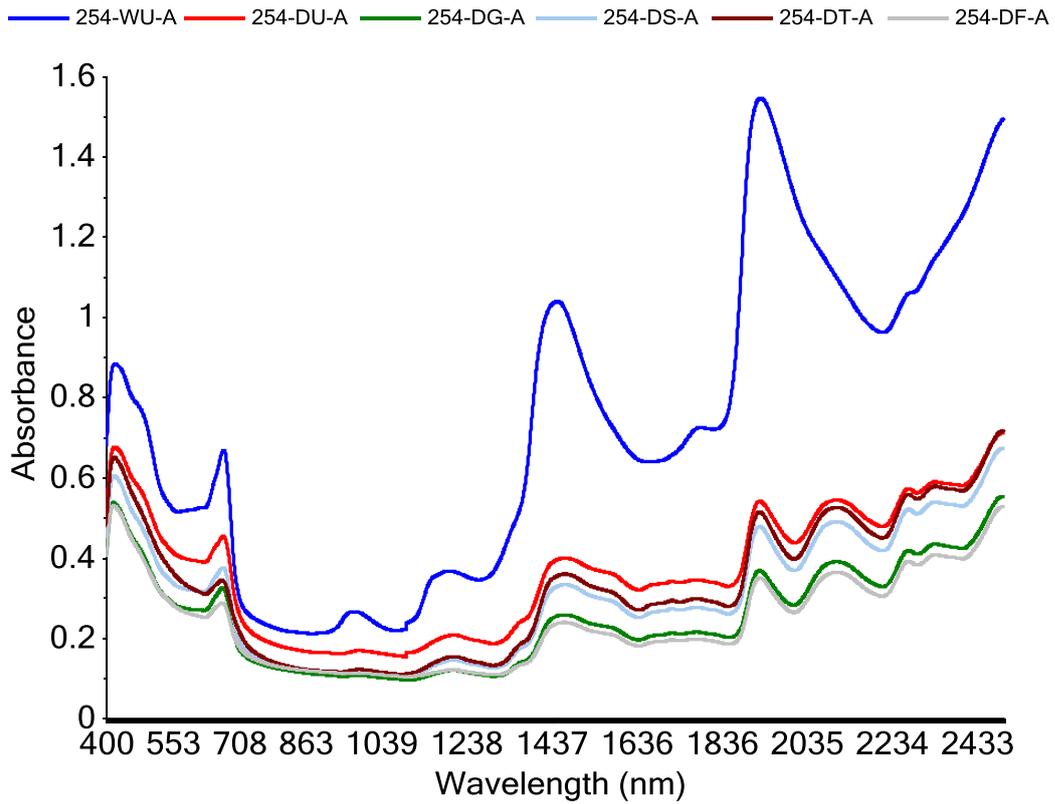


Figure E-5: Raw NIR spectra of an internode section at various sample preparation stages.

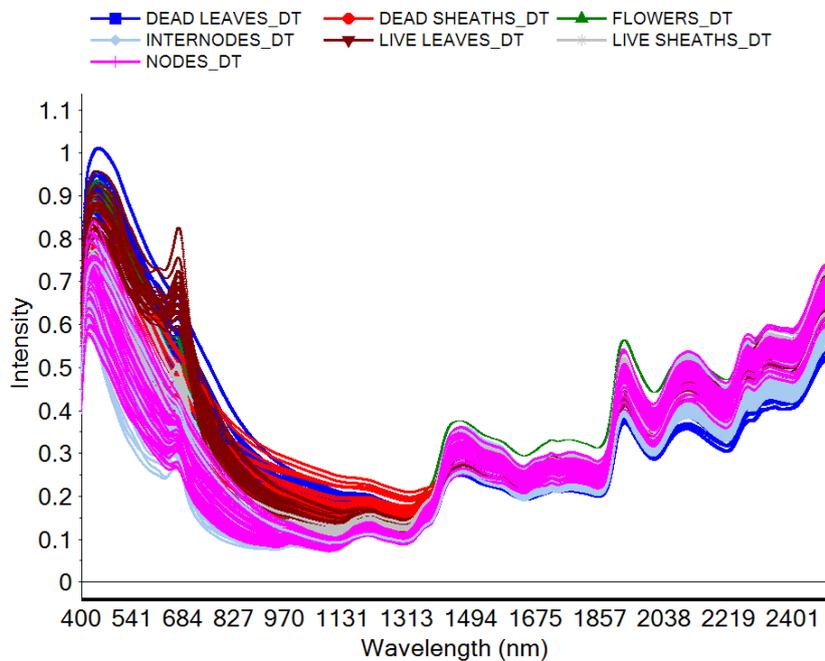


Figure E-6: The raw spectra, over 400-2500 nm, for the DT scans of various *Miscanthus* plant fractions (colour coded). The DT spectra of whole stems, WP, and HP samples are not included in this plot.

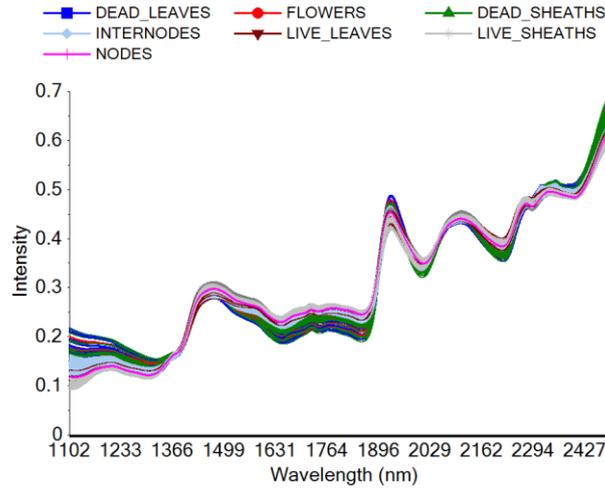


Figure E-7: The MSC-treated DT spectra, over 1100-2500 nm, for the various *Miscanthus* plant fractions (colour coded). The DT spectra of whole stems, WP, and HP samples are not included in this plot.

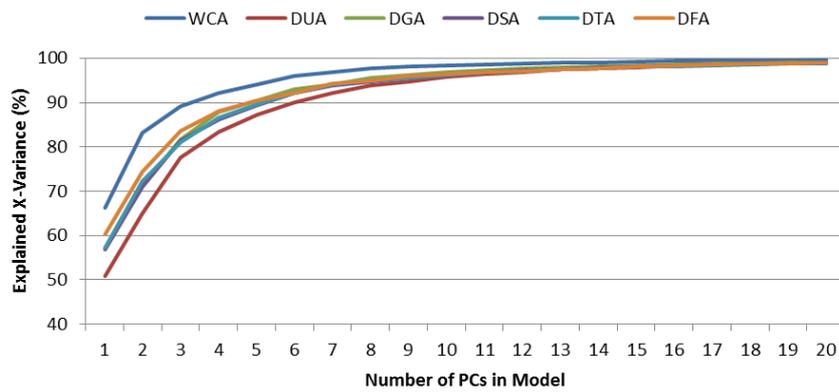


Figure E-8: An explained X-variance plot, under cross validation, with increasing number of PCs for PCA models based on the 400-2500 nm spectra for the various datasets.

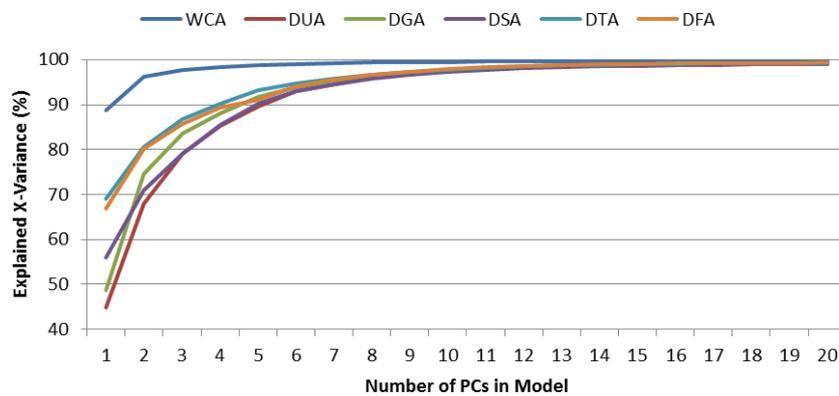


Figure E-9: An explained X-variance plot, under cross validation, with increasing number of PCs for PCA models based on the 1100-2500 nm spectra for the various datasets.

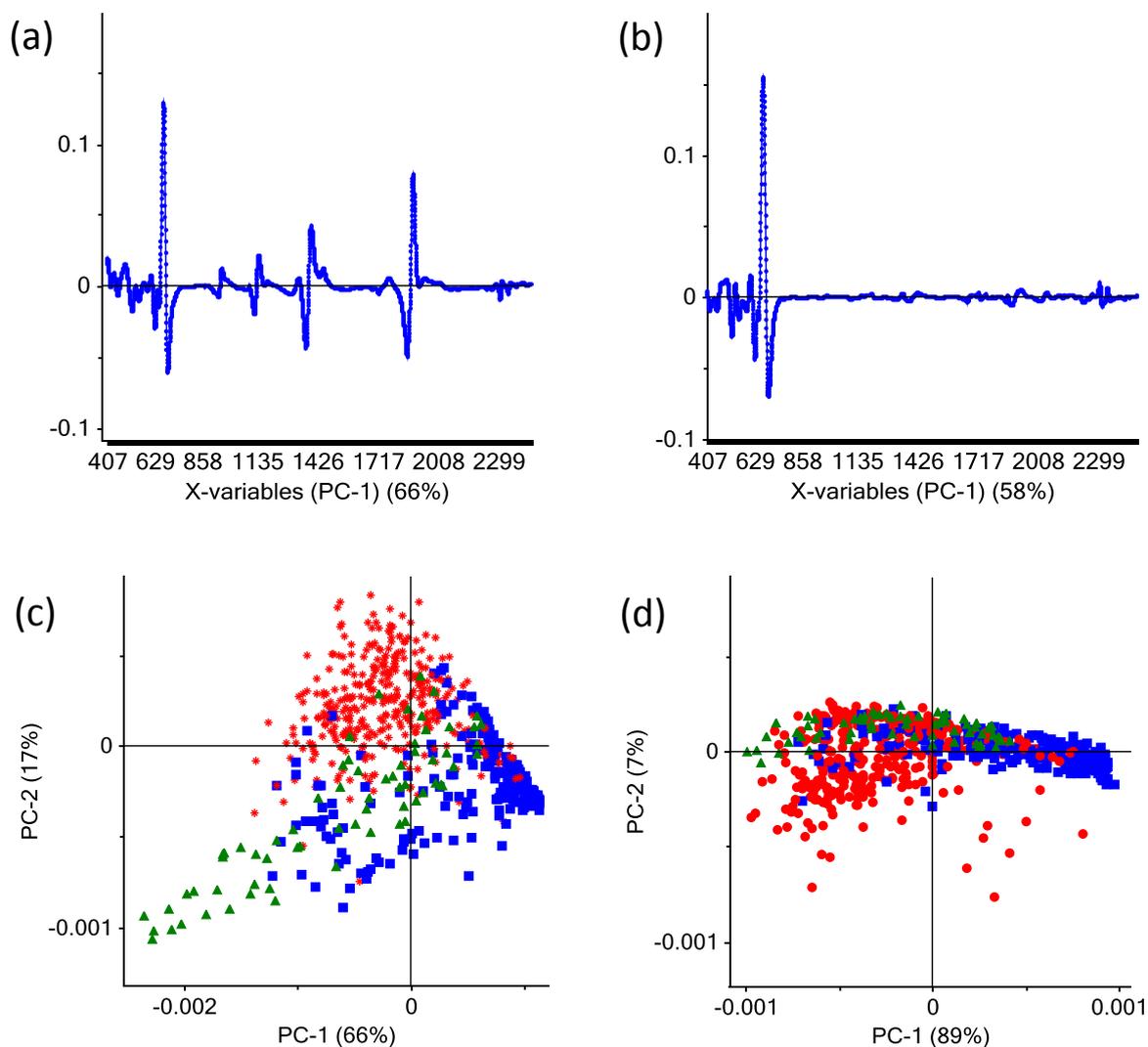


Figure E-10: Plots for PCAs on the WC/WU scans. (a) PC1 loadings plot for the WC/WU 400-2500 nm model; (b) PC1 loadings plot for the DG 400-2500 nm model; (c) PC1 vs. PC2 scores plot for the WC/WU 400-2500 nm model featuring leaves (blue squares), stems (red stars) and WP samples (green triangles); (d) same as (c) except for the 1100-2500 nm model.

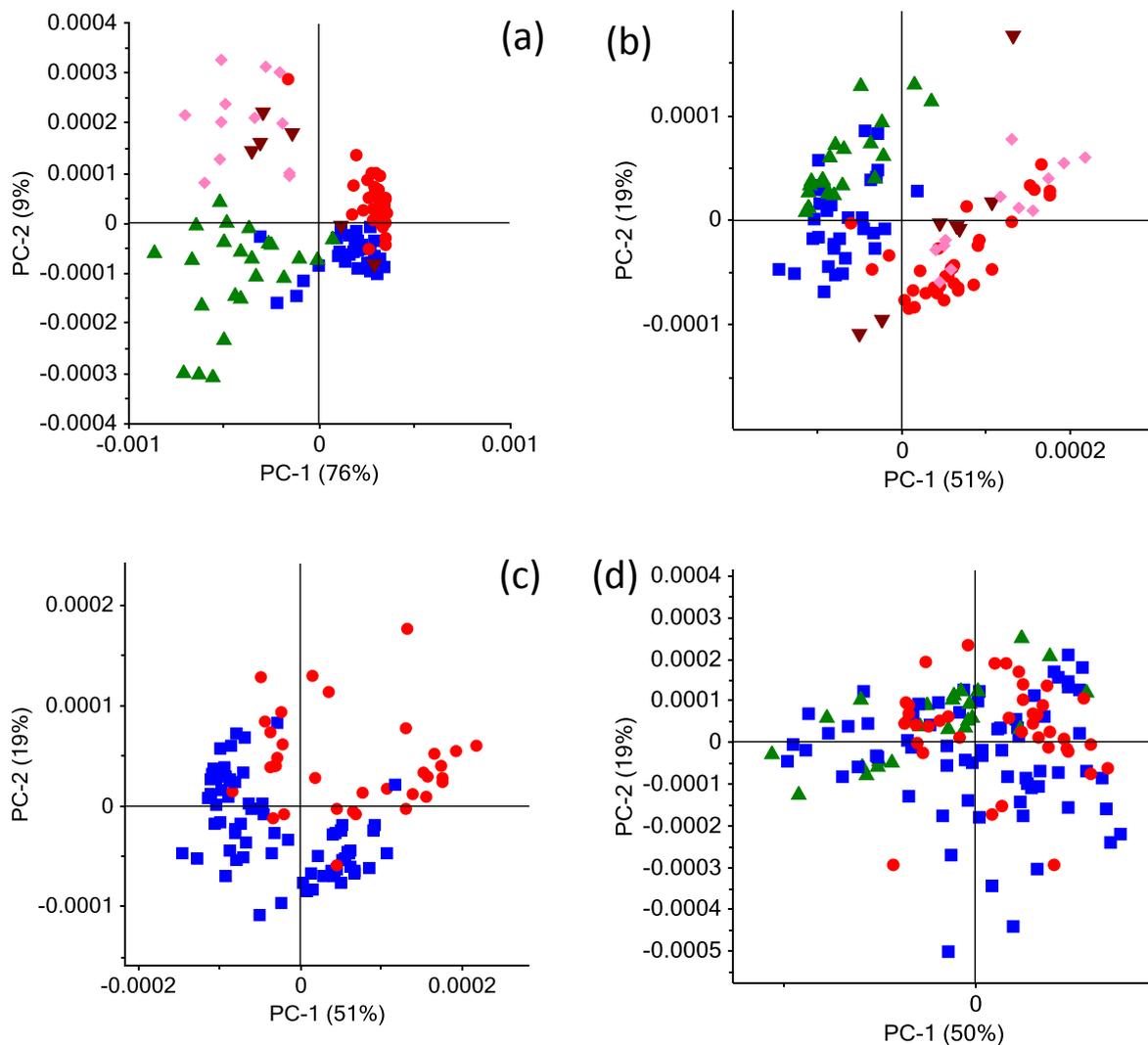


Figure E-11: Plots for PCAs on the DS scans. (a) PC1 vs. PC2 scores plot for a 400-2500 nm PCA model of the leaves in the DS dataset with dead leaves (blue squares), dead sheaths (red circles), live leaves (green triangles), live sheaths (pink diamonds) and flowers (brown inverted triangles); (b) same as (a) but for a model based on the 1100-2500 nm region; (c) the same as (b) but the samples are coloured according to whether they are from *Miscanthus x giganteus* plants (blue squares) or other varieties of *Miscanthus* (red circles); (d) a scores plot for a 400-2500 nm PCA model of the stem samples in the DS dataset with internodes (blue squares), nodes (red circles), and whole stem samples (green triangles).

Table E-11: Results for K-Means Clustering (KMC) and K-Medians Clustering (KDC) using various distance measures. The DF dataset comprised all of the samples listed in Table E-10 with the exception of the WP and HP samples.

Distance	K- Means Clustering (KMC)			K- Medians Clustering (KDC)		
	Stems (%)	Leaves (%)	Av. (%)	Stems (%)	Leaves (%)	Av. (%)
Euclidean	88.7	92.9	90.8	86.1	92.9	89.5
City-Block	93.0	92.9	93.0	86.1	92.9	89.6
Correlation	86.1	92.9	89.5	86.1	92.9	89.5
Uncentred Corr.	86.1	92.9	89.5	86.1	92.9	89.5
Spear Rank	88.7	90.5	89.6	-	-	-
Kendall's Tau	88.7	92.9	90.8	-	-	-
Cheby's Distance	86.1	94.0	90.1	86.1	92.9	89.5
Bray Curtis Distance	93.0	92.9	93.0	90.4	97.6	94.0

Table E-12: Results for the assignment of samples to one of 2 clusters using various spectral ranges and pretreatments on the DF dataset.

Wavelength (nm)	Pretreatment	Euclidean		City-Block		Bray-Curtis	
		Stems (%)	Leaves (%)	Stems (%)	Leaves (%)	Stems (%)	Leaves (%)
400-2500	NONE	86.1	92.9	86.1	92.9	90.4	97.6
400-750	NONE	87.8	92.9	87.0	92.9	86.1	94.0
400-1100	NONE	87.8	97.6	86.1	92.9	86.1	92.9
1100-2500	NONE	73.9	97.6	93.9	92.9	88.7	92.9
400-2500	SG2,2,10,10	90.4	92.9	94.8	90.5	84.3	95.2
400-1100	SG2,2,10,10	87.8	97.6	67.0	63.1	59.1	64.3
400-2500	SG1,2,8,8	81.7	70.2	85.2	92.9	92.2	90.5
400-1100	SG1,2,8,8	75.7	70.2	76.5	91.7	73.9	92.9
400-2500	SNV (400-2500 nm)	94.8	90.5	64.3	64.3	All but 1 in same cluster	
400-2500	MSC (400-2500 nm)	67.5	63.1	86.8	92.9	86.0	92.9
1100-2500	MSC (1100-2500 nm)	100	58.3	100	54.8	100	54.8
400-2500	MSC (1100-2500 nm)	88.7	92.9	90.4	92.9	90.4	95.2

Table E-13: Results for various hierarchical clustering methods tested for different wavelength regions, using different distance measures and different spectral pretreatments, on the DF dataset. The distance measure used is not relevant for Ward's Method. HSL = hierarchical single linkage, HCL = hierarchical complete linkage, HAL = hierarchical average linkage, HML = hierarchical median linkage.

Pretreatment	None		None		None		SG (2,2,10,10)	
Wavelength	1100-2500 nm		400-1100 nm		400-2500 nm		400-1100 nm	
Distance	City Block		Euclidean		Bray Curtis		Euclidean	
Linkage	Stems (%)	Leaves (%)						
HSL	All but 1 in 1 cluster		All but 2 in 1 cluster		All but 1 in 1 cluster		All but 1 in 1 cluster	
HCL	50.4	97.6	88.7	91.7	87.8	100	Most in Same Cluster	
HAL	82.6	97.6	88.7	91.7	83.5	100	Most in Same Cluster	
HML	82.6	97.6	88.7	91.7	99.1	84.5	Most in Same Cluster	
Wards	100	54.8	93.0	91.7	80.9	100	52.2	67.9

Table E-14: The classification (leaf or stem samples) for K-median clustering and hierarchical clustering linkage (HCL) methods for the various datasets.

Dataset	K Median Clustering					Hierarchical Clustering					
	Pretreat.	Wavel.	Dist.	% Stems	% Leaves	Pretreat.	Wavel.	Dist.	Link	% Stems	% Leaves
DF	None	400-2500	BC	90.4	97.6	None	400-2500	BC	HCL	87.8	100
DS	None	400-2500	BC	84.8	91.6	None	400-2500	BC	HCL	97.7	66.4
DT	None	400-2500	BC	88.5	92.6	None	400-2500	BC	HCL	96.7	76.6
DG	None	400-2500	BC	81.1	96.3	None	400-2500	BC	HCL	69.7	93.8
DH	None	400-2500	BC	100	81.5	None	400-2500	BC	HCL	100	85.2
DU	SG2,2,14,14	1100-2500	BC	90.6	66.3	SG2,2,14,14	1100-2500	BC	HCL	78.7	72.3
DV	None	1100-2500	BC	100	74.2	None	400-2500	BC	HCL	90	100
WC/WU	None	400-2500	BC	87.4	79.8	None	400-2500	BC	HCL	79.9	86.8

Table E-15: Results for the use of PLS-DA models on the DF validation set in order to discriminate between Early and Late harvest Miscanthus samples.

Model #	*1		*2		*3		*4		*5		*6		*7		*8	
Varieties	All		All		All		Giganteus		Giganteus		Giganteus		Giganteus		Giganteus	
Pre.	SG		SG		SG		SG		None		None		SG		MSC	
Specific	2,2,14,14		2,2,14,14		2,2,14,14		2,2,14,14						2,2,14,14		F,1.1-2.5	
λ (nm)	400-2500		400-1100		1100-2500		1100-2500		1100-2500		400-2500		400-2500		1100-2500	
Cal:Val	100:34		100:34		100:34		99:34 (no flower)		99:34 (no flower)		100:34		99:34 (no flower)		100:34	
# Factors	11		11		13		13		8		13		11		7	
Score	Actual		Actual		Actual		Actual		Actual		Actual		Actual		Actual	
Predict	Early	Late	Early	Late	Early	Late	Early	Late	Early	Late	Early	Late	Early	Late	Early	Late
Early	29 (90.6%)	1 (10%)	30 (93.8%)	1 (10%)	31 (96.9%)	0	23 (100%)	0	22 (95.7%)	1 (9.1%)	22 (95.7%)	1 (9.1%)	21 (91.3%)	1 (9.1%)	23 (100%)	1 (9.1%)
Late	3 (9.4%)	9 (90%)	2 (6.3%)	9 (90%)	1 (3.1%)	10 (100%)	0	11 (100%)	1 (4.3%)	10 (90.9%)	1 (4.3%)	10 (90.9%)	2 (8.7%)	10 (90.9%)	0	10 (90.9%)
Dev.	Actual		Actual		Actual		Actual		Actual		Actual		Actual		Actual	
Early	19 (59.4%)	0	19 (59.4%)	1 (10%)	26 (81.3%)	0	20 (87%)	0 (0%)	13 (56.5%)	0	19 (82.6%)	0	15 (65.2%)	0 (0%)	13 (56.5%)	0
Late	0	7 (70%)	0	4 (40%)	0	5 (50%)	0	7 (63.6%)	0	6 (54.5%)	0	10 (90.9%)	0	8 (72.7%)	0	5 (45.5%)
None	13 (40.6%)	3 (30%)	13 (40.6%)	5 (50%)	6 (18.8%)	5 (50%)	3 (13%)	4 (36.4%)	10 (43.5%)	5 (45.5%)	4 (17.4%)	1 (9.1%)	8 (34.8%)	3 (27.3%)	10 (43.5%)	6 (54.5%)

Table E-16: Validation results for the best PLS-DA models, used to discriminate between Early and Late harvest samples, for various spectral datasets.

Dataset	DF		DT		DS		DG		DU		WC+WU	
Varieties	Giganteus		Giganteus		Giganteus		Giganteus		Giganteus		Giganteus	
Pre.	SG		SG		SG		SG		None		SG	
Spec.	2,2,14,14		2,2,14,14		2,2,14,14		2,2,20,20				2,2,20,20	
λ (nm)	1100-2500		400-2500		400-2500		400-2500		400-2500		400-2500	
Cal:Val	99:34 (no flower)		115:39		124:42		92:31		96:30		269:90	
# Factors	13		12		9		14		9		14	
Score	Actual		Actual		Actual		Actual		Actual		Actual	
Predict	Early	Late	Early	Late	Early	Late	Early	Late	Early	Late	Early	Late
Early	23 (100%)	0	28 (90.3%)	0	30 (100%)	2 (16.7%)	21 (91.3%)	1 (12.5%)	18 (94.7%)	0	59 (96.7%)	0
Late	0	11 (90.9%)	3 (9.7%)	8 (100%)	0	10 (83.3%)	2 (8.7%)	7 (87.5%)	1 (5.3%)	11 (100%)	2 (3.3%)	29 (100%)
Dev.	Actual		Actual		Actual		Actual		Actual		Actual	
Early	20 (87%)	0 (0%)	22 (71%)	0	21 (70%)	1 (8.3%)	19 (82.6%)	0 (0%)	14 (73.7%)	0	49 (80.3%)	0 (0%)
Late	0	7 (63.6%)	0	7 (87.5%)	0	8 (66.7%)	0	7 (87.5%)	0	11 (100%)	0	22 (75.9%)
None	3 (13%)	4 (36.4%)	9 (29%)	1 (12.5%)	9 (30%)	3 (25%)	4 (17.4%)	1 (12.5%)	5 (26.3%)	0	12 (19.7%)	7 (24.1%)

Table E-17: Classifications, to discriminate between early and late harvest samples, for the prediction and validation sets for LDA/QDA on the WC/WU dataset.

Pre.	SG		SG		SG		SG		SG		SG	
Spec.	2,2,25,25		2,2,25,25		2,2,25,25		2,2,25,25		2,2,25,25		2,2,25,25	
λ (nm)	400-2500		400-2500		400-2500		400-2500		400-2500		400-2500	
LDA/QDA	LDA		QDA		LDA		QDA		LDA on PLS-DA		QDA on PLS-DA	
PCs Used	10		10		20		20		0		0	
Calibration	Actual		Actual		Actual		Actual		Actual		Actual	
Predict	Early	Late	Early	Late	Early	Late	Early	Late	Early	Late	Early	Late
Early	145 (84.3%)	7 (7.2%)	143 (83.1%)	2 (2.1%)	152 (88.4%)	1 (1%)	158 (91.9%)	0	168 (97.7%)	1 (1%)	157 (91.3%)	1 (1%)
Late	27 (15.7%)	90 (92.8%)	29 (16.9%)	95 (97.9%)	20 (11.6%)	96 (99%)	14 (8.1%)	97 (100%)	4 (2.3%)	96 (99%)	15 (8.7%)	96 (99%)
Validation	Actual		Actual		Actual		Actual		Actual		Actual	
Early	55 (90.2%)	2 (6.9%)	53 (86.9%)	2 (6.9%)	56 (91.8%)	1 (3.4%)	55 (90.2%)	2 (6.9%)	59 (96.7%)	1 (3.4%)	57 (93.4%)	0
Late	6 (9.8%)	27 (93.1%)	8 (13.1%)	27 (93.1%)	5 (8.2%)	28 (96.6%)	6 (9.8%)	27 (93.1%)	2 (3.3%)	28 (96.6%)	4 (6.6%)	29 (100%)

Table E-18: Validation results for the best PLS-DA models used to discriminate between giganteus samples and samples of other varieties (Not Gig), for various spectral datasets.

Dataset	DF		DS		DT		DG		DU		WC+WU	
Pre.	SG		SG		SG		SG		SG		SG	
Spec.	2,2,20,20		2,2,20,20		2,2,20,20		2,2,20,20		2,2,20,20		2,2,25,25	
λ (nm)	400-2500		400-2500		400-2500		400-2500		400-2500		400-2500	
<i>Gig:NotGig</i>	47:36		56:51		51:45		57:50		56:53		74:53	
Cal:Val	62:21		80:27		72:24		80:27		81:28		95:32	
# Factors	5		5		9		5		7		13	
Score	Actual		Actual		Actual		Actual		Actual		Actual	
<u>Predict</u>	<i>Giganteus</i>	Not Gig	<i>Giganteus</i>	Not Gig	<i>Giganteus</i>	Not Gig	<i>Giganteus</i>	Not Gig	<i>Giganteus</i>	Not Gig	<i>Giganteus</i>	Not Gig
<i>Giganteus</i>	12 (100%)	1 (11.1%)	14 (100%)	2 (15.4%)	10 (83.3%)	1 (8.3%)	14 (93.3%)	0 (0%)	13 (92.9%)	2 (14.3%)	16 (94.1%)	1 (6.7%)
Not <i>Gig</i>	0 (0%)	8 (88.9%)	0 (0%)	11 (84.6%)	2 (16.7%)	11 (91.7%)	1 (6.7%)	12 (100%)	1 (7.1%)	12 (85.7%)	1 (5.9%)	14 (93.3%)
Dev.	Actual		Actual		Actual		Actual		Actual		Actual	
<i>Giganteus</i>	10 (83.3%)	0 (0%)	13 (92.9%)	1 (7.7%)	9 (75%)	0 (0%)	8 (53.3%)	0 (0%)	11 (78.6%)	0 (0%)	16 (94.1%)	0 (0%)
Not <i>Gig</i>	0 (0%)	8 (88.9%)	0 (0%)	10 (76.9%)	0 (0%)	8 (66.7%)	0 (0%)	11 (91.7%)	0 (0%)	11 (78.6%)	0 (0%)	8 (53.3%)
None	2 (16.7%)	1 (11.1%)	1 (7.1%)	2 (15.4%)	3 (25%)	4 (33.3%)	7 (46.7%)	1 (8.3%)	3 (21.4%)	3 (21.4%)	1 (5.9%)	7 (46.7%)

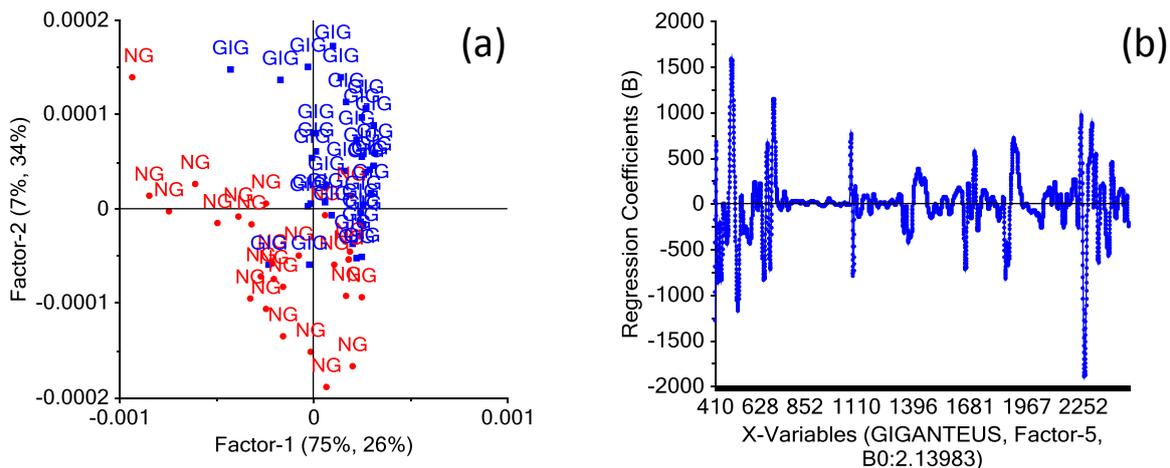


Figure E-12: Plots for the DF (giganteus/not-giganteus discriminatory-) model in Table E-18. (a) A F1 vs. F2 scores plot, GIG = giganteus, (NG) = not-giganteus; (b) a regression coefficients plot.

Table E-19: Classification matrices, according to giganteus or non-giganteus (Not Gig) variety, for the prediction and validation sets for LDA/QDA on the WC/WU dataset. There were 95 samples in the calibration set and 32 in the validation set.

Pre.	SG		SG									
Spec.	2,2,25,25		2,2,25,25		2,2,25,25		2,2,25,25		2,2,25,25		2,2,25,25	
λ (nm)	400-2500		400-2500		400-2500		400-2500		400-2500		400-2500	
LDA/QDA	LDA		QDA		LDA		QDA		LDA on PLS-DA		QDA on PLS-DA	
PCs Used	10		10		20		20		0		0	
Calibration	Actual		Actual									
Predict	Gig	Not Gig	Gig	Not Gig								
Gig	57 (100%)	2 (5.3%)	51 (89.5%)	0	55 (96.5%)	3 (7.9%)	56 (98.2%)	0	56 (98.2%)	1 (2.6%)	57 (100%)	0
Not Gig	0	36 (94.7%)	6 (10.5%)	38 (100%)	2 (3.5%)	35 (92.1%)	1 (1.8%)	38 (100%)	1 (1.8%)	37 (97.4%)	0	38 (100%)
Validation	Actual		Actual									
Gig	14 (82.4%)	1 (6.7%)	16 (94.1%)	1 (6.7%)	17 (100%)	1 (6.7%)	17 (100%)	1 (6.7%)	16 (94.1%)	1 (6.7%)	17 (100%)	0
Not Gig	3 (17.6%)	14 (93.3%)	1 (5.9%)	14 (93.3%)	0	14 (93.3%)	0	14 (93.3%)	1 (5.9%)	14 (93.3%)	0	15 (100%)

Table E-20: Validation results for the best PLS-DA models to discriminate between *M. giganteus* stem and leaf samples.

St:Lf = (number of stem samples):(number of leaf samples)

Dataset	DF		DS		DT		DG		DH		DU		DV		WC/WU	
Varieties	<i>Giganteus</i>		<i>Giganteus</i>		<i>Giganteus</i>		<i>Giganteus</i>		<i>Giganteus</i>		<i>Giganteus</i>		<i>Giganteus</i>		<i>Giganteus</i>	
Pretreat.	SG		SG		SG		SG		SG		SG		SG		SG	
Specific	2,2,20,20		2,2,20,20		2,2,20,20		2,2,20,20		2,2,20,20		2,2,20,20		2,2,20,20		2,2,20,20	
λ (nm)	400-2500		400-2500		400-2500		400-2500		400-2500		400-2500		400-2500		400-2500	
St:Lf	102:61		118:72		109:65		108:46		20:27		112:48		10:31		349:193	
Cal:Val	122:41		143:47		130:44		114:40		35:12		120:40		30:11		406:136	
Factors	5		6		7		6		3		6		5		10	
Score	Actual		Actual		Actual		Actual		Actual		Actual		Actual		Actual	
Predict	Leaf	Stem	Leaf	Stem	Leaf	Stem	Leaf	Stem	Leaf	Stem	Leaf	Stem	Leaf	Stem	Leaf	Stem
Leaf	20 (100%)	0	19 (100%)	0	16 (100%)	0	13 (100%)	0	7 (100%)	0	12 (100%)	0	9 (100%)	0	50 (100%)	1 (1.2%)
Stem	0	21 (100%)	0	28 (100%)	0	28 (100%)	0	27 (100%)	0	5 (100%)	0	27 (100%)	0	2 (100%)	0 (0%)	85 (98.8%)
Dev.	Actual		Actual		Actual		Actual		Actual		Actual		Actual		Actual	
Leaf	20 (100%)	0	18 (94.7%)	0	16 (100%)	0	13 (100%)	0	7 (100%)	0	12 (100%)	0	9 (100%)	0	48 (96%)	0 (0%)
Stem	0	20 (95.2%)	0	28 (100%)	0	28 (100%)	0	27 (100%)	0	4 (80%)	0	27 (100%)	0	2 (100%)	0 (0%)	85 (98.8%)
None	0	1 (4.8%)	1 (5.3%)	0	0	0	0	0	0	1 (20%)	0	0	0	0	2 (4%)	1 (1.2%)

Table E-21: Validation results, regarding the DF, DS, and DT datasets, for the best PLS-DA models to discriminate between the following fractions: dead leaf blades (Dead L.), dead sheaths (Dead S.), internodes (Intern.), live leaf blades (Live L.), and nodes.

Dataset	DF					DS					DT				
Varieties	<i>Giganteus</i>					<i>Giganteus</i>					<i>Giganteus</i>				
Pretreat.	SG-2,2,20,20					SG-2,2,20,20					SG-2,2,20,20				
λ (nm)	400-2500					400-2500					400-2500				
Cal:Val	110:38					129:44					121:41				
Factors	12					12					10				
Score	Actual					Actual					Actual				
<u>Predict</u>	Dead L.	Dead S.	Intern.	Live L.	Node	Dead L.	Dead S.	Intern.	Live L.	Node	Dead L.	Dead S.	Intern.	Live L.	Node
Dead L.	2 (100%)	1 (14.3%)	0	0	0	8 (100%)	1 (11.1%)	0	0	0	8 (88.9%)	0	0	2 (33.3%)	0
Dead S.	0	6 (85.7%)	0	0	0	0	8 (88.9%)	0	0	0	1 (11.1%)	4 (100%)	0	0	0
Intern.	0	0	17 (100%)	0	0	0	0	15 (100%)	0	0	0	0	14 (100%)	0 (0%)	1 (11.1%)
Live L.	0	0	0	1 (100%)	0	0	0	0	6 (100%)	0	0	0	0	4 (66.7%)	0
Node	0	0	0	0	11 (100%)	0	0	0	0	6 (100%)	0	0	0	0	8 (88.9%)
Deviation	Dead L.	Dead S.	Intern.	Live L.	Node	Dead L.	Dead S.	Intern.	Live L.	Node	Dead L.	Dead S.	Intern.	Live L.	Node
Dead L.	1 (50%)	0	0	0	0	6 (75%)	0	0	0	0	5 (55.6%)	0	0	2 (33.3%)	0
Dead S.	0	5 (71.4%)	0	0	0	0	7 (77.8%)	0	0	0	0	3 (75%)	0	0	0
Intern.	0	0	15 (88.2%)	0	0	0	0	13 (86.7%)	0	0	0	0	14 (100%)	0	1 (11.1%)
Live L.	0	0	0	1 (100%)	0	0	0	0	4 (66.7%)	0	0	0	0	2 (33.3%)	0
Node	0	0	0	0	10 (90.9%)	0	0	0	0	3 (50%)	0	0	0	0	8 (88.9%)
None	1 (50%)	2 (28.6%)	2 (11.8%)	0	1 (9.1%)	2 (25%)	2 (22.2%)	2 (13.3%)	2 (33.3%)	3 (50%)	4 (44.4%)	1 (25%)	0	2 (33.3%)	0

Table E-22: Validation results, regarding the DG, DU, and WC/WU datasets, for the best PLS-DA models to discriminate between the following fractions: dead leaf blades (Dead L.), dead sheaths (Dead S.), internodes (Intern.), live leaf blades (Live L.) and nodes.

Dataset	DG					DU					WC/WU				
Varieties	<i>Giganteus</i>					<i>Giganteus</i>					<i>Giganteus</i>				
Pretreat.	SG -2,2,20,20					SG -2,2,20,20					SG -2,2,20,20				
λ (nm)	400-2500					400-2500					400-2500				
Cal:Val	102:35					108:36					393:132				
Factors	15					11					16				
Score	Actual					Actual					Actual				
Predict	Dead L.	Dead S.	Intern.	Live L.	Node	Dead L.	Dead S.	Intern.	Live L.	Node	Dead L.	Dead S.	Intern.	Live L.	Node
Dead L.	6 (100%)	0	0	0	0	2 (100%)	0	0	0	0	15 (93.8%)	0	0	0	0
Dead S.	0	4 (100%)	0	0	0	0	4 (100%)	0	0	0	1 (6.3%)	22 (100%)	0	0	0
Intern.	0	0	15 (100%)	0	0	0	0	16 (100%)	0	0	0	0	39 (100%)	0	0
Live L.	0	0	0	1 (100%)	0	0	0	0	1 (100%)	0	0	0	0	6 (100%)	0
Node	0	0	0	0	8 (100%)	0	0	0	0	13 (100%)	0	0	0	0	49 (100%)
Deviation	Dead L.	Dead S.	Intern.	Live L.	Node	Dead L.	Dead S.	Intern.	Live L.	Node	Dead L.	Dead S.	Intern.	Live L.	Node
Dead L.	5 (83.3%)	0	0	0	0	2 (100%)	0	0	0	0	12 (75%)	0	0	0	0
Dead S.	0	4 (100%)	0	0	0	0 (0%)	3 (75%)	0	0	0	0	19 (86.4%)	0	0	0
Intern.	0	0	14 (93.3%)	0	0	0	0	16 (100%)	0	0	0	0	39 (100%)	0	0
Live L.	0	0	0	1 (100%)	0	0	0	0	0	0	0	0	0	5 (83.3%)	0
Node	0	0	0	0	8 (100%)	0	0	0	0	12 (92.3%)	0	0	0	0	49 (100%)
None	1 (16.7%)	0 (0%)	1 (6.7%)	0 (0%)	0 (0%)	0 (0%)	1 (25%)	0 (0%)	1 (100%)	1 (7.7%)	4 (25%)	3 (13.6%)	0 (0%)	1 (16.7%)	0 (0%)

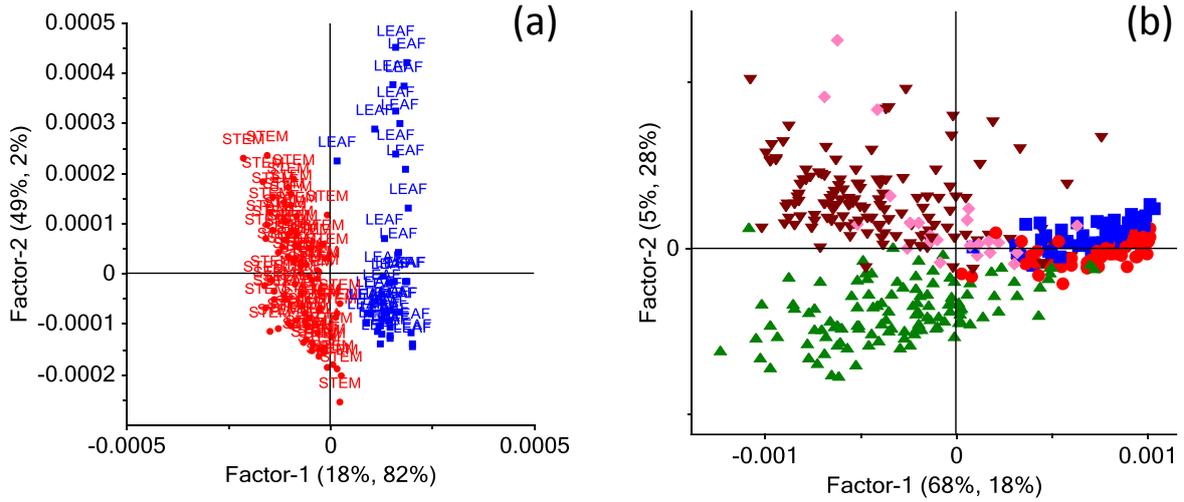


Figure E-13: Factor 1 vs. Factor 2 scores plots for plant fraction PLS-DAs. (a) PLS-DA between stem samples (red, left hand side) and leaf samples (blue, right hand side) for the DF dataset; and (b) PLS-DA between dead leaves (blue squares), dead sheaths (red circles), internodes (green triangles), nodes (brown inverted triangles), and live leaves (pink diamonds) for the WC/WU dataset.

Table E-23: LDA/QDA calibration and validation results for the discrimination, for the WC/WU datasets, between the following fractions: dead leaf blades (Dead L.), dead sheaths (Dead S.), internodes (Intern.), live leaf blades (Live L.) and nodes. In all cases the SG2,2,20,20 pretreatment and the wavelength region 400-2500 nm were used.

LDA/QDA	LDA					QDA				
Factors	20 PCs					20 PCs				
	Actual					Actual				
Calib. Set	Dead L.	Dead S.	Intern.	Live L.	Node	Dead L.	Dead S.	Intern.	Live L.	Node
Dead L.	55 (98.2%)	5 (8.1%)	0	2 (8.3%)	0	55 (98.2%)	0	0	0	0
Dead S.	0	57 (91.9%)	0	0	0	0	62 (100%)	0	0	0
Intern.	0	0	119 (100%)	0	0	0	0	119 (100%)	0	0
Live Lv.	1 (1.8%)	0	0	22 (91.7%)	0	1 (1.8%)	0	0	24 (100%)	0
Node	0	0	0	0	130 (100%)	0	0	0	0	130 (100%)
Valid. Set	Dead L.	Dead S.	Intern.	Live L.	Node	Dead L.	Dead S.	Intern.	Live L.	Node
Dead L.	15 (100%)	0	0	0	0	11 (73.3%)	0	0	0	0
Dead S.	0	22 (100%)	1 (2.5%)	0	0	2 (13.3%)	21 (95.5%)	0	0	0
Intern.	0	0	39 (97.5%)	0	0	0	0	39 (97.5%)	0	0
Live Lv.	0	0	0	6 (100%)	0	2 (13.3%)	0	0	6 (100%)	0
Node	0	0	0	0	49 (100%)	0	1 (4.5%)	1 (2.5%)	0	49 (100%)
LDA/QDA	LDA on PLS-DA Scores					QDA on PLS-DA Scores				
Factors	16 (in PLS-DA)					16 (in PLS-DA)				
	Actual					Actual				
Calib. Set	Dead L.	Dead S.	Intern.	Live L.	Node	Dead L.	Dead S.	Intern.	Live L.	Node
Dead L.	56 (100%)	5 (8.1%)	0	0	0	56 (100%)	0	0	0	0
Dead S.	0	57 (91.9%)	0	2 (8.3%)	0	0	62 (100%)	0	0	0
Intern.	0	0	119 (100%)	0	1 (0.8%)	0	0	119 (100%)	0	0
Live Lv.	0	0	0	22 (91.7%)	0	0	0	0	24 (100%)	0
Node	0	0	0	0	129 (99.2%)	0	0	0	0	130 (100%)
Valid. Set	Dead L.	Dead S.	Intern.	Live L.	Node	Dead L.	Dead S.	Intern.	Live L.	Node
Dead L.	15 (100%)	0	0	0	0	12 (80%)	0	0	0	0
Dead S.	0	22 (100%)	1 (2.5%)	0	0	1 (6.7%)	21 (95.5%)	0	0	0
Intern.	0	0	39 (97.5%)	0	0	0	0	40 (100%)	0	0
Live Lv.	0	0	0	6 (100%)	0	2 (13.3%)	0	0	6 (100%)	0
Node	0	0	0	0	49 (100%)	0	1 (4.5%)	0	0	49 (100%)

Table E-24: Classification table, for plant fraction discrimination, for the SIMCA of the WC/WU validation set. The number of PCs used in the models of each fraction are provided in brackets.

Fraction (#PCs)	Dead Leaves (5)	Dead Sheaths (5)	Internodes (6)	Live Leaves (4)	Nodes (8)	Number Wrongly Identified As
Actual #	15	22	40	6	49	
Identified						
Dead Leaves	15	1	0	0	0	1
Dead Sheaths	6	22	0	0	0	6
Internodes	4	18	39	0	18	40
Live Leaves	9	7	1	6	0	17
Nodes	10	12	14	0	49	36
Sensitivity (%)	100	100	97.5	100	100	
Specificity (%)	99.1	94.5	56.5	86.5	56.6	

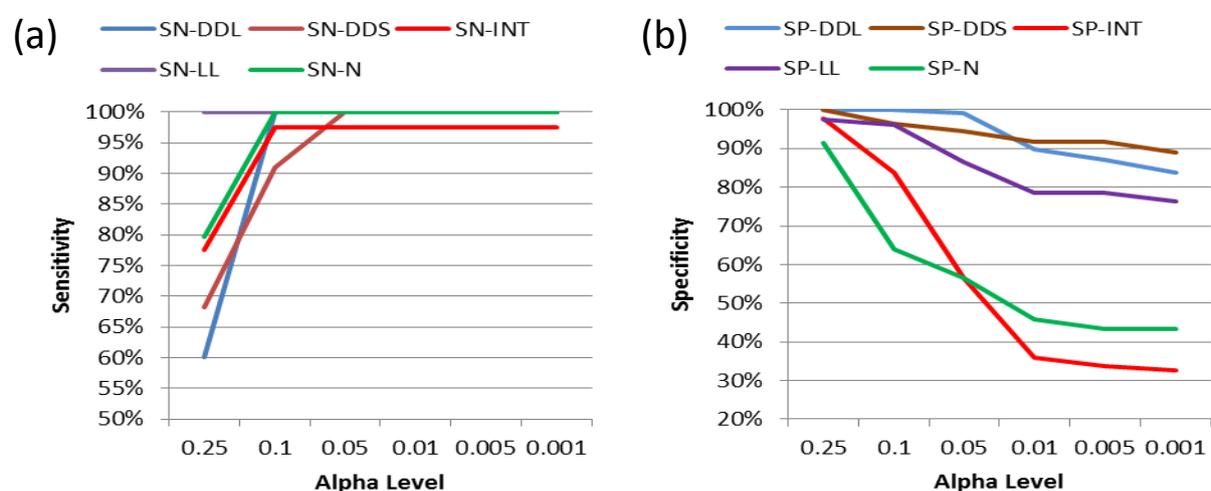


Figure E-14: SIMCA plots. (a) Sensitivity and (b) specificity values for the five different plant fractions under SIMCA classification of the WU/WC validation set. Alpha levels range from 0.25 (25%) to 0.001.

Table E-25: Results from the classification of an independent validation set using PLS-DA models designed to discriminate between plant samples that came from a plantation in its first year of growth (1-Year) against samples from older plantations (Older).

Dataset	WU/WC		WU/WC Leaves		DT	
Pretreatment	SG		None		SG	
Specific	2,2,20,20				2,2,20,20	
λ (nm)	400-2500		400-2500		400-2500	
1-Year:Older	83:312		33:112		31:92	
Cal:Val	295:100		107:38		88:35	
Factors	20		15		5	
Score	Actual		Actual		Actual	
<u>Predict</u>	1-Year	Older	1-Year	Older	1-Year	Older
1-Year	14 (77.8%)	0	8 (100%)	2 (6.7%)	8 (88.9%)	2 (7.7%)
Older	4 (22.2%)	82 (100%)	0 (0%)	28 (93.3%)	1 (11.1%)	24 (92.3%)
Dev.	Actual		Actual		Actual	
1-Year	3 (16.7%)	0	3 (37.5%)	0 (0%)	6 (66.7%)	0 (0%)
Older	0	61 (74.4%)	0 (0%)	22 (73.3%)	0 (0%)	14 (53.8%)
None	15	21 (25.6%)	5 (62.5%)	8 (26.7%)	3 (33.3%)	12 (46.2%)

Appendix F Figures and Tables for Chapter 15: Development of NIRS Quantitative Calibrations for Miscanthus Samples

Table F-1: Summary of the number of samples, for each dataset, with good analytical data for each of the constituents used in NIRS model development. The samples are classified according to whether they are of the giganteus (G), sinensis (S), or Other (O) variety. The “Sugars” category represents arabinose, galactose, rhamnose, glucose, xylose, and total sugars. The MAN_SRS constituent is listed separately because sometimes the concentration was so low that it was not detected by the HPAEC in a sample where all the other sugars were detected. KL = Klason lignin; ASL = acid soluble lignin; AIR = acid insoluble residue; AIA = acid insoluble ash; UA = uronic acids; Elemental = carbon, hydrogen, nitrogen, and sulphur.

Constit	DS			DT			DG			DH			DU			DV			WU			DF		
	G	S	O	G	S	O	G	S	O	G	S	O	G	S	O	G	S	O	G	S	O	G	S	O
Extractives	183	13	50	173	13	44	147	12	50	43	0	16	148	13	49	37	0	0	189	13	33	51	5	3
Ash	170	13	48	162	13	43	134	12	48	44	0	16	135	13	47	38	0	0	177	13	31	51	5	3
KL	162	10	25	152	10	22	132	10	25	35	0	8	132	10	25	31	0	0	167	10	16	42	5	3
ASL	172	12	35	160	12	30	140	12	35	35	0	13	140	12	35	31	0	0	175	12	21	44	5	2
AIR	169	10	36	159	10	31	139	10	36	36	0	14	139	10	36	31	0	0	174	10	21	42	5	3
AIA	168	11	34	157	11	29	138	11	34	35	0	14	137	11	34	31	0	0	172	11	19	42	5	3
SUGARS	154	11	37	147	11	33	125	11	37	35	0	14	126	11	37	30	0	0	159	11	22	44	5	3
MAN_SRS	147	9	36	140	9	33	117	9	36	35	0	14	119	9	36	30	0	0	152	9	21	43	5	3
UA	32	1	0	25	1	0	32	0	0	0	0	0	32	1	0	0	0	0	32	1	0	0	0	0
Elemental	46	4	6	37	4	5	47	3	6	0	0	0	45	4	6	0	0	0	45	4	6	45	4	2

Table F-2: Histograms, using % whole dry mass data, with associated statistics, for total sugars (TOT_SRS) and glucose (GLU_SRS) contents of Miscanthus sets comprising: DS data for samples of all Miscanthus varieties (DS ALL); DS data for Miscanthus x giganteus samples (DS GIG); DF data for samples of all Miscanthus varieties (DF ALL).

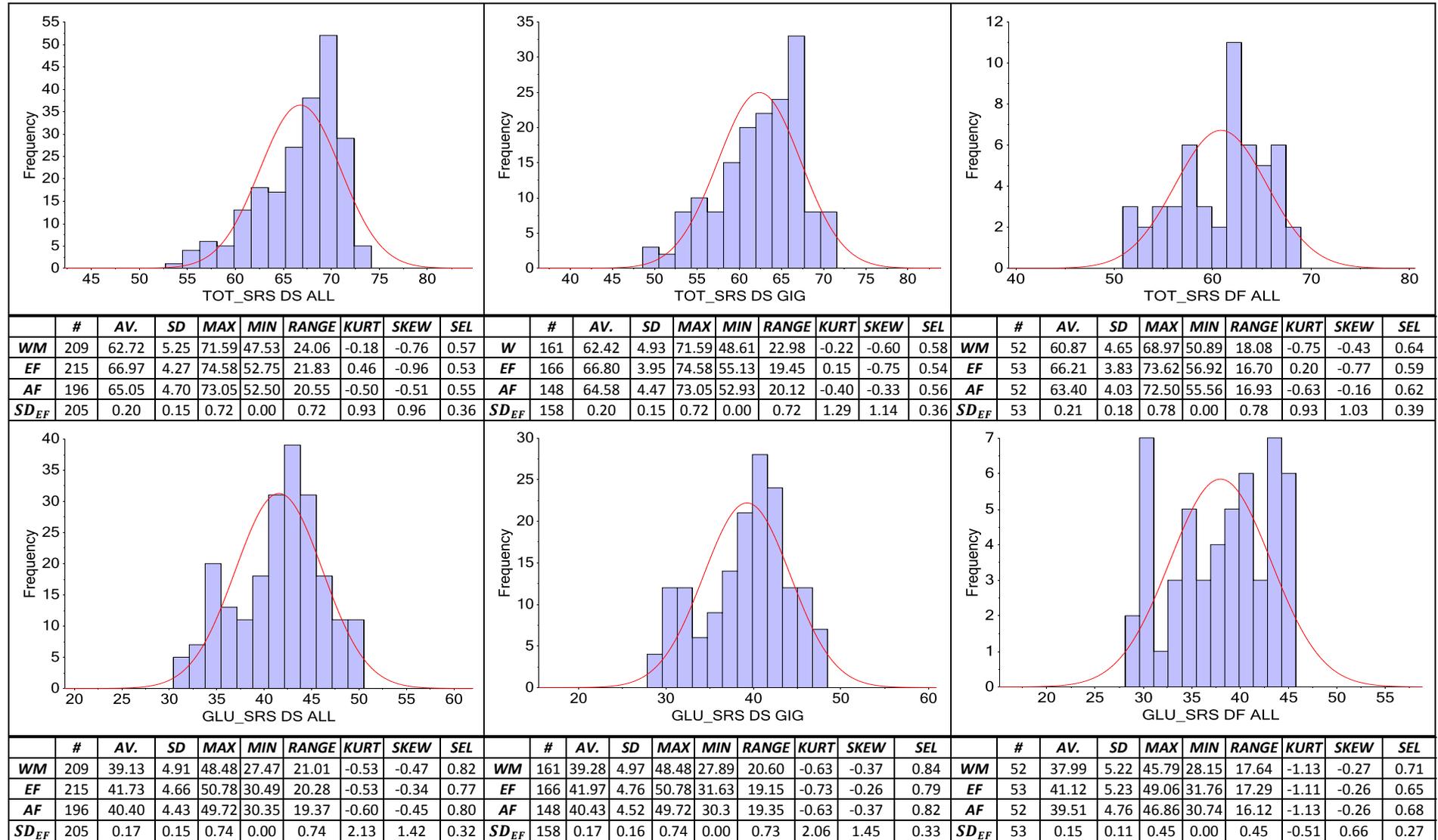


Table F-3: Histograms, with associated statistics and using % whole dry mass data, for the xylose (XYL_SRS) and the mannose (MAN_SRS) contents of sets comprising: DS data for samples of all Miscanthus varieties (DS ALL); DS data for Miscanthus x giganteus samples (DS GIG); DF data for samples of all Miscanthus varieties (DF ALL).

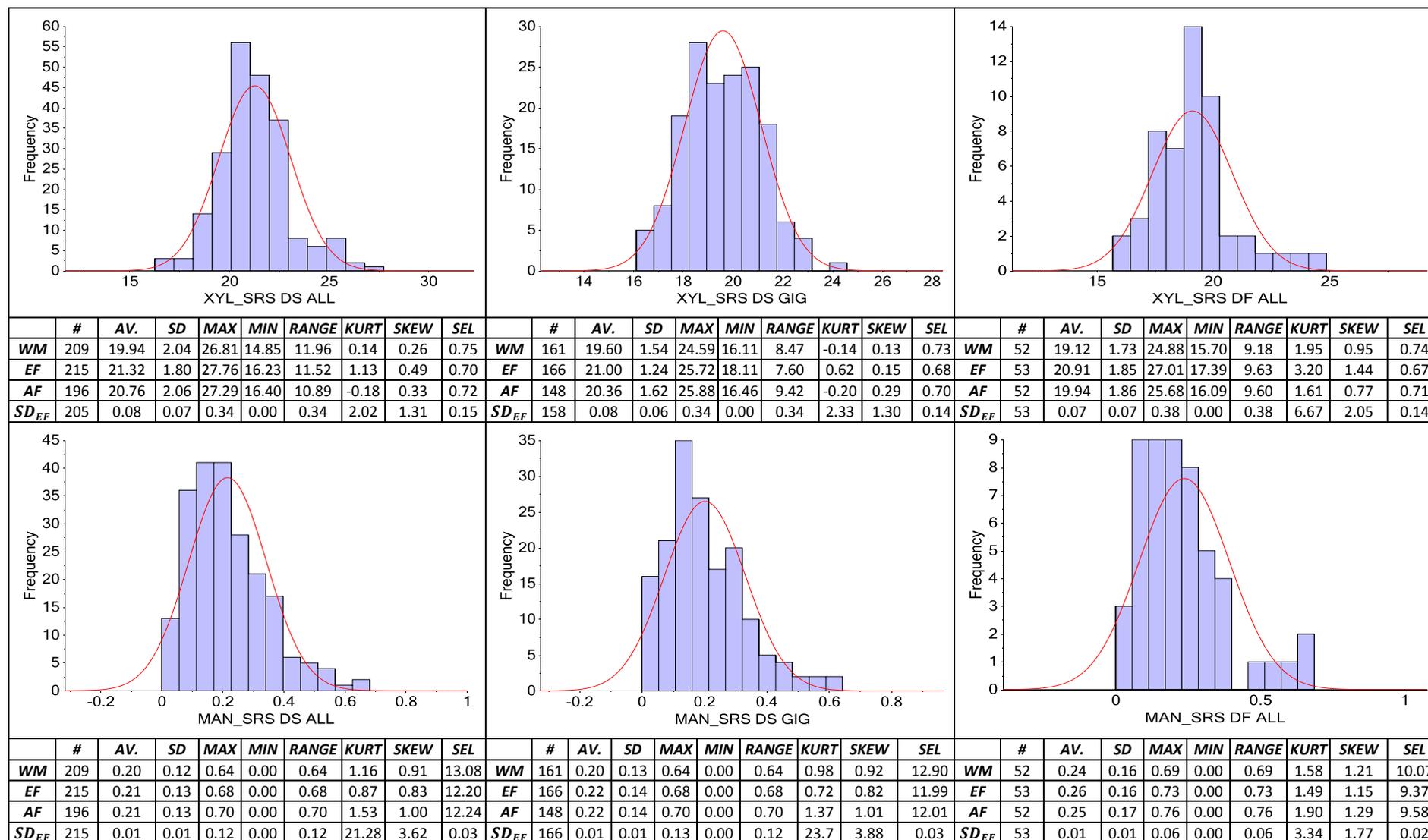


Table F-4: Histograms, with associated statistics and using % whole dry mass data, for the arabinose (ARA_SRS) and the galactose (GAL_SRS) contents of sets comprising: DS data for samples of all Miscanthus varieties (DS ALL); DS data for Miscanthus x giganteus samples (DS GIG); DF data for samples of all Miscanthus varieties (DF ALL).

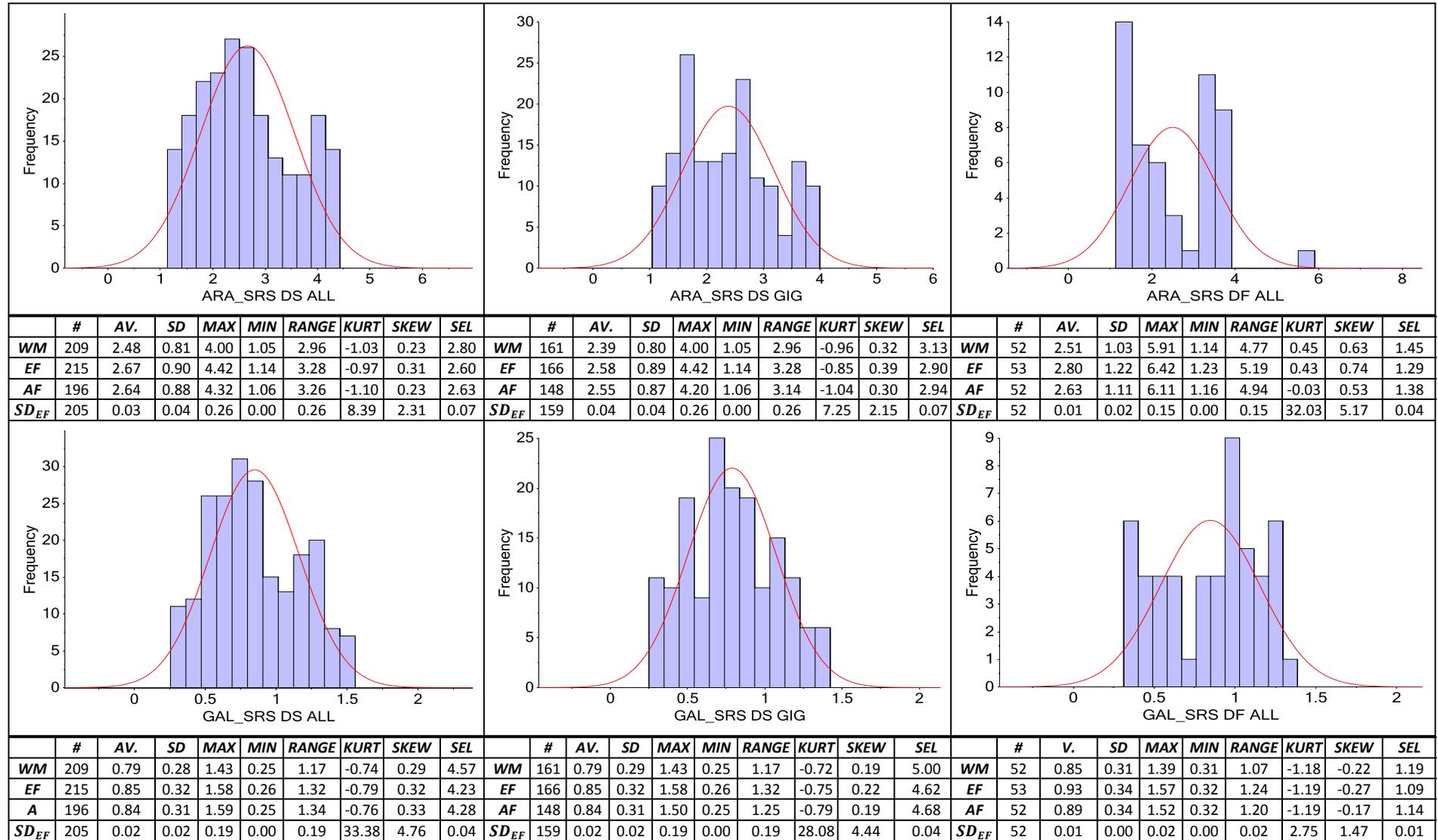


Table F-5: Histograms, with associated statistics and using % whole mass data, for the Klason lignin content (KL), and acid soluble lignin (ASL) content of sets comprising: DS data for samples of all Miscanthus varieties (DS ALL); DS data for Miscanthus x giganteus samples (DS GIG); DF data for samples of all Miscanthus varieties (DF ALL).

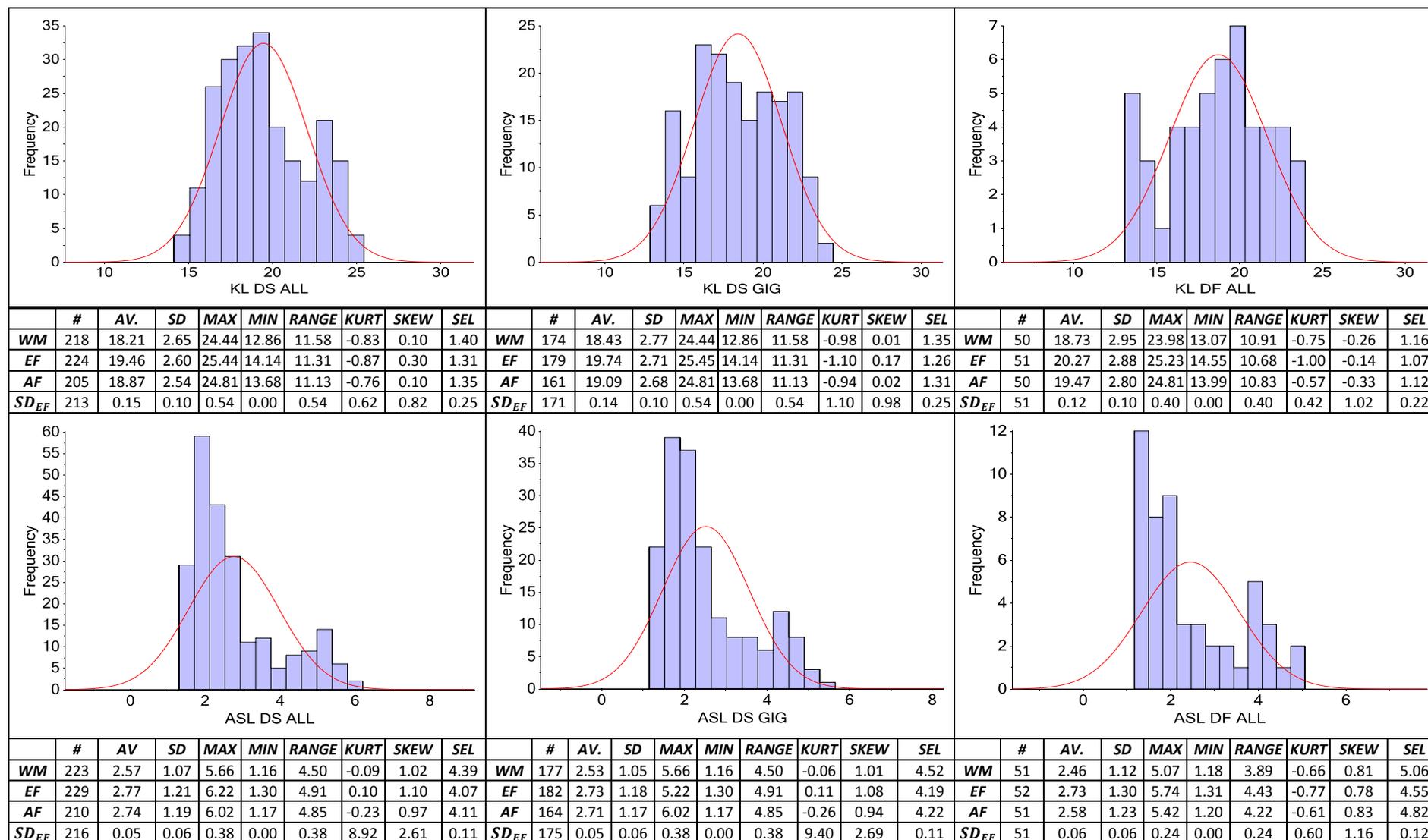


Table F-6: Histograms, with associated statistics and using % whole dry mass data, for the rhamnose (RHA_SRS) and the uronic acids (UA) contents of sets comprising: DS data for samples of all Miscanthus varieties (DS ALL); DS data for Miscanthus x giganteus samples (DS GIG); DF data for samples of all Miscanthus varieties (DF ALL).

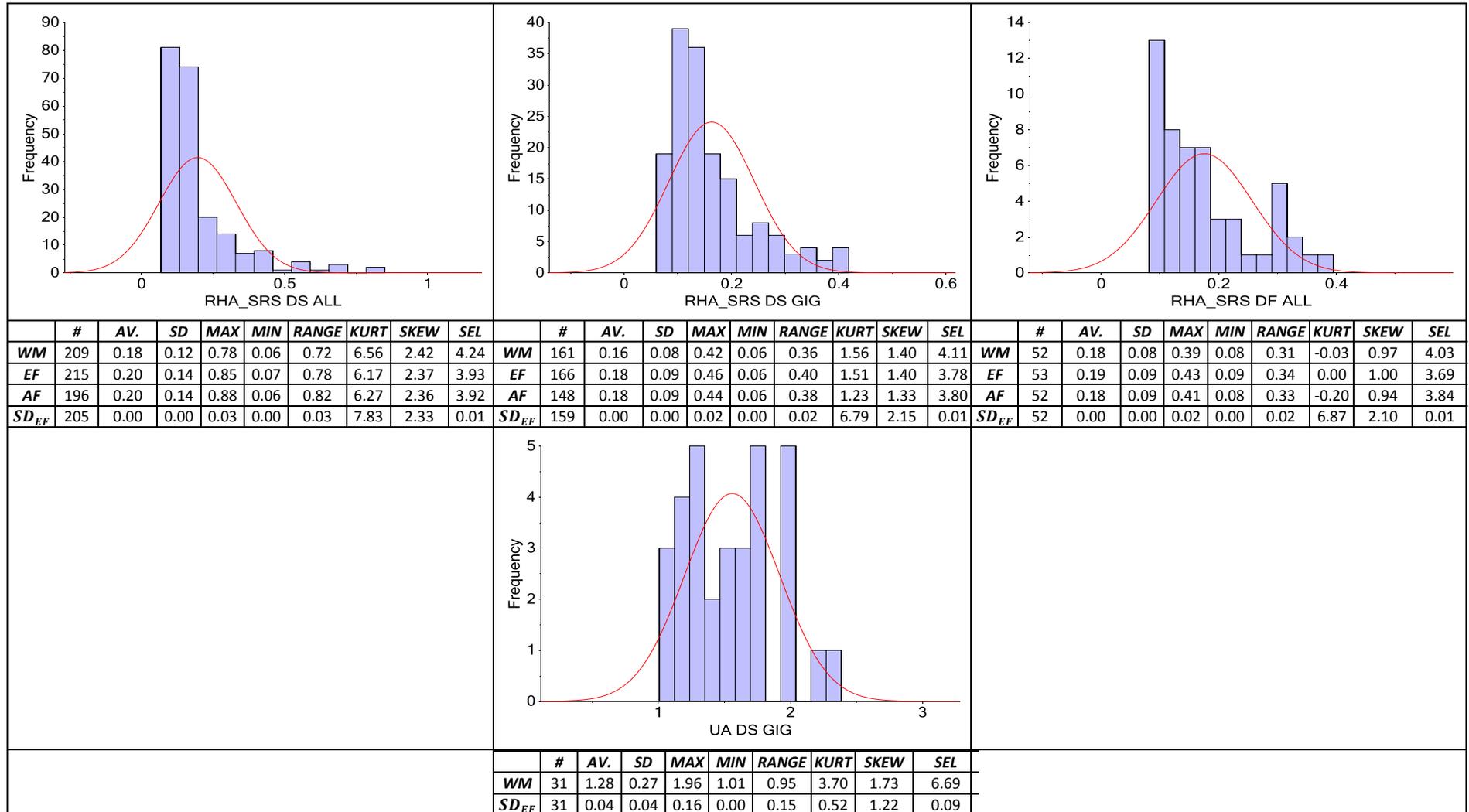


Table F-7: Histograms, with associated statistics and using % whole dry mass data, for the carbon (C) and the hydrogen (H) contents of sets comprising: DS data for samples of all Miscanthus varieties (DS ALL); DS data for Miscanthus x giganteus samples (DS GIG); DF data for samples of all Miscanthus varieties (DF ALL).

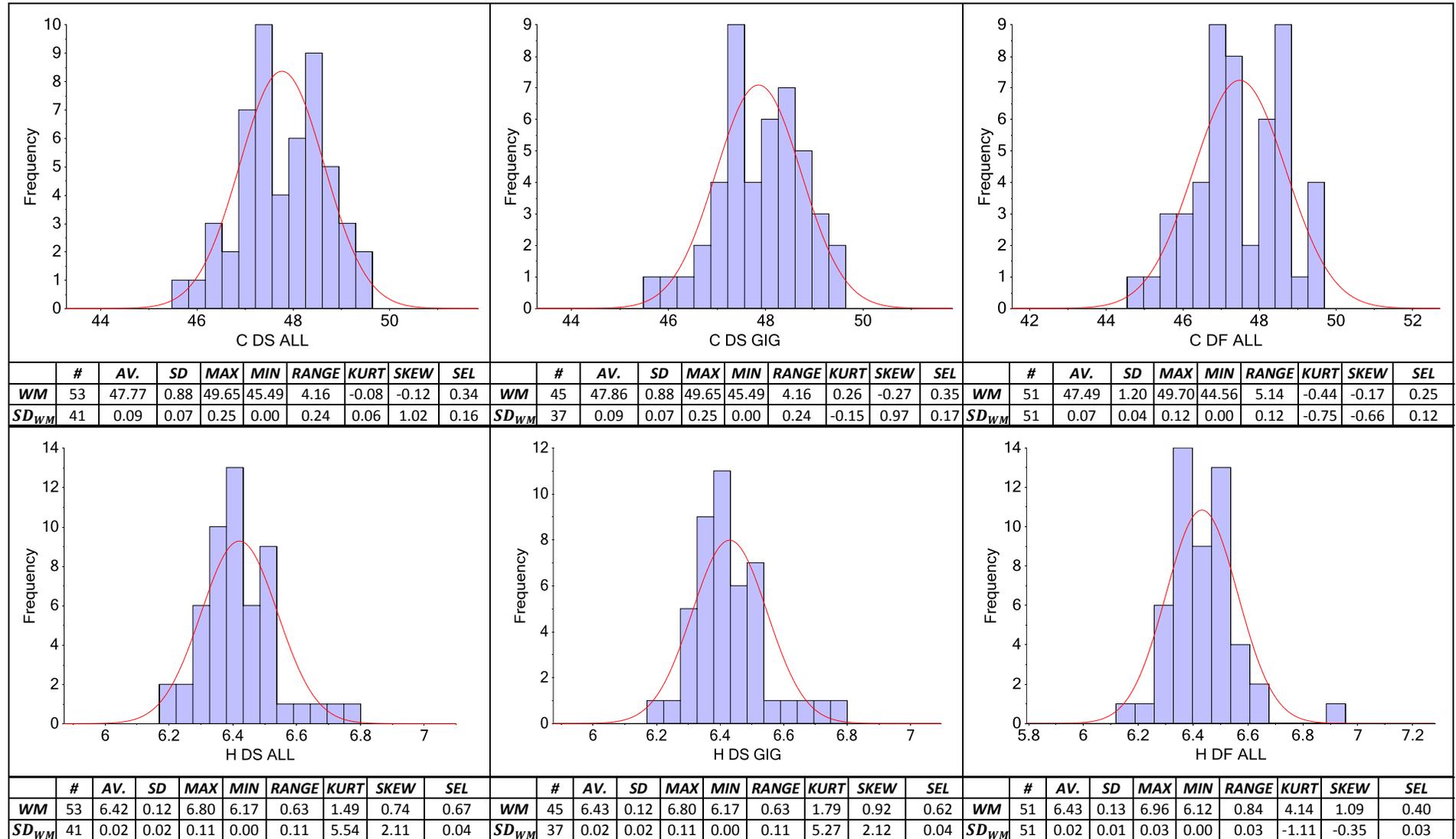


Table F-8: Histograms, with associated statistics and using % whole dry mass data, for the nitrogen (N) and the sulphur (S) contents of sets comprising: DS data for samples of all Miscanthus varieties (DS ALL); DS data for Miscanthus x giganteus samples (DS GIG); DF data for samples of all Miscanthus varieties (DF ALL).

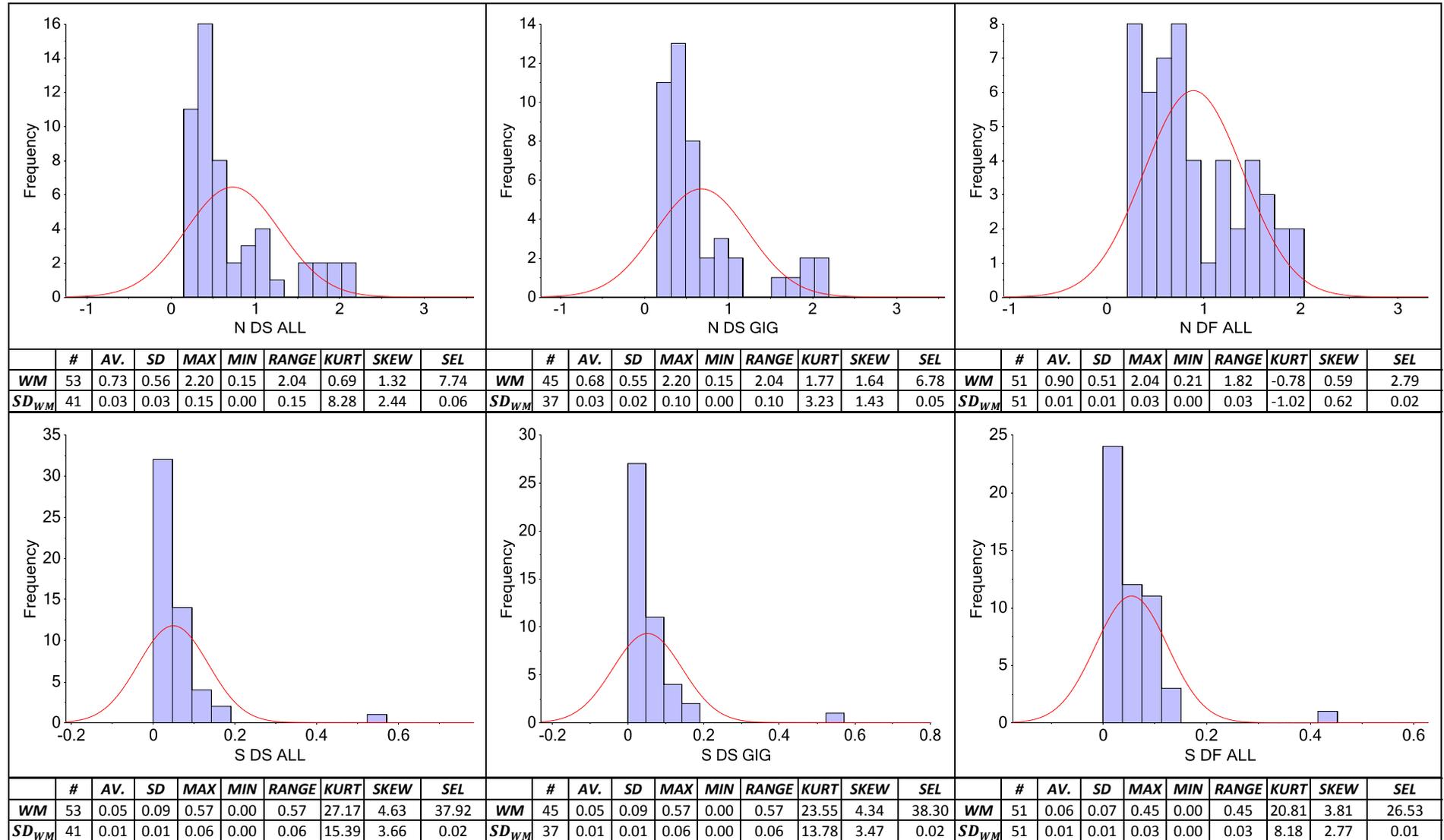


Table F-9: Histograms, with associated statistics and using whole mass data, for the 95% ethanol-soluble extractives content (EXTR_PD), and ash content of sets comprising: DS data for samples of all Miscanthus varieties (DS ALL); DS data for Miscanthus x giganteus samples (DS GIG); DF data for samples of all Miscanthus varieties (DF ALL).

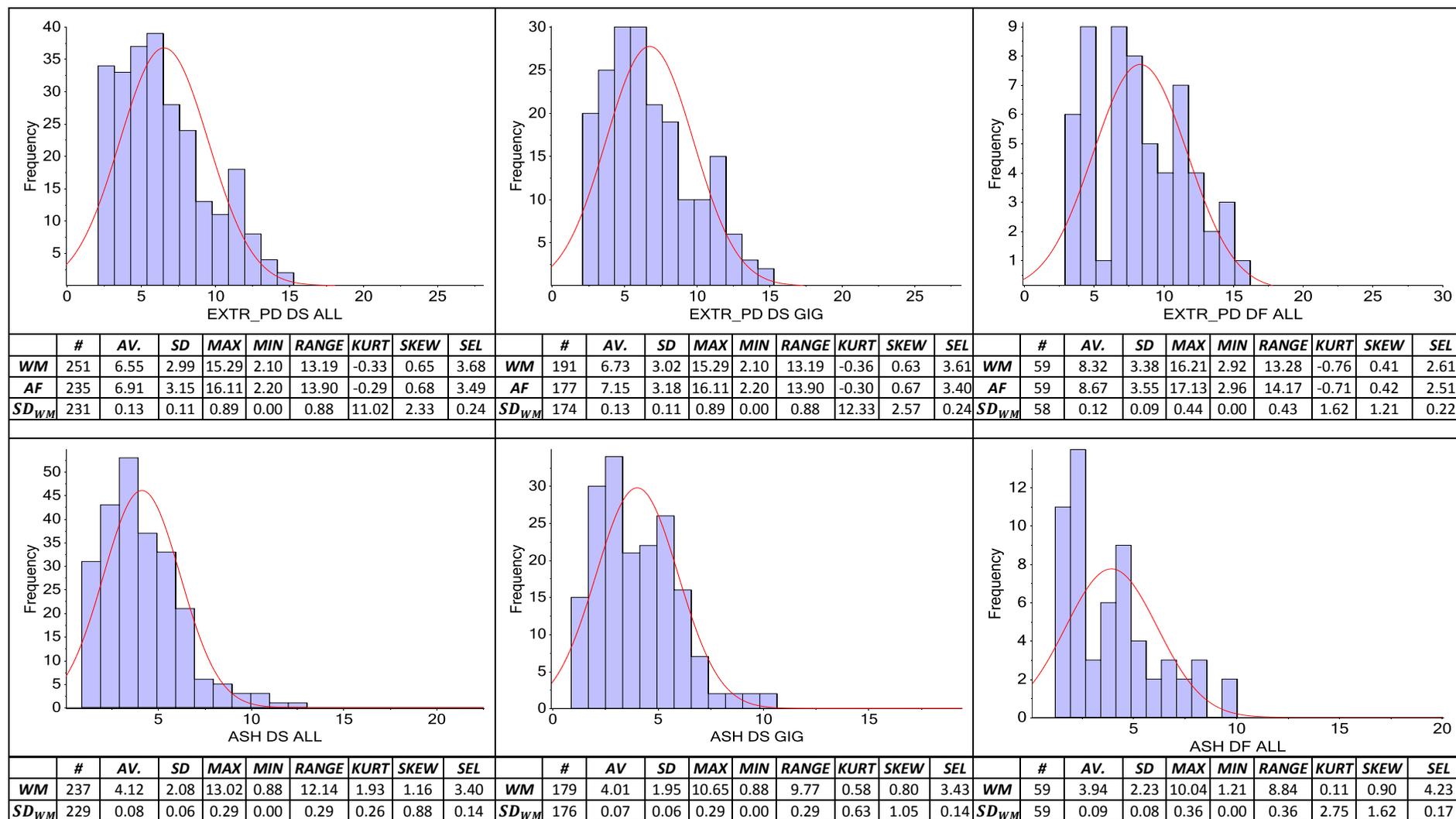
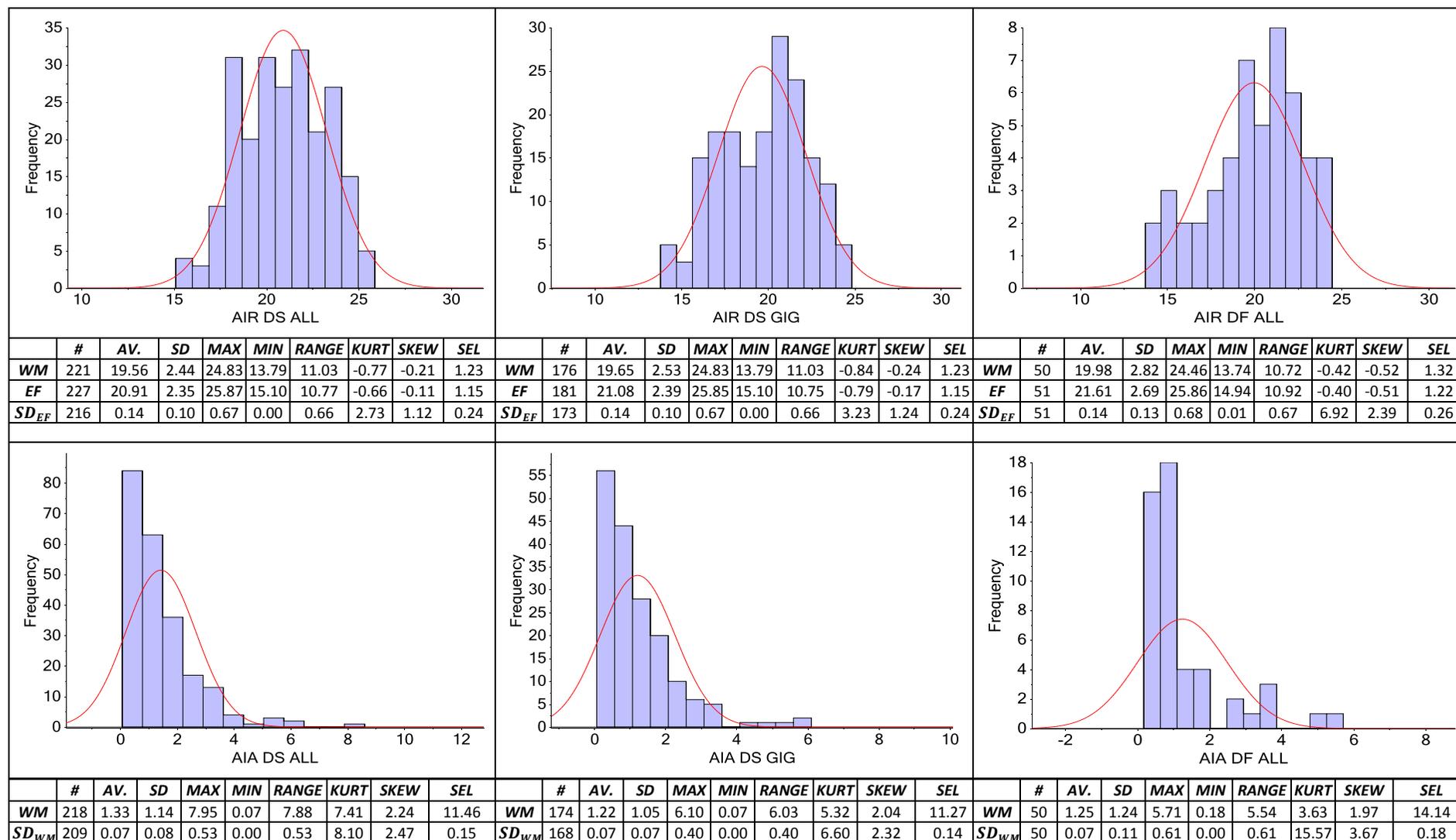


Table F-10: Histograms, with associated statistics and using whole mass data, for acid insoluble residue (AIR) and the acid insoluble ash (AIA) contents of sets comprising: DS data for samples of all Miscanthus varieties (DS ALL); DS data for Miscanthus x giganteus samples (DS GIG); DF data for samples of all Miscanthus varieties (DF ALL).



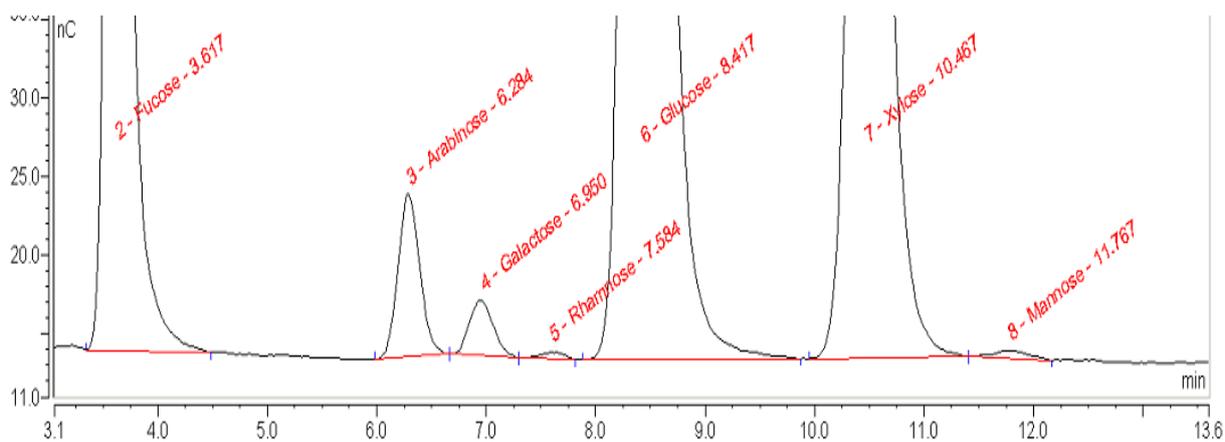


Figure F-1: A chromatogram of the hydrolysate of a *Miscanthus node* sample. The x-axis represents time in minutes and the y-axis charge in nC. The analytes are eluted in the following order: fucose (3.62 min), arabinose (6.28 min), galactose (6.95 min), rhamnase (7.58 min), glucose (8.42 min), xylose (10.47 min), mannose (11.77 min).

Table F-11: Summary data for the sugar recoveries of the 86 hydrolysis batches.

	Average Sugar Recoveries over all Batches (%)						Average Standard Deviation within a Batch (%)					
	Ara	Gal	Rha	Glu	Xyl	Man	Ara	Gal	Rha	Glu	Xyl	Man
Av	91.97	94.70	94.01	94.47	86.28	93.66	0.19	0.28	0.43	0.13	0.19	2.27
Max	93.47	96.08	96.83	95.75	87.66	102.11	0.91	1.56	3.09	0.46	0.61	10.91
Min	91.20	93.32	90.63	93.77	85.11	79.44	0.01	0.01	0.02	0.00	0.01	0.10
Range	2.27	2.76	6.21	1.97	2.56	22.67	0.90	1.55	3.07	0.46	0.60	10.81
SD	0.43	0.50	0.85	0.47	0.51	3.32	0.12	0.26	0.51	0.09	0.10	2.78

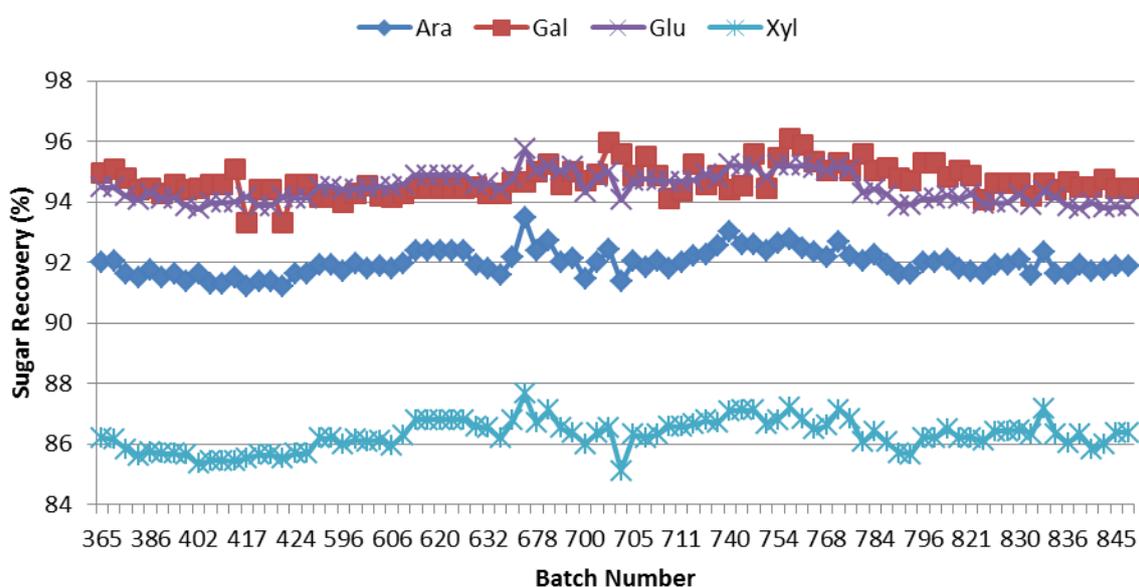


Figure F-2: Chart plotting the sugar recovery rates over the 86 batches.

Table F-12: Correlation table for selected components of the *Miscanthus x giganteus* DS samples. Absolute values greater than 0.5 are highlighted in bold.

	ASH	UA	ASA	KL	ASL	AIR	AIA	EIA	ARA_SRS	GAL_SRS	RHA_SRS	GLU_SRS	XYL_SRS	MAN_SRS	TOT_SRS	C	N
EXTR_PD	0.217	-0.118	0.387	-0.525	0.492	-0.573	-0.022	0.237	0.234	0.260	0.316	-0.586	-0.730	-0.156	-0.765	0.055	0.642
ASH		0.265	0.880	-0.726	0.771	-0.480	0.700	0.937	0.727	0.686	0.689	-0.722	-0.112	0.426	-0.575	-0.711	0.581
UA			0.337	-0.198	0.178	-0.264	-0.119	0.138	0.483	0.595	0.140	-0.346	0.556	0.487	-0.057	-0.400	-0.155
ASA				-0.739	0.884	-0.637	0.413	0.843	0.770	0.783	0.730	-0.842	-0.171	0.478	-0.709	-0.419	0.826
KL					-0.793	0.926	-0.404	-0.676	-0.707	-0.603	-0.611	0.648	0.127	-0.119	0.533	0.695	-0.613
ASL						-0.665	0.463	0.794	0.850	0.783	0.825	-0.905	-0.326	0.406	-0.808	-0.316	0.920
AIR							-0.026	-0.369	-0.513	-0.463	-0.419	0.505	0.094	0.052	0.422	0.505	-0.626
AIA								0.839	0.608	0.443	0.575	-0.447	-0.069	0.425	-0.329	-0.570	0.062
EIA									0.789	0.711	0.754	-0.756	-0.181	0.506	-0.623	-0.629	0.581
ARA_SRS										0.847	0.831	-0.819	0.001	0.553	-0.585	-0.574	0.547
GAL_SRS											0.708	-0.815	-0.089	0.709	-0.624	-0.368	0.552
RHA_SRS												-0.790	-0.170	0.422	-0.646	-0.319	0.653
GLU_SRS													0.389	-0.491	0.923	0.241	-0.855
XYL_SRS														0.130	0.700	-0.314	-0.467
MAN_SRS															-0.290	-0.038	0.096
TOT_SRS																0.033	-0.865
Carbon (C)																	0.171

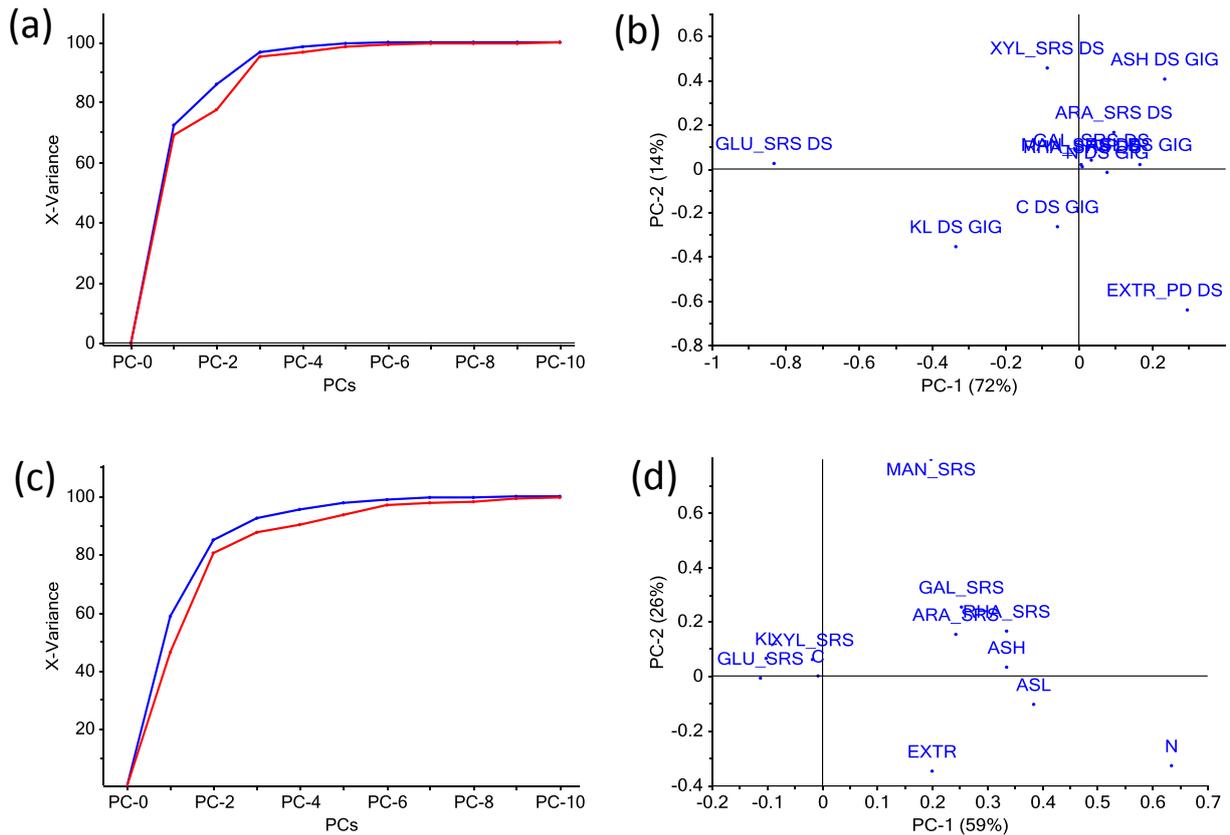


Figure F-3: Plots for a PCA based on the chemical data of *Miscanthus DS* samples. (a) Explained variance plot (blue line = in calibration, red line = in full cross validation) for the PCA model based on the original constituent data; (b) PC1 vs. PC2 loadings plot for this model; (c) and (d) – same as (a) and (b) but for the model based on the mean-normalised constituent data.

Table F-13: Statistics for the PCR for a range of components where the predictive variables are the other constituents in this table.

Component	# PCs	R^2_{Cal}	RMSEC (%)	Offset	R^2_{CV}	RMSECV (%)	Slope	Offset
EXTR_PD	10	0.9215	0.7369	0.5048	0.41	1.2295	5.7860	0.6437
ASH	11	0.8647	0.6940	0.5520	0.6414	1.2186	3.0358	1.1015
KL	10	0.9514	0.6037	0.9111	0.8564	0.9974	17.7311	0.9974
ASL	2	0.9561	2.1644	0.1061	0.9571	0.2390	2.2841	0.1276
ARA_SRS	3	0.8114	0.3354	0.4470	0.7896	0.3780	1.8646	0.5072
GAL_SRS	2	0.8296	0.1162	0.1361	0.8239	0.1277	0.6509	0.1491
RHA_SRS	7	0.8613	0.0231	0.0203	0.7622	0.0318	0.1121	0.0340
GLU_SRS	1	0.8797	1.7572	4.7825	0.8783	1.8923	31.7401	8.1259
XYL_SRS	3	0.6605	0.8511	6.6576	0.5024	1.0178	11.6445	8.0068
MAN_SRS	6	0.7730	0.0653	0.0490	0.6941	0.0840	0.1530	0.0636
Carbon (C)	9	0.8679	0.3542	6.3271	0.6897	0.5791	34.8758	13.0228
Nitrogen (N)	7	0.9318	0.1328	0.0465	0.8582	0.1849	0.6102	0.0778

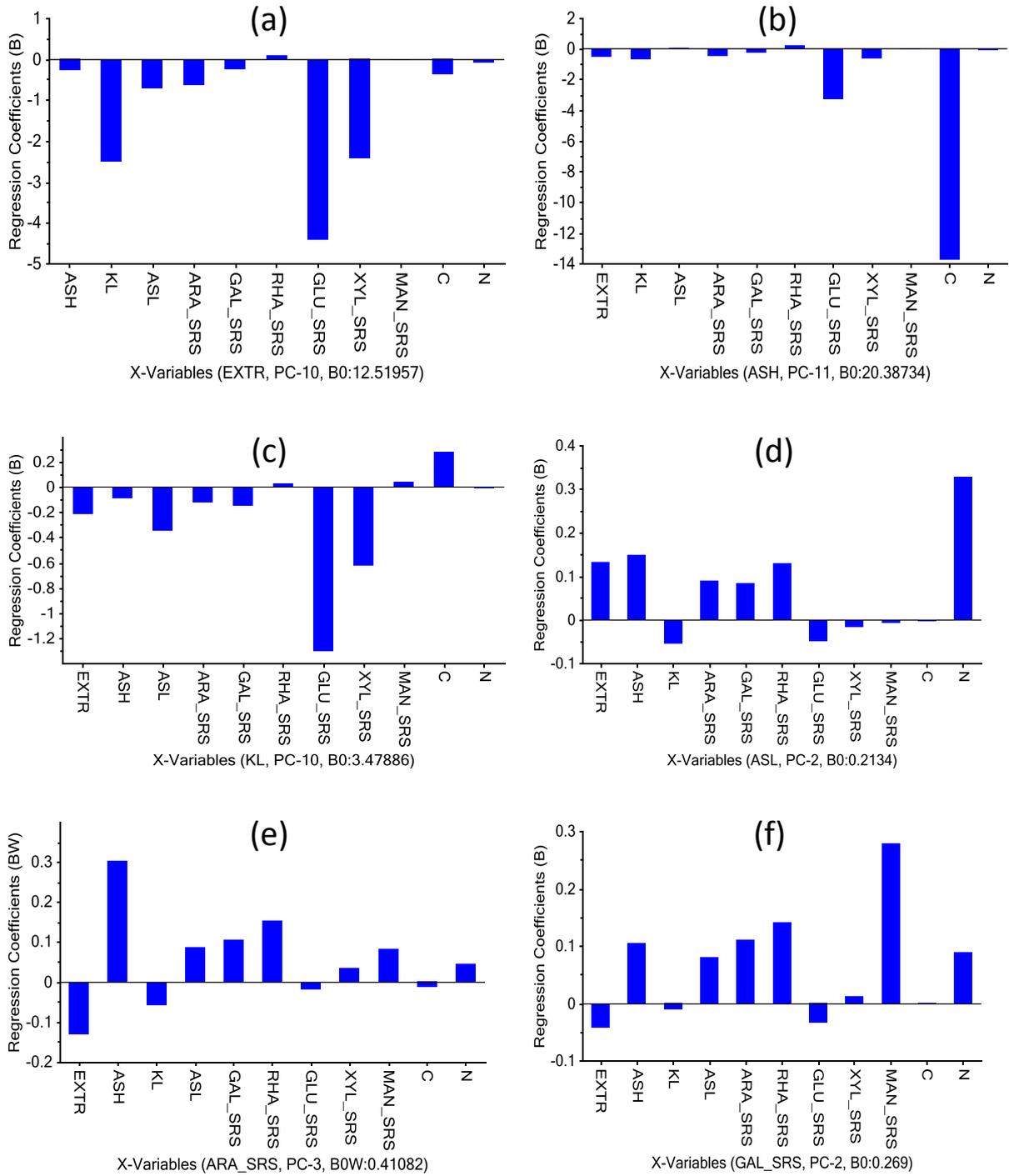


Figure F-4: Regression coefficients plots for PCR models using Miscanthus chemical data. (a) Extractives content PCR, using 10 PCs; (b) ash content PCR, using 11 PCs; (c) KL content PCR, using 10 PCs; (d) ASL content PCR, using 2 PCs; (e) ARA_SRS content PCR, using 3 PCs; (f) GAL_SRS content PCR, using 2 PCs.

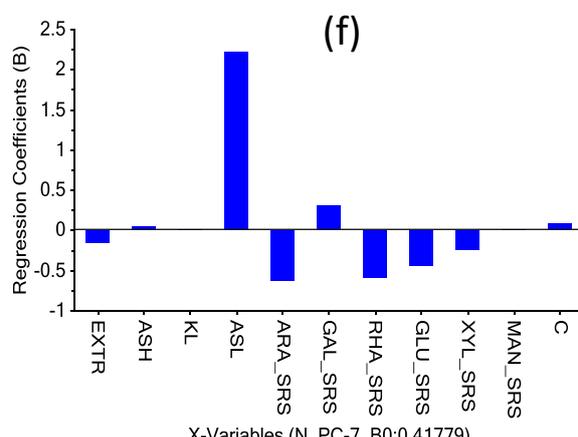
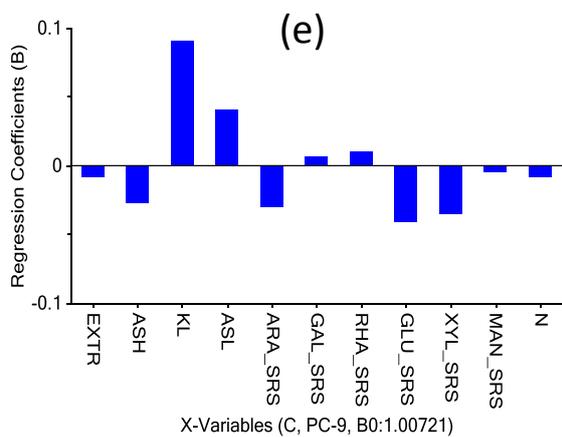
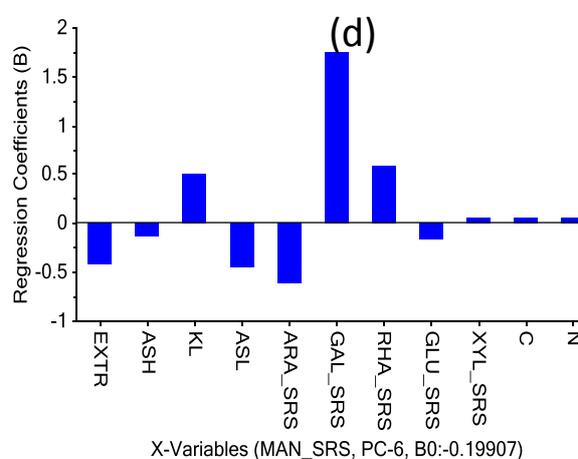
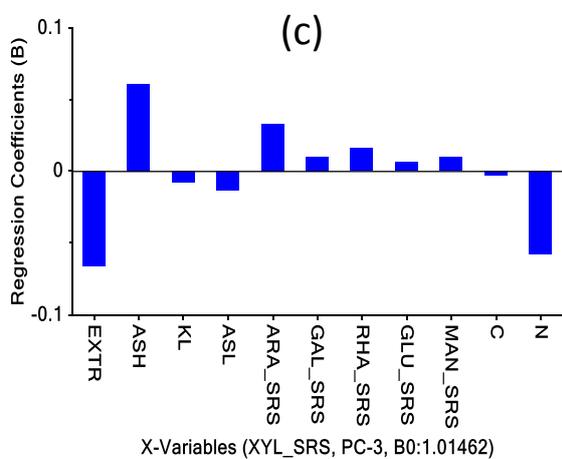
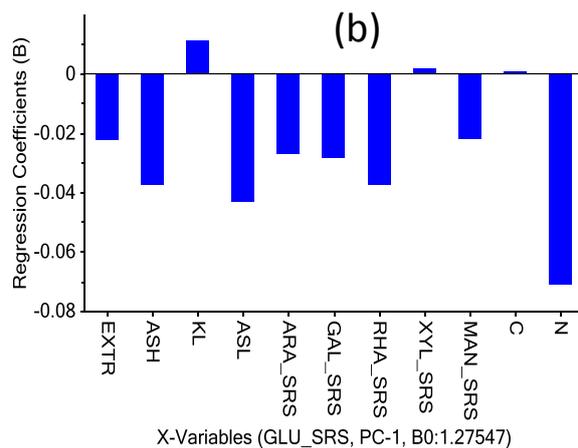
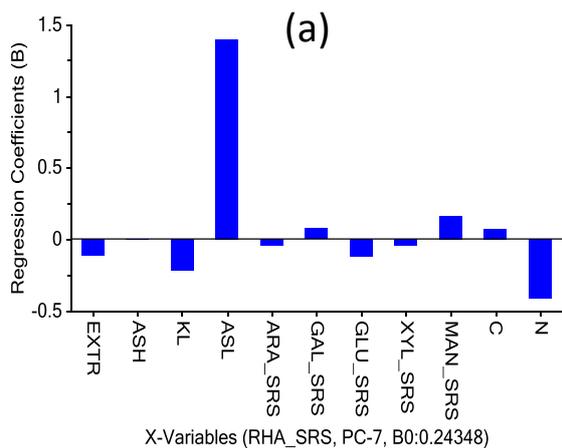


Figure F-5: Regression coefficients plots for PCR models using Miscanthus chemical data. (a) RHA_SRS content PCR, using 7 PCs; (b) GLU_SRS content PCR, using 1 PC; (c) XYL_SRS content PCR, using 3 PCs; (d) MAN_SRS content PCR, using 6 PCs; (e) Carbon (C) content PCR, using 9 PCs; (f) Nitrogen (N) content PCR, using 7 PCs.

Table F-14: Regression statistics for PLSR models for the glucose content (% whole dry mass basis) of the Miscanthus samples in the DT and WU datasets.

DT Dataset										
Pretreat.	NONE	NONE	SNV	SNVDT	MSC	EMSC	SG	SG	SG	SG
Specific			1.1-2.5	2,1.1-2.5	F,1.1-2.5	F,1.1-2.5	1,1,10,10	2,2,25,25	2,2,40,40	4,4,40,40
PLS- λ 10 ³ nm	0.4-2.5	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5
Sam. Excl.										
F-Haaland's	19	17	16	15	16	15	18	11	12	15
R^2_{calib}	0.9727	0.9789	0.9835	0.9821	0.9833	0.9842	0.9873	0.9783	0.9780	0.9848
RMSEC (%)	0.8226	0.7375	0.6384	0.6650	0.6428	0.6249	0.5604	0.7331	0.7380	0.6144
R^2_{CV}	0.9515	0.9616	0.9722	0.9714	0.9727	0.9750	0.9698	0.9650	0.9647	0.9654
RMSECV (%)	1.0989	0.9764	0.8299	0.8417	0.8235	0.7868	0.8683	0.9306	0.9350	0.9262
BIAS _{CV} (%)	-0.0156	0.0025	-0.0081	0.0029	-0.0103	0.0005	-0.0095	-0.0032	0.0040	0.0092
SECV (%)	1.1025	0.9797	0.8327	0.8446	0.8262	0.7894	0.8712	0.9338	0.9381	0.9293
RPD _{CV}	4.5279	5.0956	5.9951	5.9108	6.0424	6.3236	5.7300	5.3461	5.3214	5.3719
RER _{CV}	19.0555	21.4447	25.2303	24.8755	25.4293	26.6130	24.1147	22.4991	22.3951	22.6077
WU Dataset										
Pretreat.	NONE	SG	SG	SG	SG	SNV	SNVDT	SNVDT	MSC	
Specific		1,1,10,10	1,1,10,10	1,1,10,10	2,2,25,25	1.1-2.5	2,1.1-2.5	2,1.1-1.8	F,1.1-2.5	
PLS- λ 10 ³ nm	1.1-1.8	1.1-2.5	1.1-1.8	1.1-1.8, 2.1-2.5	1.1-1.8	1.1-2.5	1.1-2.5	1.1-1.8	1.1-2.5	
Sam. Excl.	258 2530 5125 18008									
F-Haaland's	8	11	14	13	14	12	17	15	16	
R^2_{calib}	0.9189	0.9684	0.9708	0.9711	0.9743	0.9253	0.9514	0.9402	0.9478	
RMSEC (%)	1.4437	0.9008	0.8669	0.8618	0.8126	1.3857	1.1173	1.2395	1.1588	
R^2_{CV}	0.8942	0.9454	0.9545	0.9457	0.9525	0.8906	0.9122	0.9095	0.8917	
RMSECV (%)	1.6560	1.1859	1.0823	1.1825	1.1070	1.6846	1.5117	1.5448	1.6941	
BIAS _{CV} (%)	0.1110	-0.0018	0.0002	0.0082	0.0244	0.0317	-0.0224	0.0241	-0.0453	
SECV (%)	1.6579	1.1900	1.0860	1.1865	1.1105	1.6900	1.5167	1.5499	1.6993	
RPD _{CV}	3.0689	4.2757	4.6852	4.2883	4.5815	3.0106	3.3546	3.2827	2.9941	
RER _{CV}	12.6720	17.6548	19.3456	17.7071	18.9177	12.4311	13.8515	13.5549	12.3632	

Table F-15: Regression statistics for glucose content (% whole dry mass basis) PLSR models. All spectra were transformed by SG 1,1,10,10 prior to model development.

Dataset	DS		DT		DG		DH	DU		DV	WU		DF		
Varieties	All	<i>Giganteus</i>	All	<i>Giganteus</i>	All	<i>Giganteus</i>	<i>Giganteus</i>	All	<i>Giganteus</i>	<i>Giganteus</i>	All	<i>Giganteus</i>	All	<i>Giganteus</i>	
PLS- λ 10 ³ nm	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5	1.-2.5	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5	1.1-1.8	1.1-1.8	1.1-2.5	1.1-2.5	
Samples Excluded	18002	18002			258 5125 18002	5125 18002		258 5125	258 5125		258 2530 5125 18008	258 2530 5125	5049 18008		
Calib:Valid	159:42	119:34	149:42	113:34	125:45	91:31	35:0	128:44	93:31	30:0	147:42	122:34	37:13	33:11	
F-Haaland's	16	14	18	11	16	17	5	14	8	12	14	17	11	6	
R^2_{calib}	0.9836	0.9830	0.9873	0.9826	0.9752	0.9868	0.9725	0.9658	0.9455	0.9979	0.9708	0.9740	0.9941	0.9687	
$Offset_{calib}$	0.6373	0.6706	0.4990	0.6904	0.9749	0.5261	0.9760	1.4306	2.2126	0.0747	1.1278	1.0193	0.2235	1.1974	
RMSEC (%)	0.6355	0.6545	0.5604	0.6585	0.7521	0.5433	0.6368	0.9520	1.0741	0.1798	0.8669	0.8092	0.3833	0.8952	
R^2_{CV}	0.9663	0.9625	0.9698	0.9670	0.9435	0.9456	0.9419	0.9248	0.9159	0.9681	0.9545	0.9459	0.9490	0.8937	
RMSECV (%)	0.9142	0.9748	0.8683	0.9105	1.1394	1.1110	0.9257	1.3778	1.3359	0.6998	1.0823	1.1686	1.1524	1.6667	
$BIAS_{CV}$ (%)	-0.0013	0.0001	-0.0095	-0.0113	-0.0049	-0.0055	-0.0155	0.0302	-0.0008	0.0199	0.0002	0.0071	-0.0014	0.0324	
SECV (%)	0.9171	0.9790	0.8712	0.9145	1.1440	1.1172	0.9391	1.3829	1.3431	0.7115	1.0860	1.1734	1.1683	1.6923	
RPD_{CV}	5.4357	5.1441	5.7300	5.4881	4.1917	4.2606	4.1496	3.6375	3.4441	5.5993	4.6852	4.2902	4.3478	3.0344	
RER_{CV}	22.9086	21.0387	24.1147	22.5219	18.0032	18.4358	13.9587	15.1918	14.5132	18.4235	19.3456	17.5524	15.0954	9.4462	
R^2_{pred}	0.9680	0.9654	0.9595	0.9660	0.9483	0.9272		0.8922	0.9176		0.9305	0.9751	0.9480	0.9266	
$Slope_{pred}$	0.9361	0.9808	0.9917	1.0135	0.9479	1.0034		0.9523	0.9760		0.9722	0.9766	0.9993	0.9337	
$Offset_{pred}$	2.3863	0.6809	0.3311	-0.4960	2.2459	-0.5244		2.1694	0.7443		0.8622	0.8132	0.1228	2.4775	
RMSEP (%)	0.8617	0.9100	0.9471	0.9269	0.9680	1.2439		1.2374	1.3875		1.2663	0.7736	1.3045	1.3527	
$Bias_{pred}$ (%)	-0.1700	-0.0732	-0.0017	0.0346	0.0810	-0.3829		0.2324	-0.2055		-0.2505	-0.1063	0.0971	-0.1043	
SEP (%)	0.8550	0.9207	0.9585	0.9402	0.9755	1.2031		1.2294	1.3949		1.2563	0.7778	1.3540	1.4145	
RPD_{pred}	5.5007	5.3539	4.9063	5.2428	4.3965	3.5578		2.9907	3.4082		3.7436	6.3375	4.2713	3.6887	
RER_{pred}	23.8086	20.8790	21.2359	20.4456	21.2891	14.9445		13.3996	13.3380		16.2032	24.7147	11.1912	11.1547	
All Samples in Calibration Set															
													F-Haaland's	8	8
													R^2_{calib}	0.9743	0.9807
													RMSEC (%)	0.8194	0.7012
													R^2_{CV}	0.9454	0.9532
													RMSECV (%)	1.2432	1.1167
													RPD_{CV}	4.1116	4.5236
													RER_{CV}	14.0544	14.3560

Table F-16: Regression statistics for total sugar content (% whole dry mass) PLSR models. All spectra transformed by SG 1,1,10,10 prior to model development.

Dataset	DS		DT		DG		DH	DU		DV	WU		DF	
Varieties	All	Gig	All	Gig	All	Gig	Gig	All	Gig	Gig	All	Gig	All	Gig
PLS- λ 10 ³ nm	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5	1.1-1.8	1.1-1.8	1.1-2.5	1.1-2.5
Samples Excluded	18002	18002			257 5125 18002	258 5125 18002		258 5125	258 5125		258 2530 5125 18008	258 2530 5125	18008	
Calib:Valid	159:42	119:34	149:42	113:34	125:45	91:31	35:0	128:44	93:31	30:0	147:42	122:34	38:13	33:11
F-Haaland's	14	15	13	10	16	18	3	12	10	13	17	17	11	12
R^2_{calib}	0.9821	0.9830	0.9809	0.9745	0.9788	0.9879	0.8882	0.9575	0.9422	0.9987	0.9502	0.9577	0.9894	0.9925
$Offset_{calib}$	1.1195	1.0646	1.2037	1.5990	1.3383	0.9129	6.4996	2.6895	3.6974	0.0778	3.0955	2.6284	0.6440	0.4564
RMSEC (%)	0.7105	0.6264	0.7408	0.7713	0.7678	0.5719	1.1938	1.1019	1.0547	0.1357	1.1956	0.9993	0.4683	0.4025
R^2_{CV}	0.9677	0.9607	0.9687	0.9589	0.9479	0.9438	0.8141	0.9136	0.9035	0.9455	0.9122	0.9155	0.9322	0.9174
RMSECV (%)	0.9566	0.9550	0.9474	0.9807	1.2161	1.1389	1.5400	1.5744	1.3665	0.9177	1.5930	1.4183	1.1958	1.3437
$BIAS_{CV}$ (%)	-0.0130	-0.0085	-0.0161	-0.0216	-0.0308	-0.0208	-0.0390	0.0331	-0.0056	0.1701	0.0122	0.0454	-0.0168	-0.0992
SECV (%)	0.9595	0.9590	0.9505	0.9848	1.2206	1.1450	1.5620	1.5802	1.3739	0.9172	1.5984	1.4234	1.2117	1.3608
RPD_{CV}	5.5599	5.0272	5.6532	4.9254	4.3426	4.1860	2.3192	3.3966	3.2093	4.1221	3.3647	3.4259	3.8070	3.4695
RER_{CV}	25.0796	23.9633	25.3171	23.3341	18.2553	20.0700	8.0033	15.2282	15.5173	13.6289	15.0545	16.1447	14.9225	12.2911
R^2_{pred}	0.9708	0.9659	0.9535	0.9580	0.9331	0.9237		0.8780	0.9204		0.9443	0.9589	0.8615	0.9549
$Slope_{pred}$	0.9345	0.9748	0.9548	0.9803	0.9566	1.0922		0.9072	1.0269		0.9731	0.9872	0.8691	1.0145
$Offset_{pred}$	4.1607	1.4743	2.9761	1.3855	3.0399	-6.4584		6.4699	-2.1812		1.4654	1.0844	8.1033	-1.1012
RMSEP (%)	0.8706	0.9915	1.0811	1.1095	1.1254	1.4049		1.4831	1.5891		1.2070	1.1295	1.8306	1.0259
$Bias_{pred}$ (%)	0.0188	-0.1017	0.1230	0.1515	0.2362	-0.5361		0.4920	-0.5010		-0.2317	0.2863	0.1374	-0.2129
SEP (%)	0.8809	1.0011	1.0871	1.1156	1.1128	1.3201		1.4152	1.5329		1.1989	1.1090	1.9000	1.0526
RPD_{pred}	5.7215	5.4070	4.6364	4.8520	3.8506	3.0562		2.8514	3.2975		4.2041	4.8810	2.6866	4.5249
RER_{pred}	25.2948	21.6559	20.4976	19.4331	21.6241	13.6041		13.1302	13.8278		18.5866	19.5491	7.8119	13.7245
All Samples in Calibration Set														
												F-Haaland's	8	9
												R^2_{calib}	0.9539	0.9785
												RMSEC (%)	0.9891	0.6785
												R^2_{CV}	0.8865	0.9390
												RMSECV (%)	1.5456	1.1681
												RPD_{CV}	2.9874	3.9583
												RER_{CV}	11.6131	14.1564

Table F-17: Regression statistics for xylose content PLSR models (% whole dry mass). All spectra transformed by SG 2,2,25,25 prior to model development.

Dataset	DS		DT		DG		DH	DU		DV	WU		DF	
Varieties	All	Gig	All	Gig	All	Gig	Gig	All	Gig	Gig	All	Gig	All	Gig
PLS- λ 10 ³ nm	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5	1.1-1.8	1.1-1.8	1.1-2.5	1.1-2.5
Samples Excluded	18002	18002			258 5125 18002	5125 18002		258 5125	258 5125		258 2530 5125 18008	258 2530 5125	18008	
Calib:Valid	159:42	119:34	149:42	113:34	125:45	91:31	35:0	128:44	93:31	30:0	147:42	122:34	38:13	33:11
F-Haaland's	16	17	14	12	14	14	7	14	17	4	14	13	11	7
R^2_{calib}	0.9837	0.9848	0.9778	0.9753	0.9766	0.9649	0.9132	0.9552	0.9585	0.8311	0.9066	0.8729	0.9741	0.9242
$Offset_{calib}$	0.3266	0.2979	0.4428	0.4815	0.5045	0.7013	1.6099	1.0154	0.8301	3.1458	1.8477	2.4822	0.4924	1.4452
RMSEC (%)	0.2635	0.1848	0.3062	0.2301	0.3436	0.2711	0.2985	0.4748	0.2987	0.3922	0.6342	0.5408	0.2695	0.4117
R^2_{CV}	0.9660	0.9508	0.9571	0.9578	0.9565	0.9117	0.6122	0.9178	0.8492	0.7461	0.8605	0.7839	0.8593	0.7288
RMSECV (%)	0.3805	0.3321	0.4262	0.3007	0.4547	0.4314	0.6367	0.6059	0.5730	0.4823	0.7757	0.7074	0.6307	0.7887
$BIAS_{CV}$ (%)	-0.0016	0.0000	-0.0069	-0.0031	0.0055	-0.0034	-0.0674	-0.0016	0.0005	0.0202	0.0114	-0.0064	0.0489	-0.0637
SECV (%)	0.3817	0.3335	0.4276	0.3020	0.4565	0.4338	0.6424	0.6083	0.5761	0.4901	0.7783	0.7103	0.6372	0.7983
RPD_{CV}	5.4257	4.5102	4.8266	4.8674	4.7933	3.3557	1.6000	3.4878	2.5579	1.9803	2.6760	2.1444	2.6613	1.9021
RER_{CV}	31.3364	25.4073	27.9720	22.3473	26.1357	15.5590	7.3333	16.1676	13.6038	8.1479	15.3696	11.9302	12.1160	8.8654
R^2_{pred}	0.9636	0.9485	0.9475	0.9418	0.9392	0.8712		0.9249	0.8324		0.9292	0.8988	0.6363	0.8799
$Slope_{pred}$	0.9581	1.0220	0.9495	0.9766	0.9650	0.8797		0.8965	0.8940		0.9701	0.9575	0.7212	0.8714
$Offset_{pred}$	0.9131	-0.4839	1.0627	0.4326	0.8092	2.2249		2.3403	2.1391		0.6033	0.9420	5.1254	2.4144
RMSEP (%)	0.3864	0.3955	0.4571	0.4013	0.4664	0.6405		0.6245	0.7390		0.5317	0.5412	0.6386	0.6037
$Bias_{pred}$ (%)	0.0868	-0.0483	0.0683	-0.0301	0.1073	-0.1304		0.2320	0.0887		0.0145	0.1004	-0.1890	0.0144
SEP (%)	0.3811	0.3984	0.4574	0.4062	0.4590	0.6375		0.5865	0.7458		0.5380	0.5398	0.6349	0.6330
RPD_{pred}	5.2383	4.1821	4.3648	4.1015	4.0322	2.7852		3.6281	2.4102		3.7112	3.0864	1.6329	2.8842
RER_{pred}	24.0657	16.9027	20.0524	16.5770	20.6527	11.9622		19.5001	8.5444		17.0500	12.4743	4.7094	11.3445
All Samples in Calibration Set														
												F-Haaland's	11	9
												R^2_{calib}	0.9644	0.9581
												RMSEC (%)	0.2889	0.3215
												R^2_{CV}	0.8607	0.8805
												RMSECV (%)	0.5807	0.5552
												RPD_{CV}	2.6632	2.8280
												RER_{CV}	13.3076	13.7490

Table F-18: Regression statistics for arabinose content PLSR models (% whole dry mass). All spectra transformed by SG 2,2,25,25 prior to model development.

Dataset	DS		DT		DG		DH	DU		DV	WU		DF		
Varieties	All	Gig	All	Gig	All	Gig	Gig	All	Gig	Gig	All	Gig	All	Gig	
PLS- λ 10 ³ nm	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5	1.1-1.8	1.1-1.8	1.1-2.5	1.1-2.5	
Samples Excluded	18002	18002			258 5125 18008	258 5125 18002		258 5125	258 5125		258 2530 5125 18008	258 2530 5125	18008		
Calib:Valid	159:42	119:34	149:42	113:34	125:45	91:31	35:0	128:44	93:31	30:0	147:42	122:34	38:13	33:11	
F-Haaland's	14	13	11	13	16	5	6	5	6	4	13	6	11	4	
R^2_{calib}	0.9685	0.9789	0.9647	0.9834	0.9774	0.9446	0.9877	0.9132	0.9472	0.9737	0.9460	0.9400	0.9936	0.9505	
$Offset_{calib}$	0.0792	0.0507	0.0861	0.0385	0.0569	0.1302	0.0343	0.2104	0.1188	0.0763	0.1364	0.1437	0.0158	0.1188	
RMSEC (%)	0.1436	0.1197	0.1521	0.1064	0.1157	0.1759	0.0851	0.2378	0.1760	0.1237	0.1964	0.2021	0.0730	0.2005	
R^2_{CV}	0.9408	0.9504	0.9447	0.9655	0.9441	0.9231	0.9646	0.8963	0.9180	0.9469	0.9196	0.9219	0.9318	0.9024	
RMSECV (%)	0.1972	0.1837	0.1906	0.1536	0.1822	0.2076	0.1446	0.2599	0.2199	0.1764	0.2399	0.2308	0.2410	0.2822	
$BIAS_{CV}$ (%)	0.0005	0.0031	0.0009	-0.0028	0.0057	0.0009	-0.0005	0.0025	-0.0036	0.0087	-0.0018	-0.0001	-0.0139	-0.0032	
SECV (%)	0.1978	0.1845	0.1913	0.1542	0.1828	0.2087	0.1467	0.2609	0.2211	0.1792	0.2407	0.2318	0.2438	0.2865	
RPD_{CV}	4.1012	4.4810	4.2482	5.3810	4.2241	3.6020	5.3052	3.1050	3.4811	4.3320	3.5251	3.5748	3.7886	3.1943	
RER_{CV}	14.4862	15.4143	14.7791	18.4911	15.4631	13.4269	16.6417	11.3004	12.7016	12.7199	11.9078	12.3052	10.4461	9.5141	
R^2_{pred}	0.9399	0.9424	0.9433	0.9534	0.9158	0.8984		0.8911	0.8928		0.8891	0.9093	0.9494	0.9141	
$Slope_{pred}$	0.9254	0.9611	0.9390	0.9374	0.9523	0.8566		0.9630	0.8666		0.8842	0.8945	0.8690	0.9890	
$Offset_{pred}$	0.2171	0.1576	0.1620	0.1683	0.1731	0.2890		0.1172	0.2852		0.3011	0.2899	0.3317	0.0650	
RMSEP (%)	0.1990	0.1862	0.1898	0.1580	0.2389	0.2305		0.2476	0.2320		0.2657	0.2219	0.2168	0.2553	
$Bias_{pred}$ (%)	0.0389	0.0657	0.0162	0.0205	0.0727	0.0062		0.0308	-0.0149		0.0247	0.0406	0.0216	0.0401	
SEP (%)	0.1975	0.1769	0.1914	0.1590	0.2301	0.2342		0.2485	0.2354		0.2677	0.2215	0.2246	0.2644	
RPD_{pred}	4.0701	4.1528	4.1989	4.6200	3.4158	3.1075		2.9529	3.0428		3.0022	3.3160	4.1741	3.2959	
RER_{pred}	14.9653	15.2786	15.4389	16.9978	12.8132	11.0308		11.3705	11.9051		11.0388	12.1999	11.6479	9.3811	
All Samples in Calibration Set															
													F-Haaland's	12	8
													R^2_{calib}	0.9913	0.9714
													RMSEC (%)	0.0846	0.1498
													R^2_{CV}	0.9582	0.9299
													RMSECV (%)	0.1897	0.2401
													RPD_{CV}	4.7974	3.6905
													RER_{CV}	14.2337	11.2239

Table F-19: Regression statistics for galactose content (% whole dry mass) PLSR models. All spectra transformed by SG 2,2,25,25 prior to model development.

Dataset	DS		DT		DG		DH	DU		DV	WU		DF	
Varieties	All	Gig	All	Gig	All	Gig	Gig	All	Gig	Gig	All	Gig	All	Gig
PLS- λ 10 ³ nm	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5	1.1-1.8	1.1-1.8	1.1-2.5	1.1-2.5
Samples Excluded	156 236 18002	156 236 18002	156 236	236	156 236 258 5125 18008	156 5125 18002		156 236 258 5125	156 236 258 5125		156 236 258 2530 5125 18008	236 258 2530 5125	18008	
Calib:Valid	157:42	118:42	147:42	112:34	123:45	90:31	35:0	126:44	91:31	30:0	145:42	121:34	38:13	33:11
F-Haaland's	18	14	13	12	19	4	11	5	2	8	9	8	11	4
R^2_{calib}	0.9672	0.9943	0.9242	0.9308	0.9678	0.8171	0.9887	0.8657	0.7911	0.9882	0.8940	0.8840	0.9805	0.8964
$Offset_{calib}$	0.0260	0.0378	0.0577	0.0525	0.0251	0.1427	0.0101	0.1031	0.1515	0.0104	0.0851	0.0912	0.0163	0.0859
RMSEC (%)	0.0520	0.0646	0.0728	0.0717	0.0488	0.1202	0.0232	0.1077	0.1296	0.0240	0.0947	0.0986	0.0453	0.1018
R^2_{CV}	0.9062	0.8980	0.8667	0.8550	0.8505	0.7749	0.8791	0.8384	0.7636	0.9445	0.8490	0.8138	0.8039	0.7738
RMSECV (%)	0.0886	0.0949	0.0968	0.1042	0.1060	0.1335	0.0768	0.1182	0.1379	0.0522	0.1132	0.1253	0.1466	0.1523
$BIAS_{CV}$ (%)	-0.0003	-0.0012	0.0016	0.0012	0.0013	0.0004	-0.0007	0.0007	0.0003	0.0028	-0.0002	0.0020	-0.0068	0.0071
SECV (%)	0.0889	0.0953	0.0971	0.1047	0.1064	0.1343	0.0779	0.1187	0.1387	0.0530	0.1136	0.1258	0.1484	0.1544
RPD_{CV}	3.2403	3.1261	2.7342	2.6143	2.5649	2.1053	2.8503	2.4855	2.0558	4.2408	2.5688	2.3118	2.2119	2.0797
RER_{CV}	13.1998	12.3211	11.7967	10.9376	10.7659	8.7422	9.4798	9.8917	8.4622	13.3034	10.3342	9.3287	7.2340	6.2419
R^2_{pred}	0.7914	0.7832	0.7945	0.7194	0.8655	0.7050		0.6823	0.7217		0.7869	0.6937	0.7996	0.6272
$Slope_{pred}$	0.8591	0.7443	0.7683	0.7638	0.8464	0.7931		0.8751	0.7265		0.8050	0.7010	1.1755	0.7770
$Offset_{pred}$	0.1205	0.1826	0.1842	0.2056	0.1071	0.1614		0.0804	0.1650		0.1656	0.2784	-0.1826	0.1539
RMSEP (%)	0.1205	0.1210	0.1178	0.1343	0.1049	0.1484		0.1427	0.1500		0.1209	0.1465	0.1517	0.1801
$Bias_{pred}$ (%)	0.0152	-0.0084	0.0111	0.0269	0.0040	0.0304		-0.0099	-0.0538		0.0199	0.0522	-0.0309	-0.0384
SEP (%)	0.1210	0.1222	0.1187	0.1336	0.1061	0.1477		0.1440	0.1423		0.1207	0.1389	0.1546	0.1846
RPD_{pred}	2.1599	2.1382	2.2015	1.8788	2.7228	1.8076		1.6392	1.8955		2.1640	1.8067	1.6284	1.5645
RER_{pred}	7.9522	7.8720	8.1051	7.1992	10.7116	6.9099		7.7690	7.6706		7.9671	6.9229	5.2491	4.9143
All Samples in Calibration Set														
												F-Haaland's	11	10
												R^2_{calib}	0.9488	0.9667
												RMSEC (%)	0.0691	0.0560
												R^2_{CV}	0.8052	0.8403
												RMSECV (%)	0.1369	0.1255
												RPD_{CV}	2.2318	2.4455
												RER_{CV}	7.7691	7.6247

Table F-20: Regression statistics for rhamnose content (% whole dry mass) PLSR models. All spectra transformed by SG 2,2,25,25 prior to model development.

Dataset	DS		DT		DG		DH	DU		DV	WU		DF		
Varieties	All	Gig	All	Gig	All	Gig	Gig	All	Gig	Gig	All	Gig	All	Gig	
PLS- λ 10 ³ nm	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5	1.1-1.8	1.1-1.8	1.1-2.5	1.1-2.5	
Samples Excluded	18002	18002			258 5125 18002	5125 18002		258 5125	258,5125		258 2530 5125 18008	258 2530 5125	18008		
Calib:Valid	159:42	119:34	149:42	113:34	125:45	91:31	35:0	128:44	93:31	30:0	147:42	122:34	38:13	33:11	
F-Haaland's	15	11	17	10	12	13	1	8	8	14	16	18	8	8	
R^2_{calib}	0.9574	0.9077	0.9654	0.9364	0.9148	0.9295	0.8291	0.8487	0.8679	0.9985	0.9121	0.9419	0.9505	0.9542	
$Offset_{calib}$	0.0079	0.0152	0.0062	0.0102	0.0153	0.0108	0.0376	0.0267	0.0198	0.0007	0.0166	0.0096	0.0088	0.0074	
RMSEC (%)	0.0258	0.0242	0.0221	0.0201	0.0365	0.0182	0.0408	0.0498	0.0263	0.0050	0.0386	0.0194	0.0184	0.0137	
R^2_{CV}	0.9040	0.8306	0.9100	0.8757	0.8522	0.7999	0.7967	0.7821	0.7624	0.9084	0.8423	0.8403	0.8346	0.7540	
RMSECV (%)	0.0387	0.0331	0.0357	0.0282	0.0482	0.0312	0.0445	0.0599	0.0359	0.0286	0.0519	0.0326	0.0344	0.0329	
$BIAS_{CV}$ (%)	0.0008	0.0006	-0.0007	-0.0004	0.0006	0.0002	0.0010	-0.0009	0.0013	0.0001	-0.0002	-0.0001	0.0010	0.0008	
SECV (%)	0.0388	0.0332	0.0359	0.0283	0.0484	0.0313	0.0451	0.0601	0.0360	0.0291	0.0521	0.0328	0.0348	0.0334	
RPD_{CV}	3.2225	2.4105	3.3281	2.8300	2.5917	2.2018	2.2177	2.1378	2.0222	3.2995	2.5088	2.4703	2.4037	1.9517	
RER_{CV}	18.4418	10.8026	19.9845	12.6574	14.7938	10.2463	7.1476	11.9649	8.8784	10.3520	13.7645	10.9472	8.9794	7.0403	
R^2_{pred}	0.9439	0.8743	0.9430	0.8780	0.9217	0.7448		0.8550	0.7944		0.8162	0.8451	0.5736	0.8728	
$Slope_{pred}$	1.1491	0.9372	1.1619	0.8629	0.8089	1.1560		0.8317	1.1614		1.0762	0.9490	0.7196	1.1184	
$Offset_{pred}$	-0.0085	0.0214	-0.0162	0.0213	0.0329	-0.0092		0.0208	-0.0157		-0.0034	0.0157	0.0503	-0.0290	
RMSEP (%)	0.0422	0.0311	0.0416	0.0281	0.0397	0.0360		0.0479	0.0292		0.0634	0.0337	0.0524	0.0289	
$Bias_{pred}$ (%)	0.0173	0.0114	0.0119	-0.0005	0.0028	0.0104		-0.0070	0.0061		0.0098	0.0076	0.0023	-0.0098	
SEP (%)	0.0390	0.0294	0.0403	0.0285	0.0401	0.0350		0.0479	0.0290		0.0634	0.0333	0.0544	0.0286	
RPD_{pred}	3.1497	2.7714	3.0447	2.8605	3.2946	1.4401		2.6200	1.6329		1.9367	2.4425	1.4687	2.2569	
RER_{pred}	13.6218	11.4091	13.1676	11.7758	17.1228	5.3472		12.4348	7.2484		8.3757	10.0551	4.1465	7.5093	
All Samples in Calibration Set															
													F-Haaland's	10	9
													R^2_{calib}	0.9449	0.9413
													RMSEC (%)	0.0191	0.0154
													R^2_{CV}	0.8127	0.7842
													RMSECV (%)	0.0366	0.0302
													RPD_{CV}	2.2200	2.1046
													RER_{CV}	8.4653	7.7041

Table F-21: Regression statistics for mannose content (% whole dry mass) PLSR models. All spectra transformed by SG 2,2,25,25 prior to model development.

Dataset	DS		DT		DG		DH	DU		DV	WU		DF	
Varieties	All	Gig	All	Gig	All	Gig	Gig	All	Gig	Gig	All	Gig	All	Gig
PLS-λ 10 ³ nm	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5	1.1-1.8	1.1-1.8	1.1-2.5	1.1-2.5
Samples Excluded	18002	18002		236	258 5125 18008	258 5125 18002		258 5125	258 5125		258 2530 5125 18008	258 2530 5125	18008	
Calib:Valid	159:42	119:34	149:42	112:34	125:45	91:31	35:0	128:44	93:31	30:0	147:42	122:34	38:13	33:11
F-Haaland's	10	11	5	6	11	9	3	11	18	5	10	9	4	8
R^2_{calib}	0.7877	0.8297	0.6803	0.7582	0.7803	0.7935	0.7756	0.8312	0.9342	0.9325	0.8025	0.7776	0.8127	0.9121
$Offset_{calib}$	0.0428	0.0357	0.0602	0.0463	0.0450	0.0449	0.0436	0.0334	0.0131	0.0121	0.0396	0.0457	0.0444	0.0223
RMSEC (%)	0.0599	0.0567	0.0639	0.0577	0.0584	0.0651	0.0315	0.0550	0.0368	0.0139	0.0584	0.0623	0.0664	0.0490
R^2_{CV}	0.6775	0.6940	0.6118	0.6871	0.6422	0.6323	0.6516	0.7142	0.7045	0.7849	0.7079	0.6789	0.6678	0.6971
RMSECV (%)	0.0743	0.0766	0.0706	0.0658	0.0750	0.0878	0.0395	0.0720	0.0811	0.0250	0.0714	0.0751	0.0897	0.0943
$BIAS_{CV}$ (%)	0.0002	-0.0005	-0.0011	-0.0002	-0.0009	0.0002	-0.0014	-0.0012	-0.0011	-0.0020	0.0000	-0.0007	0.0036	-0.0019
SECV (%)	0.0745	0.0769	0.0708	0.0661	0.0753	0.0883	0.0400	0.0723	0.0815	0.0253	0.0717	0.0754	0.0908	0.0957
RPD_{CV}	1.7488	1.7924	1.6016	1.7840	1.6598	1.6313	1.6887	1.8577	1.7719	2.1482	1.8392	1.7599	1.7121	1.7540
RER_{CV}	8.6309	8.3613	8.8164	9.4395	8.5365	7.2850		8.8974	7.8929	8.0635	8.9733	8.5333	7.5175	7.1809
R^2_{pred}	0.6094	0.6010	0.5019	0.4073	0.7510	0.5686		0.6861	0.5599		0.5746	0.5132	0.7888	0.7916
$Slope_{pred}$	0.7114	0.6802	0.5487	0.5178	0.7535	0.6215		0.7915	0.5932		0.7254	0.5794	0.8025	0.9513
$Offset_{pred}$	0.0495	0.0660	0.0762	0.0837	0.0659	0.0786		0.0118	0.0952		0.0544	0.0809	0.0315	-0.0043
RMSEP (%)	0.0647	0.0701	0.0719	0.0855	0.0718	0.0754		0.0685	0.0791		0.0694	0.0766	0.0763	0.0788
$Bias_{pred}$ (%)	-0.0013	0.0103	-0.0033	-0.0003	0.0263	0.0243		-0.0277	0.0088		0.0060	0.0076	-0.0170	-0.0154
SEP (%)	0.0655	0.0704	0.0727	0.0868	0.0676	0.0726		0.0634	0.0799		0.0699	0.0774	0.0774	0.0811
RPD_{pred}	1.5662	1.5629	1.4107	1.2673	2.0040	1.5139		1.7404	1.5040		1.4665	1.4208	2.1750	2.0385
RER_{pred}	5.9007	7.4724	5.3146	6.0592	9.2301	4.9966		6.8015	5.3981		5.5248	6.7930	8.1202	7.1629
All Samples in Calibration Set														
												F-Haaland's	5	10
												R^2_{calib}	0.8340	0.9512
												RMSEC (%)	0.0634	0.0362
												R^2_{CV}	0.7482	0.7880
												RMSECV (%)	0.0797	0.0772
												RPD_{CV}	1.9540	2.1230
												RER_{CV}	8.5420	8.8077

Table F-22: Regression statistics for Klason lignin (KL) content (% whole dry mass) PLSR models. All spectra transformed by SG 1,1,10,10 prior to model development.

Dataset	DS		DT		DG		DH	DU		DV	WU		DF	
Varieties	All	Gig	All	Gig	All	Gig	Gig	All	Gig	Gig	All	Gig	All	Gig
PLS- λ 10 ³ nm	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5	1.1-1.8	1.1-1.8	1.1-2.5	1.1-2.5
Samples Excluded	18002 18008	18002	18008		258 5125 18002	258 5125 18002		250 258 5125 18008	250 258 5125		258 2530 5125 18008	258 2530 5125	18008	
Calib:Valid	172:39	136:31	157:39	126:31	120:43	104:31	35:0	134:44	105:29	31:0	159:39	138:31	36:13	31:11
F-Haaland's	13	8	7	8	6	8	7	8	8	7	10	10	9	8
R^2_{calib}	0.9750	0.9691	0.9595	0.9710	0.9560	0.9707	0.9920	0.9500	0.9604	0.9863	0.9494	0.9411	0.9929	0.9922
$Offset_{calib}$	0.4546	0.5728	0.7462	0.5895	0.7995	0.5493	0.1349	1.0546	0.7481	0.2260	0.9204	1.0931	0.1346	0.1509
RMSEC (%)	0.4106	0.4770	0.5325	0.4925	0.5538	0.4470	0.2506	0.6031	0.5007	0.3211	0.6245	0.6793	0.2469	0.2541
R^2_{CV}	0.9565	0.9583	0.9439	0.9568	0.9440	0.9541	0.9669	0.9243	0.9448	0.9594	0.9320	0.9173	0.9705	0.9659
RMSECV (%)	0.5428	0.5544	0.6278	0.5775	0.6252	0.5596	0.5101	0.6952	0.5910	0.5529	0.7249	0.8055	0.5071	0.5342
$BIAS_{CV}$ (%)	-0.0046	-0.0082	-0.0039	-0.0065	-0.0026	0.0061	0.0095	-0.0044	-0.0035	-0.0265	0.0179	0.0009	0.0087	-0.0059
SECV (%)	0.5444	0.5564	0.6298	0.5798	0.6278	0.5623	0.5174	0.6978	0.5939	0.5614	0.7270	0.8085	0.5142	0.5430
RPD_{CV}	4.7871	4.8942	4.2180	4.8111	4.2249	4.6676	5.4963	3.6341	4.2550	4.9626	3.8321	3.4735	5.7598	5.4003
RER_{CV}	20.2905	20.8108	17.5396	19.9707	18.4437	20.5908	19.2656	16.5930	18.3601	17.7575	15.1947	14.3213	21.2125	20.0870
R^2_{pred}	0.9495	0.9783	0.9502	0.9747	0.9067	0.9532		0.9465	0.9666		0.8787	0.9576	0.9671	0.9593
$Slope_{pred}$	0.8684	0.9378	0.8899	0.8988	0.9888	1.0238		0.9338	0.9548		0.8623	0.9330	1.0507	0.9096
$Offset_{pred}$	2.4642	1.1499	2.1936	1.8517	0.1505	-0.5493		1.3533	0.8215		2.6534	1.3987	-0.9464	1.6801
RMSEP (%)	0.6319	0.4167	0.6384	0.4812	0.6786	0.6121		0.5889	0.5198		0.9257	0.5980	0.5888	0.5492
$Bias_{pred}$ (%)	0.0887	0.0394	0.2051	0.0470	-0.0645	-0.0914		0.1184	-0.0429		0.1664	0.2033	-0.0007	0.0049
SEP (%)	0.6338	0.4217	0.6125	0.4868	0.6835	0.6152		0.5835	0.5272		0.9226	0.5716	0.6129	0.5759
RPD_{pred}	4.1746	6.5379	4.3199	5.6632	3.1502	4.3824		4.3162	5.4617		2.8680	4.8233	4.9901	4.8071
RER_{pred}	15.8067	21.3495	16.3570	18.4932	11.5579	15.9996		17.7044	20.9808		10.8594	15.7505	17.1471	15.8585
All Samples in Calibration Set														
												F-Haaland's	8	7
												R^2_{calib}	0.9878	0.9869
												RMSEC (%)	0.3234	0.3259
												R^2_{CV}	0.9714	0.9657
												RMSECV (%)	0.5048	0.5407
												RPD_{CV}	5.7959	5.2725
												RER_{CV}	21.3839	19.9321

Table F-23: Regression statistics for acid soluble lignin (ASL) content (% whole dry mass). All spectra transformed by SG 2,2,25,25 prior to model development.

Dataset	DS		DT		DG		DH	DU		DV	WU		DF		
Varieties	All	Gig	All	Gig	All	Gig	Gig	All	Gig	Gig	All	Gig	All	Gig	
PLS- λ 10 ³ nm	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5	1.1-1.8	1.1-1.8	1.1-2.5	1.1-2.5	
Samples Excluded	18002 18008	18002			258 5125	258 5125 18002		258 5125	258 5125		258 2530 5125 18008	258 2530 5125	18008		
Calib:Valid	176:41	137:34	161:41	126:34	142:43	106:31	35:0	142:43	108:30	31:0	163:41	138:34	37:13	33:11	
F-Haaland's	10	8	7	9	14	12	2	7	12	2	10	17	6	6	
R^2_{calib}	0.9635	0.9660	0.9671	0.9811	0.9762	0.9789	0.9589	0.9351	0.9701	0.9401	0.9337	0.9709	0.9801	0.9727	
$Offset_{calib}$	0.0946	0.0857	0.0826	0.0460	0.0610	0.0512	0.1292	0.1608	0.0696	0.1932	0.1791	0.0741	0.0480	0.0653	
RMSEC (%)	0.2037	0.1932	0.1911	0.1430	0.1651	0.1429	0.2154	0.2674	0.1610	0.2572	0.2886	0.1796	0.1530	0.1853	
R^2_{CV}	0.9492	0.9535	0.9556	0.9681	0.9524	0.9543	0.9264	0.9123	0.9321	0.9117	0.9029	0.9397	0.9463	0.9162	
RMSECV (%)	0.2406	0.2258	0.2218	0.1857	0.2337	0.2104	0.2920	0.3113	0.2442	0.3127	0.3453	0.2601	0.2519	0.3273	
$BIAS_{CV}$ (%)	-0.0002	-0.0031	0.0012	-0.0020	0.0051	-0.0008	-0.0099	-0.0068	0.0086	0.0076	0.0052	0.0044	0.0067	0.0105	
SECV (%)	0.2413	0.2266	0.2225	0.1865	0.2345	0.2114	0.2961	0.3124	0.2452	0.3178	0.3463	0.2610	0.2553	0.3322	
RPD_{CV}	4.4327	4.6390	4.7470	5.5979	4.5767	4.6755	3.6401	3.3736	3.8133	3.3630	3.2005	4.0499	4.3113	3.4286	
RER_{CV}	16.3724	19.8636	17.7600	24.1381	18.7492	21.2877	11.0610	14.4103	16.1147	9.7288	11.4091	17.2442	15.2459	11.5044	
R^2_{pred}	0.9477	0.9689	0.9434	0.9723	0.8739	0.9793		0.9070	0.9680		0.8540	0.9285	0.9100	0.9756	
$Slope_{pred}$	0.8670	0.9471	0.8812	0.8877	1.0334	1.1333		0.8275	0.9637		0.8175	0.9128	0.9070	0.9958	
$Offset_{pred}$	0.2753	0.1239	0.2699	0.2594	0.0009	-0.2037		0.3280	0.1146		0.4045	0.2573	0.3090	0.0351	
RMSEP (%)	0.2710	0.1923	0.2698	0.2047	0.3301	0.2115		0.3039	0.2002		0.4217	0.2908	0.3547	0.1523	
$Bias_{pred}$ (%)	-0.0566	-0.0134	-0.0265	-0.0318	0.0728	0.0772		-0.0755	0.0268		-0.0511	0.0311	0.0786	0.0254	
SEP (%)	0.2683	0.1947	0.2718	0.2052	0.3257	0.2001		0.2979	0.2018		0.4238	0.2935	0.3600	0.1575	
RPD_{pred}	4.1106	5.6263	4.0582	5.3381	2.5377	4.7161		3.1624	5.5902		2.6028	3.7323	3.3335	6.3538	
RER_{pred}	16.7736	17.3855	16.5598	16.4951	11.8684	19.7450		13.6067	22.3066		10.6211	11.5330	9.4089	19.8052	
All Samples in Calibration Set															
													F-Haaland's	6	6
													R^2_{calib}	0.9677	0.9734
													RMSEC (%)	0.1984	0.1766
													R^2_{CV}	0.9332	0.9473
													RMSECV (%)	0.2911	0.2544
													RPD_{CV}	3.7927	4.2559
													RER_{CV}	13.2379	15.0299

Table F-24: Regression statistics for uronic acids content (% whole dry mass). All spectra transformed by SG 2,2,25,25 prior to model development. Only (full) cross validation was used to test the models.

Dataset	DS	DT	DG	DU	WU
Varieties	Gig	Gig	Gig	Gig	Gig
PLS- λ 10 ³ nm	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5
Calib:Valid	31:0	24:0	31:0	32:0	31:0
F-Haaland's	8	6	11	7	6
R^2_{calib}	0.9684	0.9306	0.9856	0.8985	0.8281
$Offset_{calib}$	0.0491	0.1032	0.0224	0.1572	0.2674
RMSEC (%)	0.0639	0.0812	0.0431	0.1134	0.1491
R^2_{CV}	0.8670	0.7703	0.8396	0.7161	0.5685
RMSECV (%)	0.1312	0.1508	0.1467	0.1915	0.2457
$BIAS_{CV}$ (%)	0.0031	0.0047	-0.0095	-0.0057	-0.0003
SECV (%)	0.1334	0.1540	0.1488	0.1945	0.2498
RPD_{CV}	2.7408	2.0442	2.4573	1.8600	1.4636
RER_{CV}	10.3412	6.4702	9.2714	7.0911	5.5221

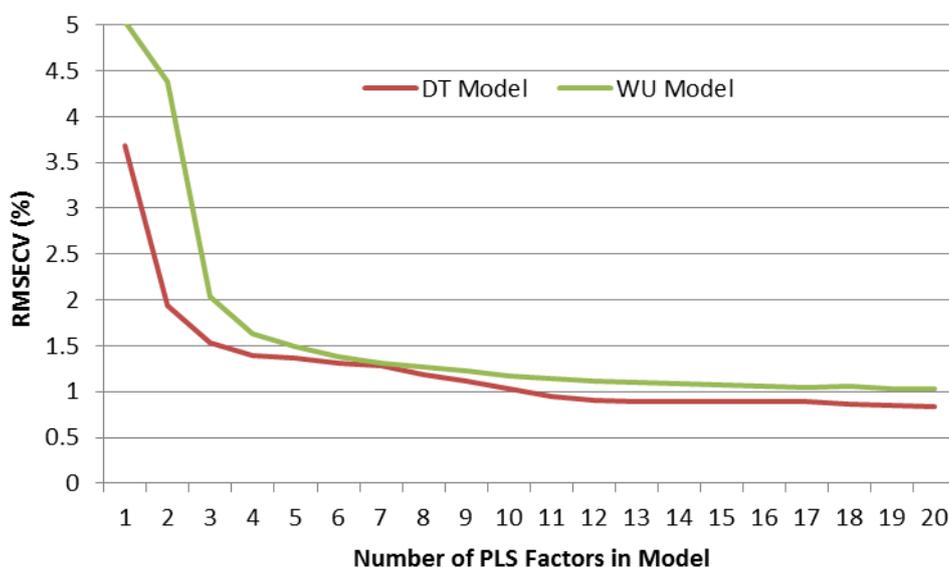


Figure F-6: RMSECV plotted as a function of number of PLS factors in the models developed for the glucose content (% whole dry mass basis) of the DT and WU datasets. For both datasets the spectra were transformed by SG-1,1,10,10. For the DT model the 1100-2500 nm region was used for calibration, while the 1100-1800 nm region was used for the WU model.

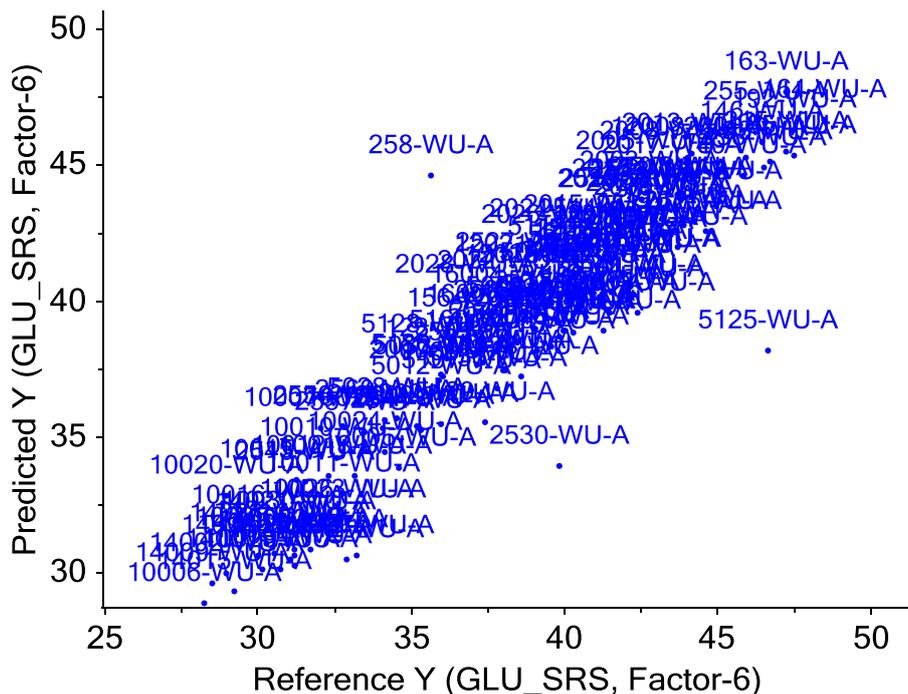


Figure F-7: A predicted glucose content vs. reference glucose content for a model developed on the WU spectra. Samples 258 (internode), 5125 (node), and 2530 (WP) are clear outliers compared with the other 147 samples in the calibration set.

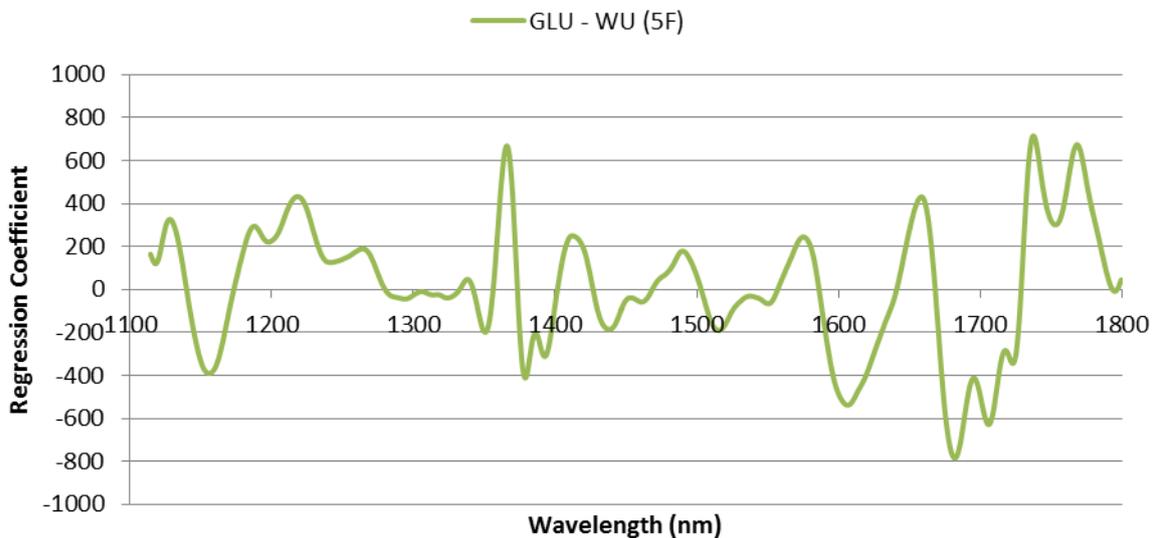


Figure F-8: A regression coefficient plot for a 5 PLS factor model for the glucose content of WU samples. The spectra were pretreated by SG-1,1,10,10 prior to model development.

Table F-25: Regression statistics for ethanol-soluble extractives (EXTR) content (% whole dry mass). All spectra were transformed by SG 2,2,25,25 prior to model development.

Dataset	DS		DT		DG		DH	DU		DV	WU		DF		
Varieties	All	Gig	All	Gig	All	Gig	Gig	All	Gig	Gig	All	Gig	All	Gig	
PLS-λ 10 ³ nm	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5	
Samples Excluded	18002 18004 18008	18002			258 5125 18008	258 5125 18002		258 5125	258 5125		258 2530 5125 18004 18008	258 2530 5125	18008		
Calib:Valid	201:42	148:34	188:42	139:34	161:45	113:31	43:0	164:44	115:31	37:0	188:42	152:34	45:13	40:11	
F-Haaland's	15	15	14	13	13	15	7	13	11	8	17	16	10	9	
R^2_{calib}	0.9614	0.9669	0.9600	0.9617	0.9464	0.9645	0.9635	0.9155	0.8940	0.9817	0.9298	0.9238	0.9752	0.9704	
$Offset_{calib}$	0.2487	0.2186	0.2542	0.2510	0.3280	0.2101	0.3328	0.5119	0.6178	0.1762	0.4770	0.5112	0.2109	0.2375	
RMSEC (%)	0.5771	0.5332	0.5951	0.5879	0.6233	0.4861	0.6363	0.7514	0.7481	0.4487	0.7786	0.8311	0.5502	0.5898	
R^2_{CV}	0.9333	0.9301	0.9324	0.9261	0.9070	0.8968	0.9095	0.8631	0.8143	0.9263	0.8421	0.7872	0.8914	0.8826	
RMSECV (%)	0.7611	0.7764	0.7742	0.8189	0.8221	0.8345	1.0022	0.9599	0.9975	0.9133	1.1795	1.4372	1.1687	1.1823	
$BIAS_{CV}$ (%)	0.0104	0.0048	-0.0089	-0.0023	-0.0002	0.0183	0.0298	0.0008	0.0008	-0.0666	0.0244	0.0060	0.0107	0.0379	
SECV (%)	0.7629	0.7790	0.7762	0.8219	0.8247	0.8380	1.0137	0.9628	1.0019	0.9234	1.1824	1.4420	1.1819	1.1968	
RPD_{CV}	3.8597	3.7746	3.8426	3.6698	3.2742	3.0941	3.3242	2.6931	2.3039	3.6374	2.4920	2.0944	2.9923	2.9006	
RER_{CV}	17.2139	16.8584	16.9876	16.0430	13.6869	12.9993	12.8161	11.7225	10.8194	14.0690	11.1519	9.1441	11.2405	10.0106	
R^2_{pred}	0.9291	0.9471	0.9199	0.9497	0.8025	0.9075		0.8605	0.8684		0.8088	0.8910	0.8504	0.9163	
$Slope_{pred}$	0.9904	0.9722	1.0010	0.9276	0.8621	1.0001		0.8361	0.9068		0.8957	0.9735	0.9153	0.9391	
$Offset_{pred}$	-0.0487	0.1953	-0.2254	0.5020	0.7570	0.0947		0.6893	0.6738		0.3651	-0.1007	0.7910	0.4957	
RMSEP (%)	0.8253	0.7218	0.9082	0.7037	0.9995	0.7192		0.9772	1.0681		1.3830	1.1042	1.1222	1.0703	
$Bias_{pred}$ (%)	-0.1149	-0.0023	-0.2182	-0.0124	-0.0133	0.0955		-0.2667	0.0506		-0.3547	-0.2891	0.1425	-0.0118	
SEP (%)	0.8272	0.7327	0.8922	0.7141	1.0107	0.7246		0.9509	1.0846		1.3530	1.0817	1.1586	1.1225	
RPD_{pred}	3.6523	4.3225	3.3859	4.4348	1.8690	3.1322		2.6706	2.7391		2.2328	2.9278	2.5432	3.4445	
RER_{pred}	12.8103	15.9174	11.8758	16.3308	2.2255	10.3294		10.2142	8.9730		7.8316	10.7816	9.4834	10.9982	
All Samples in Calibration Set															
													F-Haaland's	11	8
													R^2_{calib}	0.9646	0.9612
													RMSEC (%)	0.6307	0.6873
													R^2_{CV}	0.9021	0.9080
													RMSECV (%)	1.0664	1.0792
													RPD_{CV}	3.1444	3.2342
													RER_{CV}	12.3633	12.1982

Table F-26: Regression statistics for ash content (% whole dry mass basis). All spectra were transformed by SG 2,2,25,25 prior to model development.

Dataset	DS		DT		DG		DH	DU		DV	WU		DF		
Varieties	All	Gig	All	Gig	All	Gig	Gig	All	Gig	Gig	All	Gig	All	Gig	
PLS- λ 10 ³ nm	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5	
Samples Excluded	18002 18008	18002			258 5125 18002 18008	258 5125 18002		258 5125	258 5125		258 2530 5125 18008	258 2530 5125	18008		
Calib:Valid	190:39	141:28	179:39	134:28	149:41	102:29	44:0	153:39	105:28	38:0	178:39	146:28	45:13	40:11	
F-Haaland's	13	13	14	12	14	14	8	11	11	10	14	12	11	12	
R^2_{calib}	0.9694	0.9680	0.9625	0.9627	0.9719	0.9695	0.9465	0.9294	0.9378	0.9770	0.8936	0.9017	0.9855	0.9910	
$Offset_{calib}$	0.1274	0.1288	0.1487	0.1424	0.1181	0.1225	0.2373	0.3153	0.2428	0.1032	0.4451	0.3951	0.0546	0.0348	
RMSEC (%)	0.3676	0.3482	0.3882	0.3450	0.3808	0.3626	0.3322	0.6339	0.5298	0.2023	0.7076	0.6068	0.2363	0.2116	
R^2_{CV}	0.9510	0.9405	0.9393	0.9339	0.9532	0.9334	0.8070	0.8925	0.8914	0.8521	0.8045	0.8017	0.9143	0.9353	
RMSECV (%)	0.4653	0.4747	0.4937	0.4592	0.4920	0.5366	0.6367	0.7503	0.7021	0.5188	0.9628	0.8645	0.5785	0.5715	
$BIAS_{CV}$ (%)	0.0046	0.0019	0.0108	-0.0051	0.0138	-0.0111	0.0716	-0.0009	-0.0206	-0.0712	-0.0233	-0.0322	0.0500	0.0607	
SECV (%)	0.4665	0.4764	0.4950	0.4609	0.4934	0.5391	0.6400	0.7528	0.7052	0.5208	0.9653	0.8669	0.5829	0.5755	
RPD_{CV}	4.5143	4.0993	4.0598	3.8892	4.6215	3.8730	2.2698	3.0481	3.0265	2.5929	2.2536	2.2398	3.4096	3.9213	
RER_{CV}	26.0271	20.5107	24.5273	18.8640	24.6049	18.1240	10.6488	16.1275	13.8563	10.8830	12.5776	11.2720	12.3537	14.1937	
R^2_{pred}	0.9368	0.9663	0.9189	0.9787	0.9114	0.9402		0.8789	0.9175		0.8236	0.8377	0.9729	0.9608	
$Slope_{pred}$	0.8827	0.9605	0.8753	0.8799	0.9451	0.9765		0.9815	1.0390		0.8061	0.8619	0.8448	0.8702	
$Offset_{pred}$	0.4158	0.3474	0.3891	0.6139	0.1974	0.2742		-0.0688	-0.1001		1.1451	0.5950	0.3918	0.4066	
RMSEP (%)	0.5249	0.4324	0.5946	0.3962	0.5253	0.5369		0.6840	0.6384		0.9293	0.8582	0.6827	0.5267	
$Bias_{pred}$ (%)	-0.0529	0.1867	-0.1089	0.1256	0.0099	0.1968		-0.1400	0.0512		0.3707	0.0338	-0.3154	-0.1380	
SEP (%)	0.5290	0.3972	0.5922	0.3827	0.5317	0.5084		0.6783	0.6481		0.8634	0.8732	0.6302	0.5331	
RPD_{pred}	3.8819	5.4443	3.4678	5.6511	3.3365	4.0414		2.7410	3.1855		2.3787	2.4765	4.7689	4.5757	
RER_{pred}	17.3967	21.9680	15.5408	22.8025	15.3669	13.8121		10.3493	13.9419		10.6603	9.9929	13.5194	15.9277	
All Samples in Calibration Set															
													F-Haaland's	12	12
													R^2_{calib}	0.9832	0.9867
													RMSEC (%)	0.2892	0.2595
													R^2_{CV}	0.9419	0.9417
													RMSECV (%)	0.5468	0.5550
													RPD_{CV}	4.0846	4.0753
													RER_{CV}	16.0465	15.8190

Table F-27: Regression statistics for acid insoluble residue (AIR) content (% whole dry mass basis). All spectra were transformed by SG 1,1,10,10 prior to model development.

Dataset	DS		DT		DG		DH	DU		DV	WU		DF		
Varieties	All	Gig	All	Gig	All	Gig	Gig	All	Gig	Gig	All	Gig	All	Gig	
PLS- λ 10 ³ nm	1,1-2,5	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5	
Samples Excluded	18002 18008	18002	18008		258 5125 18002 18008	258 5125 18002		250 258 5125 18008	250 258 5125		258 2530 5125 18008	258 2530 5125	18008		
Calib:Valid	172:41	135:33	158:41	126:33	137:44	105:31	36:0	138:43	106:30	31:0	160:41	138:33	36:13	31:11	
F-Haaland's	17	19	12	14	8	8	8	8	12	10	12	12	8	7	
R^2_{calib}	0.9729	0.9819	0.9622	0.9676	0.9474	0.9554	0.9870	0.9325	0.9544	0.9940	0.9226	0.9245	0.9797	0.9782	
$Offset_{calib}$	0.5308	0.3574	0.7481	0.6468	1.0340	0.8919	0.2387	1.3419	0.9166	0.1074	1.5158	1.4965	0.4053	0.4519	
RMSEC (%)	0.3944	0.3357	0.4621	0.4466	0.5738	0.5177	0.2753	0.6079	0.5014	0.1848	0.7009	0.6996	0.3800	0.4037	
R^2_{CV}	0.9453	0.9427	0.9370	0.9283	0.9175	0.9331	0.9182	0.9040	0.9175	0.9652	0.8768	0.8630	0.9172	0.9157	
RMSECV (%)	0.5616	0.6001	0.5970	0.6671	0.7199	0.6347	0.7038	0.7258	0.6759	0.4466	0.8852	0.9439	0.7686	0.7944	
$BIAS_{CV}$ (%)	0.0004	-0.0253	-0.0048	0.0140	-0.0163	0.0060	0.0649	-0.0081	0.0119	-0.0043	-0.0043	0.0035	0.0302	0.0355	
SECV (%)	0.5633	0.6018	0.5989	0.6696	0.7224	0.6377	0.7107	0.7284	0.6790	0.4539	0.8879	0.9473	0.7789	0.8068	
RPD_{CV}	4.2684	4.1632	3.9816	3.7182	3.4749	3.8628	3.4436	3.2241	3.4743	5.3484	2.8459	2.6974	3.4740	3.4450	
RER_{CV}	19.0400	18.3343	17.9067	16.4770	15.2742	17.3022	12.6756	15.1485	16.0034	19.8462	12.0778	11.6473	13.5566	13.2852	
R^2_{pred}	0.9284	0.9550	0.9240	0.9602	0.8983	0.9269		0.9009	0.9324		0.8188	0.9237	0.8916	0.8343	
$Slope_{pred}$	0.8917	1.0202	0.8517	0.8980	0.9605	1.0215		0.9387	0.9248		0.7892	0.9050	0.8823	0.8410	
$Offset_{pred}$	2.2868	-0.4596	3.0623	2.0120	0.7911	-0.5781		1.3020	1.7118		4.4092	1.6586	2.1392	3.1677	
RMSEP (%)	0.6956	0.5422	0.7346	0.5112	0.6084	0.7240		0.7370	0.6761		1.1098	0.6844	1.0423	0.9891	
$Bias_{pred}$ (%)	0.2108	-0.0768	0.2192	0.0759	-0.0149	-0.1454		0.0827	0.1935		0.3677	-0.1456	-0.2410	0.0346	
SEP (%)	0.6711	0.5450	0.7098	0.5134	0.6152	0.7209		0.7410	0.6589		1.0602	0.6791	1.0555	1.0367	
RPD_{pred}	3.7004	4.4994	3.4988	4.7766	3.0710	3.4769		3.1520	3.8437		2.3424	3.6111	3.0364	2.4562	
RER_{pred}	14.1711	16.5364	13.3988	17.5553	11.4736	13.9401		13.9293	16.1333		8.9706	13.2717	9.9381	7.7216	
All Samples in Calibration Set															
													F-Haaland's	7	6
													R^2_{calib}	0.9559	0.9446
													RMSEC (%)	0.5849	0.6341
													R^2_{CV}	0.9014	0.8669
													RMSECV (%)	0.8929	1.0070
													RPD_{CV}	3.1200	2.6765
													RER_{CV}	11.8844	10.5176

Table F-28: Regression statistics for acid insoluble ash (AIA) content (% whole dry mass basis). All spectra were transformed by SG 2,2,25,25 prior to model development.

Dataset	DS		DT		DG		DH	DU		DV	WU		DF	
Varieties	All	Gig	All	Gig	All	Gig	Gig	All	Gig	Gig	All	Gig	All	Gig
PLS- λ 10 ³ nm	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5
Samples Excluded	18002 18008	18002			258 5125 18002	258 5125 18002		257 5125	258 5125		258 2530 5125 18008	258 2530 5125	18008	
Calib:Valid	172:39	136:31	158:39	126:31	136:43	104:31	35:0	136:44	106:29	31:0	160:39	138:31	36:13	31:11
F-Haaland's	15	15	14	14	15	15	5	11	14	14	14	13	5	9
R^2_{calib}	0.9650	0.9690	0.9479	0.9702	0.9712	0.9753	0.8413	0.9123	0.9496	0.9971	0.9350	0.8882	0.8940	0.9713
$Offset_{calib}$	0.0489	0.0386	0.0689	0.0350	0.0390	0.0308	0.2224	0.1181	0.0619	0.0042	0.0636	0.1413	0.1224	0.0371
RMSEC (%)	0.2240	0.1942	0.2425	0.1549	0.2124	0.1880	0.3100	0.3806	0.2663	0.0325	0.4280	0.3701	0.3130	0.2196
R^2_{CV}	0.9353	0.9305	0.8994	0.9388	0.9348	0.9319	0.6493	0.8695	0.8801	0.9371	0.8979	0.7747	0.7774	0.8693
RMSECV (%)	0.3050	0.2914	0.3370	0.2229	0.3200	0.3131	0.4644	0.4647	0.4140	0.1528	0.5809	0.5313	0.4580	0.4717
$BIAS_{CV}$ (%)	0.0021	0.0007	-0.0019	0.0034	0.0084	-0.0041	0.0048	0.0012	0.0083	0.0010	0.0018	0.0048	-0.0026	0.0419
SECV (%)	0.3058	0.2925	0.3381	0.2237	0.3211	0.3146	0.4712	0.4665	0.4159	0.1553	0.5827	0.5332	0.4645	0.4776
RPD_{CV}	3.9242	3.7879	3.1519	4.0258	3.9151	3.8230	1.6755	2.7660	2.8664	3.9785	3.0975	2.0832	2.0990	2.7572
RER_{CV}	25.4511	20.6343	23.0226	21.9218	24.5502	18.8683	8.2005	16.8980	14.5113	15.0361	30.6484	11.3176	7.7043	11.5281
R^2_{pred}	0.9227	0.9501	0.8871	0.9389	0.9269	0.8920		0.8574	0.7586		0.6141	0.8204	0.8385	0.8025
$Slope_{pred}$	1.0306	0.9925	0.9119	0.9838	1.0006	0.9496		0.9968	0.9207		0.8561	1.0949	0.6119	0.8788
$Offset_{pred}$	-0.0494	0.0356	0.0678	0.0022	-0.0136	0.1223		-0.0419	0.2241		0.5562	-0.2060	0.2854	0.2174
RMSEP (%)	0.2406	0.1819	0.2710	0.1995	0.2782	0.2329		0.3966	0.5089		0.6865	0.4244	0.8928	0.5915
$Bias_{pred}$ (%)	-0.0169	0.0273	-0.0257	-0.0156	-0.0129	0.0777		-0.0458	0.1450		0.4035	-0.1021	-0.3198	0.0746
SEP (%)	0.2431	0.1828	0.2733	0.2022	0.2812	0.2232		0.3985	0.4965		0.5626	0.4187	0.8676	0.6154
RPD_{pred}	3.3350	4.3962	2.9675	3.9756	3.5598	2.9921		2.4596	1.9033		1.4414	1.9195	2.1189	2.2097
RER_{pred}	13.3088	17.1633	11.8424	15.5211	16.5796	12.6582		9.5970	6.3135		5.7520	7.4938	6.2356	7.7419
All Samples in Calibration Set														
												F-Haaland's	6	7
												R^2_{calib}	0.9155	0.9291
												RMSEC (%)	0.3606	0.3452
												R^2_{CV}	0.8397	0.8234
												RMSECV (%)	0.5070	0.5582
												RPD_{CV}	2.4469	2.3230
												RER_{CV}	10.8073	9.7984

Table F-29: Regression statistics for the carbon and nitrogen contents (% whole dry mass). Spectra transformed by SG 2,2,25,25 prior to model development.

Dataset	Carbon Content					Nitrogen Content						
	DS	DT	DG	DU	DF	DS	DT	DG	DU	DF	WU	WU _{pred}
Varieties	All	All	All	All	All	All	All	All	All	All	All	All
PLS- λ 10 ³ nm	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5
Samples Excluded						5034		5034	5034		5034	2530
Calib:Valid	45:11	35:11	45:11	44:11	34:10	44:11	35:11	44:11	43:11	34:10	43:11	207:50
F-Haaland's	11	6	11	8	9	6	6	7	4	9	5	12
R^2_{calib}	0.9377	0.8505	0.9013	0.8231	0.9480	0.9570	0.9723	0.9200	0.8476	0.9913	0.8659	0.9594
$Offset_{calib}$	2.9748	7.1350	4.7113	8.4395	2.4641	0.0326	0.0213	0.0609	0.1097	0.0081	0.1014	0.0226
RMSEC (%)	0.2166	0.3600	0.2721	0.3674	0.2707	0.1034	0.0890	0.1407	0.1817	0.0502	0.1836	0.0753
R^2_{CV}	0.7533	0.5610	0.5767	0.5525	0.7466	0.8991	0.9252	0.7932	0.6916	0.9622	0.6725	0.9399
RMSECV (%)	0.4355	0.6237	0.5736	0.5972	0.6095	0.1585	0.1462	0.2268	0.2593	0.1061	0.2882	0.0917
$BIAS_{CV}$ (%)	-0.0170	0.0167	0.0490	-0.0002	-0.0073	0.0013	0.0031	-0.0141	0.0056	0.0040	-0.0187	0.0000
SECV (%)	0.4401	0.6326	0.5779	0.6041	0.6186	0.1603	0.1483	0.2290	0.2624	0.1076	0.2910	0.0919
RPD_{CV}	1.9935	1.4933	1.5154	1.4627	1.9486	3.1470	3.6551	2.1972	1.7947	5.0731	1.7439	4.0765
RER_{CV}	9.0768	6.3138	6.9111	6.6117	7.4102	12.3931	13.3950	8.6781	7.5739	16.9453	6.8289	22.1536
R^2_{pred}	0.7336	0.6699	0.7799	0.7869	0.8321	0.9898	0.9878	0.9797	0.8739	0.9163	0.8919	0.9117
$Slope_{pred}$	1.0351	0.9941	0.9171	1.1151	1.0012	0.9017	0.9142	0.8783	0.6732	0.9767	0.6318	0.7957
$Offset_{pred}$	-1.8235	0.0576	3.8637	-5.6649	-0.0731	0.0663	0.1114	0.1232	0.1877	0.0423	0.2128	0.0548
RMSEP (%)	0.4729	0.5565	0.3855	0.4641	0.3701	0.0873	0.1014	0.1211	0.2822	0.1209	0.2813	0.1974
$Bias_{pred}$ (%)	-0.1299	-0.2273	-0.1387	-0.1269	-0.0178	-0.0005	0.0532	0.0405	-0.0384	0.0273	-0.0374	-0.0906
SEP (%)	0.4769	0.5328	0.3772	0.4682	0.3896	0.0915	0.0905	0.1197	0.2932	0.1241	0.2924	0.1771
RPD_{pred}	1.6007	1.4329	2.0237	1.6903	2.2235	7.4495	7.5291	5.6961	2.4098	3.3771	2.3316	3.1150
RER_{pred}	5.0946	4.5607	6.4408	5.1898	6.6257	21.8540	22.0875	16.7103	6.8197	11.8100	6.8400	11.5448
All Samples in Calibration Set												
F-Haaland's	10	6	10	9	7	6	6	6	4	11	8	14
R^2_{calib}	0.9289	0.8677	0.8732	0.8408	0.9086	0.9634	0.9742	0.9299	0.8681	0.9936	0.9277	0.9691
RMSEC (%)	0.2314	0.3327	0.3085	0.3454	0.3393	0.1020	0.0908	0.1410	0.1869	0.0417	0.1441	0.0740
R^2_{CV}	0.7832	0.7324	0.6662	0.7062	0.7966	0.9331	0.9530	0.8759	0.8032	0.9683	0.8343	0.9551
RMSECV (%)	0.3990	0.4805	0.5181	0.4847	0.5366	0.1385	0.1235	0.1828	0.2366	0.0957	0.2121	0.0900
RPD_{CV}	2.1749	1.9060	1.6721	1.7865	2.0942	3.8513	4.5782	2.9139	2.1762	5.4608	2.5292	4.6751
RER_{CV}	10.3469	8.5826	7.9670	8.5170	8.4532	14.6299	16.3755	11.0901	8.5639	18.8378	9.5596	23.9783

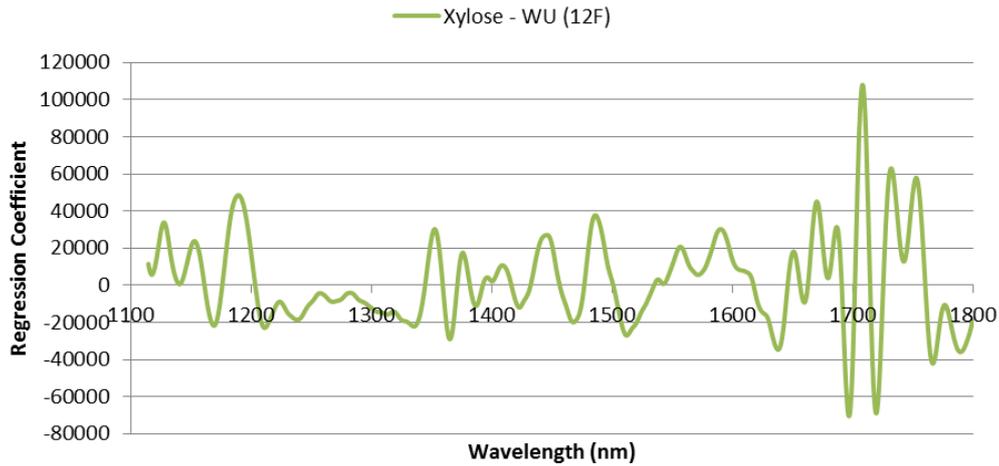


Figure F-9: Regression coefficients plot for the 12-factor PLS model for the xylose content of WU samples. The spectra were pretreated with SG-2,2,25,25 prior to model development.

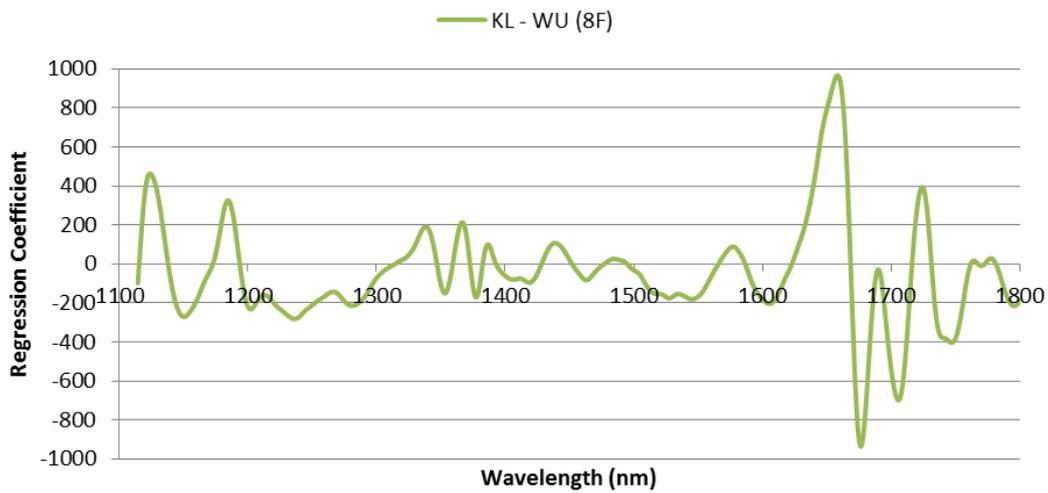


Figure F-10: Regression coefficients plot for an 8-factor PLS model for the Klason lignin (KL) content of WU samples. All spectra were transformed by SG-1,1,10,10 prior to model development.

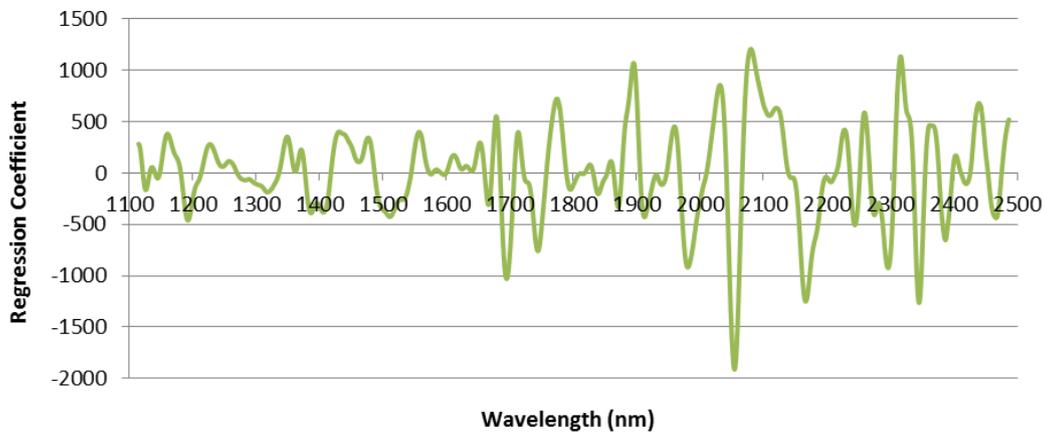


Figure F-11: Regression coefficients plot for the DT PLS model (6 factors) for nitrogen content. This model includes all samples in the calibration and was made on spectra pretreated with SG-2,2,25,25.

Table F-30: Regression statistics for the prediction of the glucose (GLU_SRS), total sugars (TOT_SRS), xylose (XYL_SRS), rhamnose (RHA_SRS), galactose (GAL_SRS), arabinose (ARA_SRS), and mannose (MAN_SRS) contents of the DF samples using the models developed on the DS and DT spectra. In the first case the number of PLS factors determined for the DS/DT model using Haaland's criterion is used; in the second the number that give the maximum R^2_{pred} is used.

Constituent	GLU_SRS		TOT_SRS		XYL_SRS		RHA_SRS		GAL_SRS		ARA_SRS		MAN_SRS	
	DS	DT	DS	DT	DS	DT	DS	DT	DS	DT	DS	DT	DS	DT
Pretreat.	SG	SG	SG	SG	SG	SG	SG	SG	SG	SG	SG	SG	SG	SG
Specific	1,1,10,10	1,1,10,10	1,1,10,10	1,1,10,10	2,2,25,25	2,2,25,25	2,2,25,25	2,2,25,25	2,2,25,25	2,2,25,25	2,2,25,25	2,2,25,25	2,2,25,25	2,2,25,25
Valid Sample	52	52	51	51	51	51	51	51	51	51	51	51	51	51
Number of PLS Factors For the DS/DT Model Determined by Haaland's Criterion														
F-Model	16	18	14	13	16	14	13	7	18	13	12	11	10	5
R^2_{pred}	0.9557	0.9139	0.9481	0.8934	0.8850	0.8597	0.4985	0.6374	0.7092	0.7969	0.9025	0.8491	0.6790	0.6587
$Slope_{pred}$	0.7714	0.6246	0.7762	0.5963	0.7611	0.8398	0.5014	0.7499	0.8957	0.6705	0.8048	0.6293	0.7150	0.4873
$Offset_{pred}$	10.3464	15.5160	15.1680	25.3435	4.8319	3.0363	0.1235	0.0125	-0.1042	0.2841	0.3214	1.0311	0.1295	0.2864
RMSEP (%)	2.2117	2.5147	2.0512	2.2465	0.6281	0.5740	0.0679	0.0593	0.2619	0.1444	0.3368	0.4337	0.1076	0.1901
$Bias_{pred}$	1.6608	1.2562	1.5544	0.7851	0.2910	-0.0080	0.0360	-0.0314	-0.1921	0.0065	-0.1548	0.1269	0.0613	0.1636
SEP (%)	1.4749	2.1997	1.3517	2.1257	0.5622	0.5796	0.0581	0.0507	0.1798	0.1457	0.3021	0.4188	0.0893	0.0977
RPD_{pred}	3.5382	2.3723	3.4700	2.2065	2.7485	2.6657	1.4121	1.6170	1.7151	2.1169	3.0417	2.1938	1.7597	1.6097
RER_{pred}	11.9577	8.0175	13.3770	8.5063	13.7340	13.3203	5.3848	6.1661	5.9704	7.3688	9.0248	6.5090	7.6924	7.0369
Optimum Number of PLS Factors in Model for Maximum R^2_{pred}														
F-Model	15	9	18	20	15	16	17	14	4	17	11	14	11	16
R^2_{pred}	0.9585	0.9404	0.9625	0.9581	0.9050	0.8904	0.7480	0.7903	0.7419	0.8649	0.9069	0.8691	0.6953	0.7746
$Slope_{pred}$	0.7691	0.6596	0.8010	0.8090	0.7879	0.8103	0.6686	0.6705	0.7060	0.6233	0.8496	0.6623	0.6947	0.6774
$Offset_{pred}$	10.3481	13.5436	13.5792	12.7083	4.2066	3.6805	0.0697	0.0472	0.3420	0.2880	0.3034	0.7334	0.0688	0.0775
RMSEP (%)	2.1444	2.0513	1.8866	1.6078	0.5371	0.5283	0.0430	0.0402	0.1820	0.1406	0.2900	0.3965	0.0860	0.0759
$Bias_{pred}$	1.5781	0.6142	1.4717	1.0867	0.1749	0.0761	0.0115	-0.0107	0.0942	-0.0295	-0.0634	-0.0902	-0.0043	0.0003
SEP (%)	1.4660	1.9763	1.1921	1.1967	0.5129	0.5280	0.0419	0.0392	0.1572	0.1388	0.2858	0.3899	0.0868	0.0766
RPD_{pred}	3.5597	2.6405	3.9346	3.9194	3.0125	2.9262	1.9597	2.0949	1.9617	2.2217	3.2152	2.3564	1.8115	2.0516
RER_{pred}	12.0304	8.9240	15.1683	15.1096	15.0531	14.6219	7.4727	7.9881	6.8287	7.7339	9.5393	6.9913	7.9190	8.9687

Table F-31: Regression statistics for the prediction of the Klason lignin (KL), acid soluble lignin (ASL), acid insoluble residue (AIR), ash, acid insoluble ash (AIA), 95% ethanol soluble extractives (EXTR), and nitrogen contents of the DF samples using the models developed on the DS and DT spectra. In the first case the number of PLS factors determined for the DS/DT model using Haaland's criterion is used; in the second the number that give the maximum R^2_{pred} is used.

Constituent	KL		ASL		AIR		ASH		AIA		EXTR		NITROGEN	
	DS	DT	DS	DT	DS	DT	DS	DT	DS	DT	DS	DT	DS	DT
Pretreat.	SG	SG	SG	SG	SG	SG	SG	SG	SG	SG	SG	SG	SG	SG
Specific	1,1,10,10	1,1,10,10	2,2,25,25	2,2,25,25	1,1,10,10	1,1,10,10	2,2,25,25	2,2,25,25	2,2,25,25	2,2,25,25	2,2,25,25	2,2,25,25	2,2,25,25	2,2,25,25
Valid Sample	49	49	50	50	49	49	58	58	49	49	58	58	47	47
Number of PLS Factors For the DS/DT Model Determined by Haaland's Criterion														
F-Model	13	7	10	7	17	12	13	14	15	14	15	14	6	6
R^2_{pred}	0.9576	0.9757	0.9239	0.9168	0.9492	0.9257	0.9140	0.8936	0.9397	0.9110	0.9304	0.8713	0.9459	0.9017
$Slope_{pred}$	0.7568	0.7167	0.7999	0.7342	0.8199	0.6311	0.7654	0.6385	0.9019	0.6887	0.7020	0.6963	0.6036	0.5916
$Offset_{pred}$	4.8435	5.8184	0.4501	0.6963	4.5656	8.3365	0.7744	1.2648	0.3014	0.7343	0.6667	1.4105	0.3711	0.5342
RMSEP (%)	0.8940	1.0214	0.3382	0.3849	1.2012	1.4797	0.7555	0.9581	0.3559	0.5805	2.1741	1.7666	0.2057	0.2889
$Bias_{pred}$ (%)	0.2756	0.4963	-0.0368	0.0433	0.9551	0.9411	-0.1526	-0.1638	0.1776	0.3414	-1.8146	-1.1183	0.0377	0.1908
SEP (%)	0.8593	0.9019	0.3396	0.3863	0.7360	1.1536	0.7464	0.9522	0.3116	0.4743	1.2080	1.3796	0.2044	0.2193
RPD_{pred}	3.4402	3.2777	3.2834	2.8921	3.8232	2.4391	3.0133	2.3619	4.0215	2.6422	2.8212	2.4702	2.3702	2.2091
RER_{pred}	12.6925	12.0930	11.4602	10.0755	14.5628	9.2905	11.8379	9.2791	17.7619	11.6701	10.9977	9.6296	8.2827	7.7197
Optimum Number of PLS Factors in Model for Maximum R^2_{pred}														
F-Model	6	8	14	5	16	15	20	13	19	15	18	17	10	8
R^2_{pred}	0.9698	0.9780	0.9331	0.9239	0.9523	0.9455	0.9460	0.9217	0.9477	0.9346	0.9406	0.8980	0.9689	0.9258
$Slope_{pred}$	0.8601	0.7280	0.6933	0.7497	0.7982	0.7139	0.7778	0.6687	0.8770	0.7068	0.7203	0.7327	0.7264	0.6474
$Offset_{pred}$	2.8507	5.7957	0.7427	0.4589	4.9871	7.1746	1.0851	1.0723	0.4654	0.7210	0.4678	0.2892	0.3065	0.5410
RMSEP (%)	0.6441	1.0985	0.3957	0.3970	1.2043	1.7131	0.6783	0.8893	0.4300	0.5561	2.1752	2.2935	0.1641	0.3099
$Bias_{pred}$ (%)	0.2239	0.6862	-0.0035	-0.1560	0.9417	1.4394	0.2069	-0.2371	0.3102	0.3509	-1.8613	-1.9362	0.0764	0.2445
SEP (%)	0.6102	0.8666	0.3997	0.3686	0.7585	0.9385	0.6516	0.8646	0.3008	0.4358	1.1354	1.2400	0.1468	0.1925
RPD_{pred}	4.8445	3.4111	2.7894	3.0303	3.7099	2.9982	3.4514	2.6012	4.1664	2.8756	3.0014	2.7483	3.3004	2.5164
RER_{pred}	17.8739	12.5852	9.7360	10.5570	14.1313	11.4203	13.5590	10.2188	18.4017	12.7009	11.7001	10.7134	11.5331	8.7933

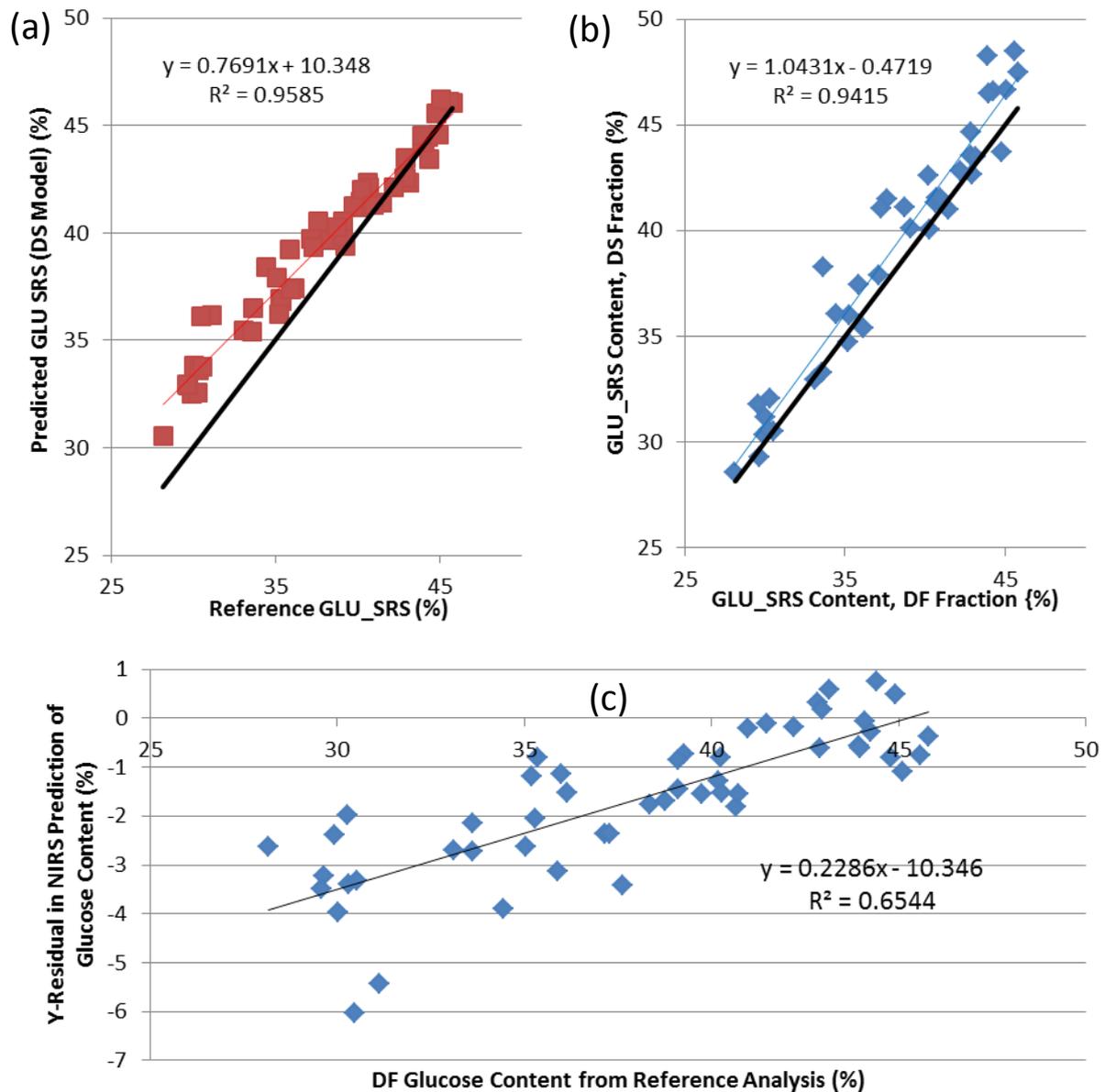


Figure F-12: Plots comparing DS and DF data for glucose content (% whole dry mass basis). (a) The reference glucose content for the DF samples vs. the predicted content using the GLU_SRS model based on the DS scans, the black line represents a 1:1 relationship; (b) glucose content of the DF samples vs. glucose content of the DS samples; (c) the y-residual (actual DF minus predicted DF content) in NIRS prediction of glucose content plotted according to the extractives content of the DF sample.

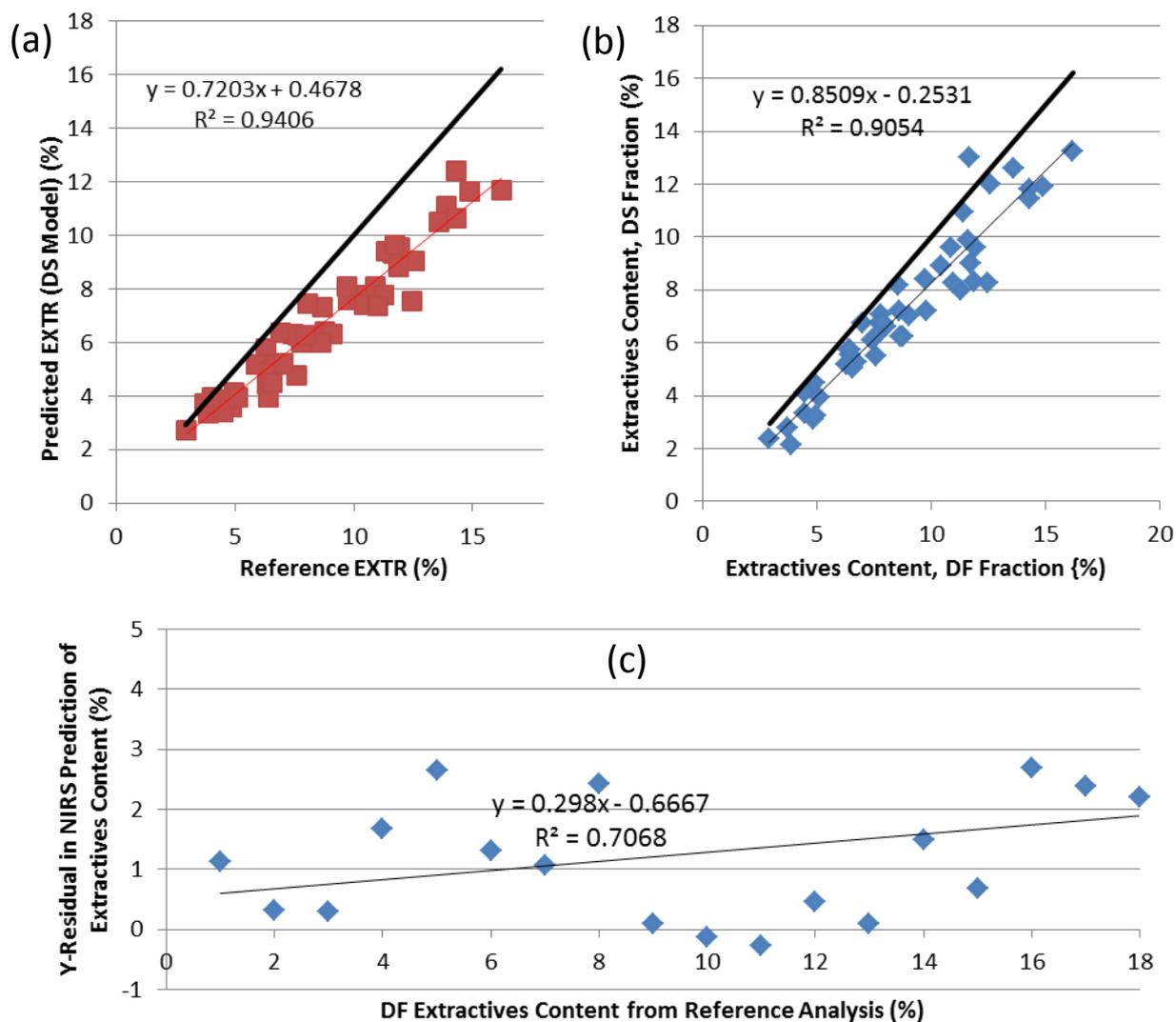


Figure F-13: Plots comparing DS and DF data for extractives (EXTR_PD) content (% whole dry mass basis). (a) Reference extractives content for the DF samples vs. the predicted content using the EXTR_PD model based on the DS scans, the black line represents a 1:1 relationship; (b) extractives content of the DF samples vs. extractives content of the DS samples; (c) the y-residual (actual DF minus predicted DF content) in NIRS prediction of extractives content plotted according to the extractives content of the DF sample.

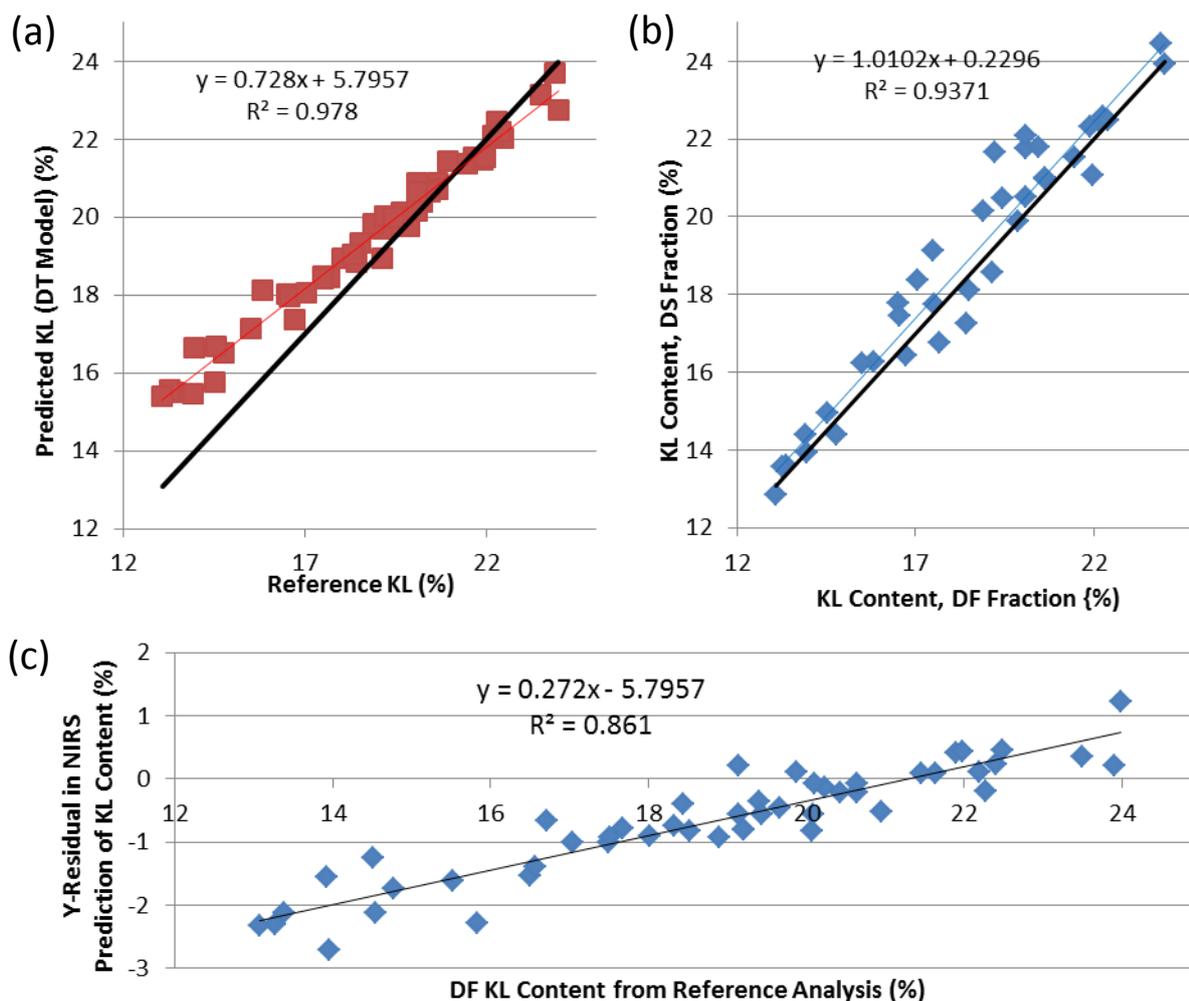


Figure F-14: Plots comparing DS and DF data for Klason lignin (KL) content (% whole dry mass basis). (a) Reference KL content for DF samples vs. predicted content using the KL model based on the DT scans; (b) KL content of the DF samples vs. KL content of the DS samples; (c) the y-residual (actual DF minus predicted DF content) in NIRS prediction of KL content plotted according to the DF KL content.

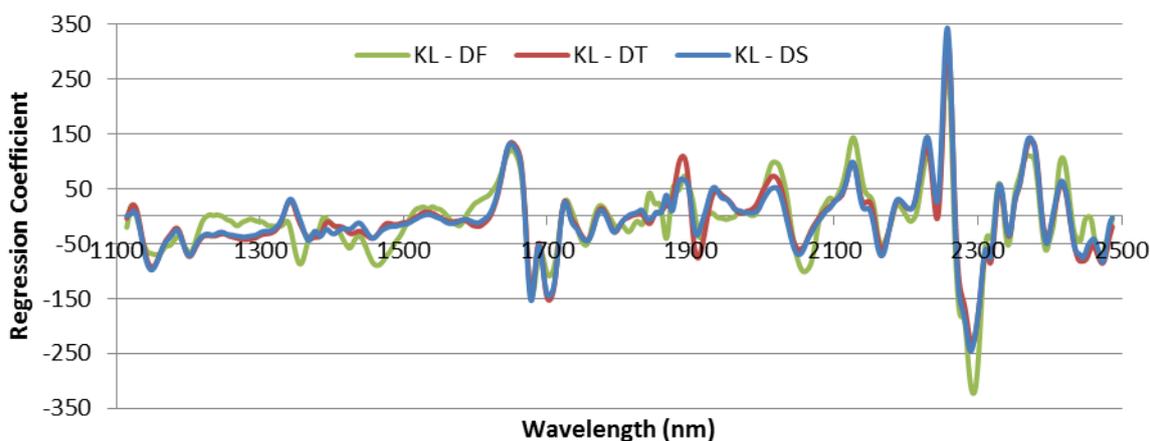


Figure F-15: Regression coefficients plot for DF, DT, and DS Klason lignin (KL) models. Each model has 6 PLS factors responsible for the regression coefficients here.

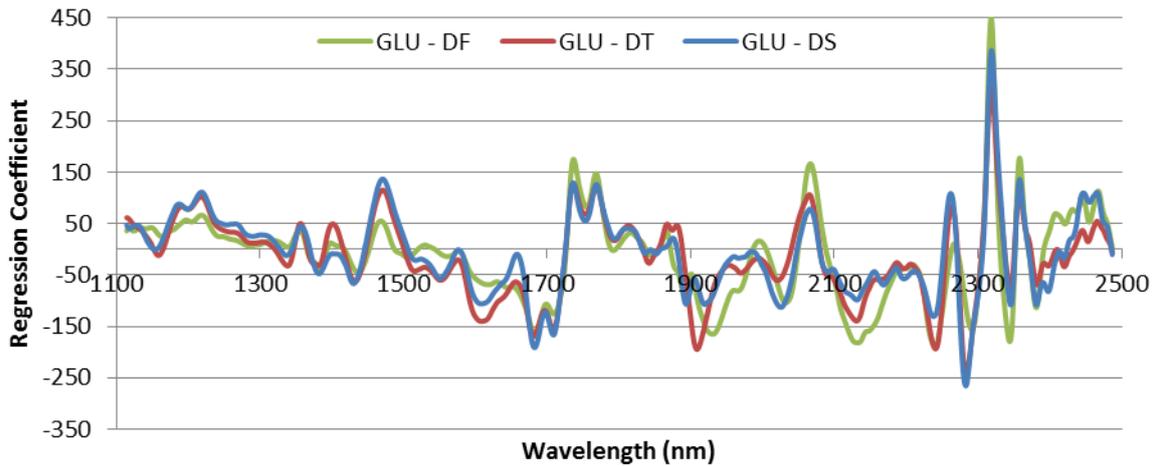


Figure F-16: Regression coefficients plot for DF, DT, and DS glucose (GLU) models. Each model has 5 PLS factors responsible for the regression coefficients here.

Table F-32: Comparison of PLSR regression statistics for WU models using only the DS data (the standard WU model) and WU models using the weighted average of the constituent values for the DS and DF fractions (WU_{DG}). The same samples are in the calibration set of both types of model for each constituent. The models are based on all *Miscanthus* varieties

Constit.	GLU_SRS		XYL_SRS		KL		ASL	
	WU	WU _{DG}	WU	WU _{DG}	WU	WU _{DG}	WU	WU _{DG}
Pretreat.	SG	SG	SG	SG	SG	SG	SG	SG
Specific	1,1,10,10	1,1,10,10	2,2,25,25	2,2,25,25	1,1,10,10	1,1,10,10	2,2,25,25	2,2,25,25
PLS- λ 10 ³ nm	1.1-1.8	1.1-1.8	1.1-1.8	1.1-1.8	1.1-1.8	1.1-1.8	1.1-1.8	1.1-1.8
Sam. Excl.	5125	5125	5125	5125	5125	5125	5125	5125
Calib:Valid	36:0	36:0	36:0	36:0	35:0	35:0	36:0	36:0
F-Haalands	4	4	8	8	4	4	8	10
R^2_{calib}	0.9341	0.9395	0.9163	0.9402	0.9031	0.9145	0.9759	0.9897
$Offset_{calib}$	2.5773	2.3348	1.6015	1.1384	1.8369	1.6076	0.0632	0.0267
RMSEC (%)	1.4813	1.3739	0.4817	0.3862	0.9933	0.9127	0.2065	0.1299
R^2_{CV}	0.8927	0.8973	0.5874	0.6801	0.8102	0.8300	0.8560	0.8759
RMSECV (%)	1.8901	1.7904	1.1090	0.9223	1.4089	1.3075	0.5081	0.4533
$BIAS_{CV}$ (%)	-0.0052	0.0000	-0.0834	-0.0803	-0.0364	-0.0389	0.0100	0.0046
SECV (%)	1.9169	1.8157	1.1215	0.9318	1.4290	1.3260	0.5152	0.4597
RPD_{CV}	3.0523	3.1202	1.5054	1.7185	2.2660	2.3889	2.6185	2.8208
RER_{CV}	10.4017	10.3958	5.9234	6.8905	8.1024	8.4758	8.4650	9.0493

Table F-33: Comparisons of the standard WU models and the WU_{DGP} models for predicting the compositions of the WU_{DG} samples. These models include samples of several *Miscanthus* varieties. Predictions for glucose (GLU_SRS), xylose (XYL_SRS), rhamnose (RHA_SRS), galactose (GAL_SRS), arabinose (ARA_SRS), and mannose (MAN_SRS) are presented.

Dataset	GLU_SRS		TOT_SRS		XYL_SRS		RHA_SRS		GAL_SRS		ARA_SRS		MAN_SRS	
	WU	WU _{DGP}												
Pretreat.	SG	SG												
Specific	1,1,10,10	1,1,10,10	1,1,10,10	1,1,10,10	2,2,25,25	2,2,25,25	2,2,25,25	2,2,25,25	2,2,25,25	2,2,25,25	2,2,25,25	2,2,25,25	2,2,25,25	2,2,25,25
PLS- λ 10 ³ nm	1.1-1.8	1.1-1.8	1.1-1.8	1.1-1.8	1.1-1.8	1.1-1.8	1.1-1.8	1.1-1.8	1.1-1.8	1.1-1.8	1.1-1.8	1.1-1.8	1.1-1.8	1.1-1.8
Sam. Excl.	5125	5125	5125	5125	5125	5125	5125	5125	5125	5125	5125	5125	5125	5125
Calib:Valid	109:35	109:35	109:35	109:35	109:35	109:35	109:35	109:35	109:35	109:35	109:35	109:35	109:35	109:35
F-Haaland's	11	11	16	16	14	14	16	18	8	9	13	9	8	11
R^2_{calib}	0.9614	0.9610	0.9516	0.9590	0.9131	0.9217	0.9066	0.9410	0.8804	0.9037	0.9433	0.9314	0.7615	0.8590
$Offset_{calib}$	1.5247	1.5149	3.0675	2.5675	1.7440	1.5519	0.0157	0.0103	0.0949	0.0785	0.1410	0.1736	0.0479	0.0304
RMSEC (%)	0.9600	0.9298	1.1192	0.9914	0.5682	0.5039	0.0330	0.0240	0.1008	0.0890	0.1974	0.2097	0.0640	0.0451
R^2_{CV}	0.9409	0.9394	0.9079	0.9227	0.8451	0.8764	0.7816	0.8403	0.8166	0.8577	0.8994	0.9001	0.6654	0.7737
RMSECV (%)	1.1894	1.1601	1.5534	1.3709	0.7633	0.6359	0.0513	0.0403	0.1252	0.1084	0.2639	0.2539	0.0761	0.0574
$BIAS_{CV}$ (%)	-0.0123	0.0152	-0.0137	-0.0201	-0.0148	0.0097	-0.0031	0.0000	-0.0023	-0.0024	0.0035	-0.0002	0.0009	-0.0019
SECV (%)	1.1948	1.1654	1.5605	1.3771	0.7667	0.6387	0.0515	0.0405	0.1257	0.1089	0.2651	0.2551	0.0765	0.0576
RPD_{CV}	4.1093	4.0601	3.2766	3.5742	2.5256	2.8327	2.1105	2.4534	2.3284	2.6464	3.1403	3.1546	1.7212	2.0918
RER_{CV}	16.7751	16.5637	14.7265	15.4740	11.6801	12.4949	11.6334	13.5662	9.1188	10.7840	11.1473	11.7382	8.1631	10.1162
R^2_{pred}	0.9529	0.9572	0.9131	0.9306	0.8548	0.8582	0.7953	0.8715	0.8167	0.8506	0.9194	0.9391	0.7311	0.7593
$Slope_{pred}$	1.0737	1.0438	1.0680	1.0503	0.8757	0.8406	0.7350	0.7013	0.9682	1.0002	1.0040	0.9597	0.8404	0.9093
$Offset_{pred}$	-2.3305	-1.8938	-3.7168	-3.3847	2.4825	3.0008	0.0338	0.0461	-0.0084	-0.0036	-0.0922	0.0449	0.0231	0.0196
RMSEP (%)	1.4777	1.2572	1.8369	1.5696	0.6221	0.6043	0.0602	0.0521	0.1365	0.1208	0.2785	0.2266	0.0568	0.0548
$Bias_{pred}$	0.5297	-0.1925	0.4507	-0.3000	0.1181	-0.0323	-0.0146	-0.0084	-0.0330	-0.0034	-0.0829	-0.0480	-0.0080	0.0019
SEP (%)	1.3997	1.2605	1.8067	1.5631	0.6197	0.6122	0.0593	0.0521	0.1344	0.1225	0.2697	0.2247	0.0570	0.0555
RPD_{pred}	4.0009	4.4426	2.9721	3.4352	2.6199	2.6522	2.1862	2.4868	2.1748	2.3861	3.3627	4.0371	1.8725	1.9232
RER_{pred}	13.4858	14.9747	10.3298	11.9395	10.3606	10.4882	10.0534	11.4357	7.3833	8.1003	9.8624	11.8401	6.3981	6.5713

Table F-34: Further comparisons of the standard WU models and the WU_{DGP} models for predicting the compositions of the WU_{DG} samples. These models include samples of several *Miscanthus* varieties. Predictions for Klason lignin (KL), acid soluble lignin (ASL), acid insoluble residue (AIR), ash, acid insoluble ash (AIA), 95% ethanol-soluble-extractives (EXTR), and nitrogen content are presented.

Dataset	KL		ASL		AIR		ASH		AIA		EXTR		NITROGEN	
	WU	WU_{DGP}	WU	WU_{DGP}	WU	WU_{DGP}								
Pretreat.	SG	SG	SG	SG										
Specific	1,1,10,10	1,1,10,10	2,2,25,25	2,2,25,25	1,1,10,10	1,1,10,10	2,2,25,25	2,2,25,25	2,2,25,25	2,2,25,25	2,2,25,25	2,2,25,25	2,2,25,25	2,2,25,25
PLS- λ 10 ³ nm	1.1-1.8	1.1-1.8	1.1-1.8	1.1-1.8	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5
Sam. Excl.	2515 5125	2515 5125	5125	5125	10020 5125	5125 10020	5125	5125	5125	5125	5125	5125		
Calib:Valid	104:32	104:32	108:32	108:32	105:33	105:33	99:31	99:31	105:32	105:32	109:35	109:35	30:9	30:9
F-Haaland's	6	8	5	5	12	12	11	12	13	13	18	18	7	7
R^2_{calib}	0.9150	0.9352	0.8940	0.8958	0.9139	0.9111	0.9158	0.9365	0.8887	0.9079	0.9619	0.9598	0.9715	0.9645
$Offset_{calib}$	1.5544	1.1812	0.2597	0.2577	1.6914	1.7367	0.3389	0.2583	0.1495	0.1191	0.2427	0.2819	0.0183	0.0248
RMSEC (%)	0.7471	0.6204	0.3211	0.3103	0.6765	0.6638	0.6711	0.5588	0.4218	0.3648	0.5449	0.5783	0.0903	0.1006
R^2_{CV}	0.8928	0.9054	0.8725	0.8729	0.8138	0.8275	0.7987	0.8687	0.7400	0.7990	0.8118	0.8101	0.8926	0.8478
RMSECV (%)	0.8415	0.7508	0.3523	0.3428	1.0090	0.9297	1.0519	0.8054	0.6510	0.5428	1.2726	1.3209	0.1783	0.2102
$BIAS_{CV}$ (%)	-0.0040	-0.0030	-0.0019	-0.0027	-0.0126	-0.0152	-0.0424	-0.0487	-0.0103	-0.0069	0.0241	-0.0095	-0.0184	-0.0282
SECV (%)	0.8456	0.7544	0.3540	0.3444	1.0137	0.9340	1.0564	0.8081	0.6541	0.5454	1.2783	1.3269	0.1803	0.2119
RPD_{CV}	3.0459	3.2468	2.7994	2.8043	2.2850	2.3944	2.2004	2.7572	1.9422	2.2150	2.1953	2.1834	3.0176	2.5630
RER_{CV}	11.8101	13.5906	11.8997	11.5531	9.9139	10.4219	11.1430	13.3897	12.0512	13.1441	10.3151	9.8300	11.0184	8.6695
R^2_{pred}	0.8896	0.9290	0.8962	0.9001	0.8795	0.8873	0.8651	0.8808	0.7962	0.8060	0.8338	0.8511	0.7791	0.7850
$Slope_{pred}$	0.9201	0.9288	0.8521	0.8379	0.8514	0.8198	1.0492	1.0602	1.0471	1.0121	0.9562	1.0084	1.1390	1.1147
$Offset_{pred}$	1.4650	1.2844	0.2458	0.3048	2.7337	3.2932	-0.1763	-0.1319	-0.1411	-0.1086	0.1099	0.2625	-0.2533	-0.1654
RMSEP (%)	1.0830	0.8656	0.4510	0.4413	1.0212	1.0279	0.9477	0.9023	0.6215	0.5825	1.2267	1.2313	0.2828	0.2501
$Bias_{pred}$	-0.0471	-0.0616	-0.1370	-0.1148	-0.2488	-0.3231	0.0182	0.1064	-0.0869	-0.0947	-0.2118	0.3240	-0.1232	-0.0581
SEP (%)	1.0993	0.8772	0.4366	0.4330	1.0058	0.9910	0.9632	0.9108	0.6253	0.5839	1.2260	1.2052	0.2700	0.2580
RPD_{pred}	2.9960	3.7542	3.0722	3.0979	2.8702	2.9131	2.3966	2.5345	1.8803	2.0135	2.3302	2.3703	1.6073	1.6821
RER_{pred}	10.2244	12.8119	9.5279	9.6075	10.6384	10.7973	9.1031	9.6269	7.6370	8.1779	9.5482	9.7124	4.5887	4.8024

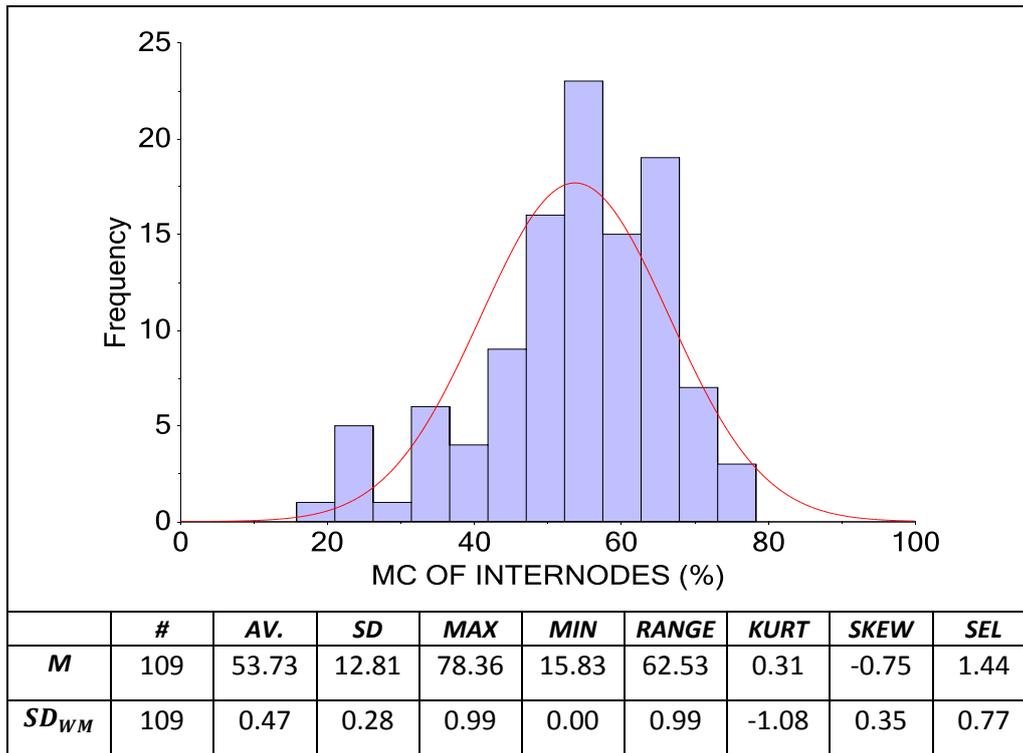


Figure F-17: A histogram, with associated statistics, for the wet-basis moisture content (MC) of the internode samples in the global set used for calibration set and validation set sample selection.

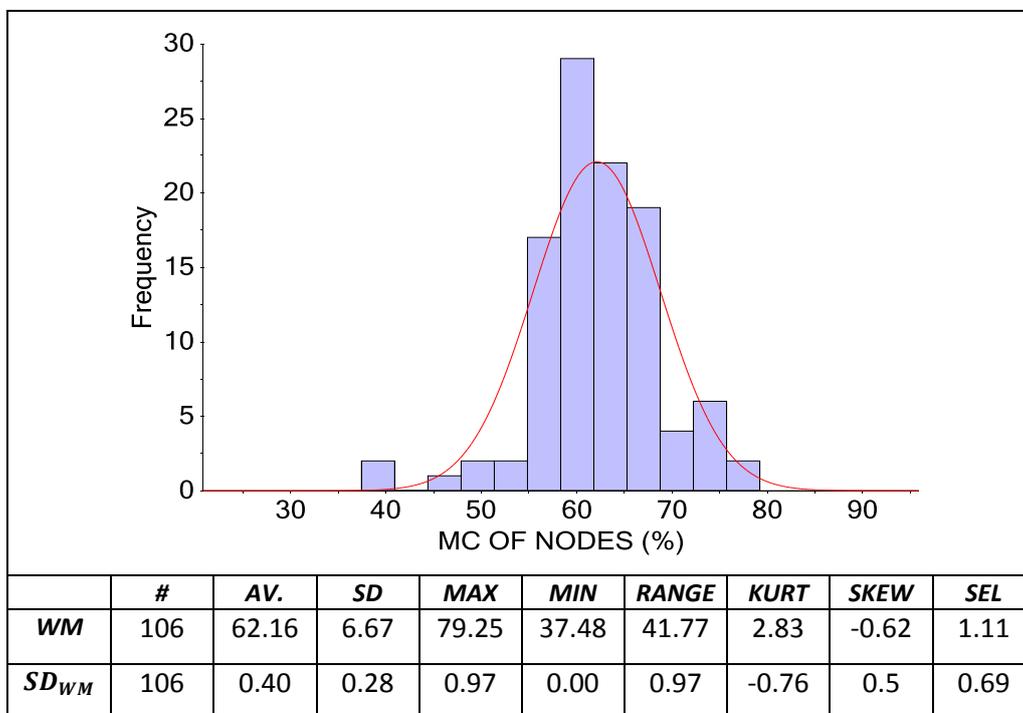


Figure F-18: A histogram, with associated statistics, for the wet-basis moisture content (MC) of the node samples in the global set used for calibration set and validation set sample selection.

Table F-35: Regression statistics for PLSR models for the moisture content (% wet basis) of internodes, nodes, and a set comprising leaf and stem samples.

Model	Internodes Model	Nodes Model	Leaves and Stems
Pretreat.	SG	SNVDT	SG
Specific	1,1,10,10	2,1.1-2.5	1,1,10,10
PLS- λ 10 ³ nm	1.1-2.5	1.1-2.5	1.1-2.5
Samples Excluded	195,158		12024
Calib:Valid	80:27	80:26	64:21
F-Haaland's	6	9	5
R^2_{calib}	0.9742	0.9691	0.9911
$Offset_{calib}$	1.4142	1.9003	0.3930
RMSEC (%)	1.8802	1.1990	1.8977
R^2_{CV}	0.9592	0.9505	0.9875
RMSECV (%)	2.3780	1.5293	2.2538
$BIAS_{CV}$ (%)	-0.0044	-0.0642	-0.0231
SECV (%)	2.3930	1.5376	2.2715
RPD_{CV}	4.9240	4.4658	8.9475
RER_{CV}	26.1319	27.1660	26.8099
R^2_{pred}	0.9529	0.9527	0.9836
$Slope_{pred}$	0.9812	1.0091	1.0101
$Offset_{pred}$	2.0246	-1.1632	0.2064
RMSEP (%)	3.4395	1.3896	2.5254
$Bias_{Pred}$	1.0582	-0.5808	0.6702
SEP (%)	3.3350	1.2874	2.4950
RPD_{pred}	4.5688	4.4422	7.6360
RER_{pred}	15.4643	18.0421	23.7959
All Samples in The Calibration Set			
F-Haaland's	5	12	6
R^2_{calib}	0.9682	0.9773	0.9916
RMSEC (%)	2.2648	1.0003	1.8099
R^2_{CV}	0.9619	0.9600	0.9889
RMSECV (%)	2.5283	1.3568	2.1617
RPD_{CV}	5.0226	4.8903	9.1631
RER_{CV}	24.6176	30.6413	28.0151

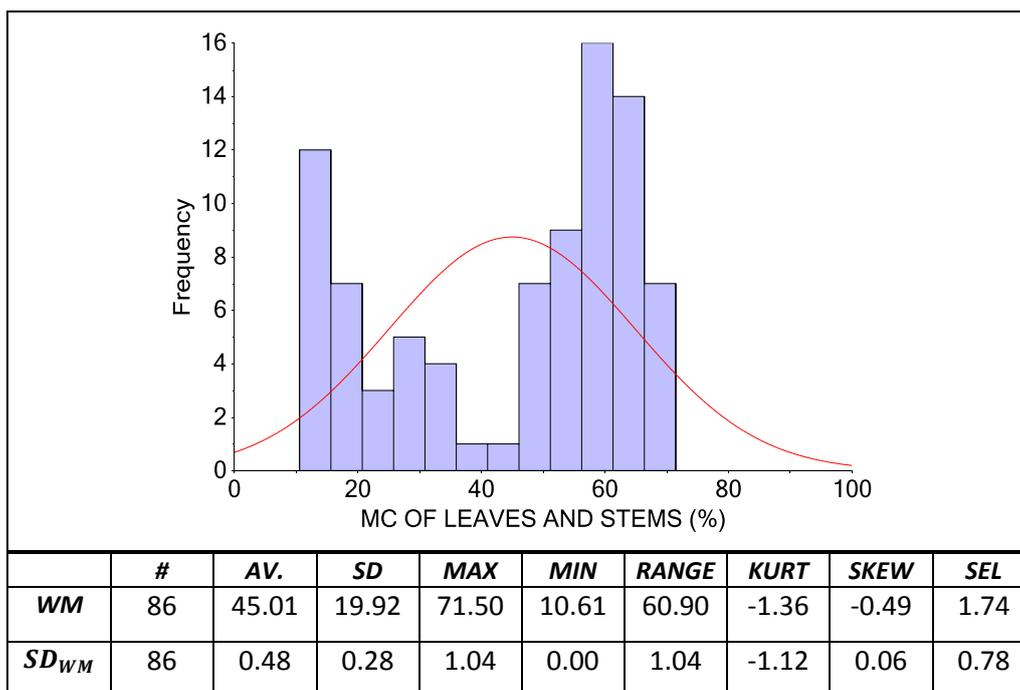


Figure F-19: A histogram, with associated statistics, for the wet-basis moisture content (MC) of the stem and leaf samples in the global set used for calibration set and validation set sample selection.

Table F-36: Summary statistics for the moisture contents (% wet basis) for the different sample sets.

Samples	#	Av (%)	SD (%)	Max (%)	Min (%)	Range (%)
DS Samples						
DS-E	207	7.59	1.04	10.62	5.87	4.74
DS-E Dishes	149	8.25	1.58	11.75	5.64	6.11
E-H (DS)	240	7.79	1.08	11.20	5.57	5.63
DF Samples						
DF-E	53	7.07	0.68	8.27	5.48	2.79
DF-E Dishes	39	7.19	0.85	9.02	5.40	3.62
E-H (DF)	48	7.10	0.80	8.94	5.55	3.39

Table F-37: Regression statistics for the chosen calibrations for moisture content (% wet basis) in the extractives/hydrolysis analysis. RMSEC and RMSECV values are in % (wet basis).

Samples	Pre.	λ (nm)	Calib:Valid	F	R^2_{cal}	RMSEC	RMSECV	R^2_{pred}	RMSEP	RPD_{pred}	RER_{pred}
DS Samples											
DS-E	SNVDT 2,1100- 2500nm	1100-2500	156:51	15	0.976	0.173	0.209	0.956	0.216	4.713	20.472
DS-E Dishes	MSC	1100-2500	112:37	3	0.965	0.286	0.298	0.974	0.282	6.29	20.79
E-H (DS)	MSC	1100-2500	180:60	6	0.959	0.216	0.227	0.942	0.2785	4.138	20.180
E-H (DS) [Cluster]	MSC	1100-2500	65:25	4	0.976	0.218	0.236	0.976	0.295	4.664	16.554
DF Samples											
DF-E	MSC	1100-2500	41:12	4	0.837	0.283	0.372	0.662	0.402	1.755	5.808
DF-E Dishes	MSC	1100-2500	30:9	3	0.903	0.305	0.313	0.829	0.230	2.43	7.26
E-H (DF)	MSC	1100-2500	36:12	2	0.873	0.294	0.310	0.875	0.304	2.67	8.700

Table F-38: Differences in predicted values for glucose, KL, and ash for replicate scans within the WU, DU, DV, DG, DH, DS, DT, and DF datasets. See Section 13.3.4 for description of terms.

Constituent	Dataset	Bias	Av. Abs. Diff	Av (%)	SD	Max	Max (%)	Min	Min (%)
GLU_SRS	WU	0.171	1.085	2.789%	0.678	3.975	10.381%	0.038	0.110%
	DU	0.192	0.876	2.213%	0.630	4.852	14.946%	0.046	0.124%
	DV	0.115	0.946	2.626%	0.650	2.808	7.211%	0.154	0.353%
	DG	0.092	0.375	0.934%	0.371	2.218	5.113%	0.003	0.007%
	DH	0.078	0.187	0.492%	0.216	1.248	4.614%	0.002	0.006%
	DS	0.304	0.409	1.029%	0.326	1.857	4.292%	0.001	0.003%
	DT	0.054	0.348	0.883%	0.259	1.207	3.255%	0.004	0.009%
	DF	0.078	0.187	0.492%	0.216	1.248	4.614%	0.002	0.006%
KL	WU	0.043	0.529	2.880%	0.371	2.387	10.198%	0.024	0.144%
	DU	-0.069	0.437	2.379%	0.297	1.593	7.603%	0.021	0.111%
	DV	0.061	0.422	2.552%	0.338	1.845	9.806%	0.036	0.189%
	DG	0.026	0.150	0.801%	0.153	0.995	4.933%	0.000	0.002%
	DH	-0.016	0.209	1.195%	0.178	0.914	4.574%	0.001	0.005%
	DS	0.047	0.165	0.886%	0.162	0.860	5.084%	0.000	0.001%
	DT	-0.047	0.130	0.695%	0.148	1.543	7.492%	0.000	0.002%
	DF	-0.064	0.088	0.495%	0.072	0.430	2.499%	0.000	0.000%
Ash	WU	0.076	0.775	25.015%	0.533	3.414	306.879%	0.030	0.878%
	DU	-0.130	0.517	17.590%	0.480	5.001	215.118%	0.027	0.752%
	DV	-0.113	0.403	10.541%	0.310	1.868	54.935%	0.036	0.920%
	DG	-0.012	0.188	6.898%	0.170	1.117	137.117%	0.000	0.002%
	DH	-0.045	0.195	4.792%	0.156	0.913	17.815%	0.000	0.009%
	DS	-0.105	0.166	5.267%	0.134	0.650	38.820%	0.002	0.040%
	DT	-0.032	0.147	5.031%	0.121	0.660	54.863%	0.003	0.054%
	DF	-0.079	0.102	3.124%	0.091	0.725	21.238%	0.001	0.019%

Table F-39: Differences in predicted glucose values (% whole dry mass), according to plant fraction, for replicate scans within the WU dataset. See Section 13.3.4 for description of terms.

Fraction	No.	Bias	Av. Abs. Diff	Av (%)	SD	Max	Max (%)	Min	Min (%)
Dead Leaf Blades	78	0.059	0.877	2.619%	0.548	2.710	8.532%	0.053	0.162%
Dead Leaf Sheaths	91	-0.057	1.078	2.621%	0.716	3.501	9.076%	0.094	0.224%
Live Leaf Blades	36	0.035	0.733	2.412%	0.540	2.171	7.333%	0.061	0.208%
Live Leaf Sheaths	13	0.252	1.212	3.066%	0.659	2.273	5.700%	0.258	0.699%
Internodes	188	0.177	1.039	2.400%	0.646	3.560	9.340%	0.079	0.194%
Nodes	178	-0.027	0.940	2.618%	0.585	3.587	10.780%	0.044	0.129%
Whole Plant	53	0.138	1.314	3.350%	0.793	3.975	10.381%	0.038	0.110%

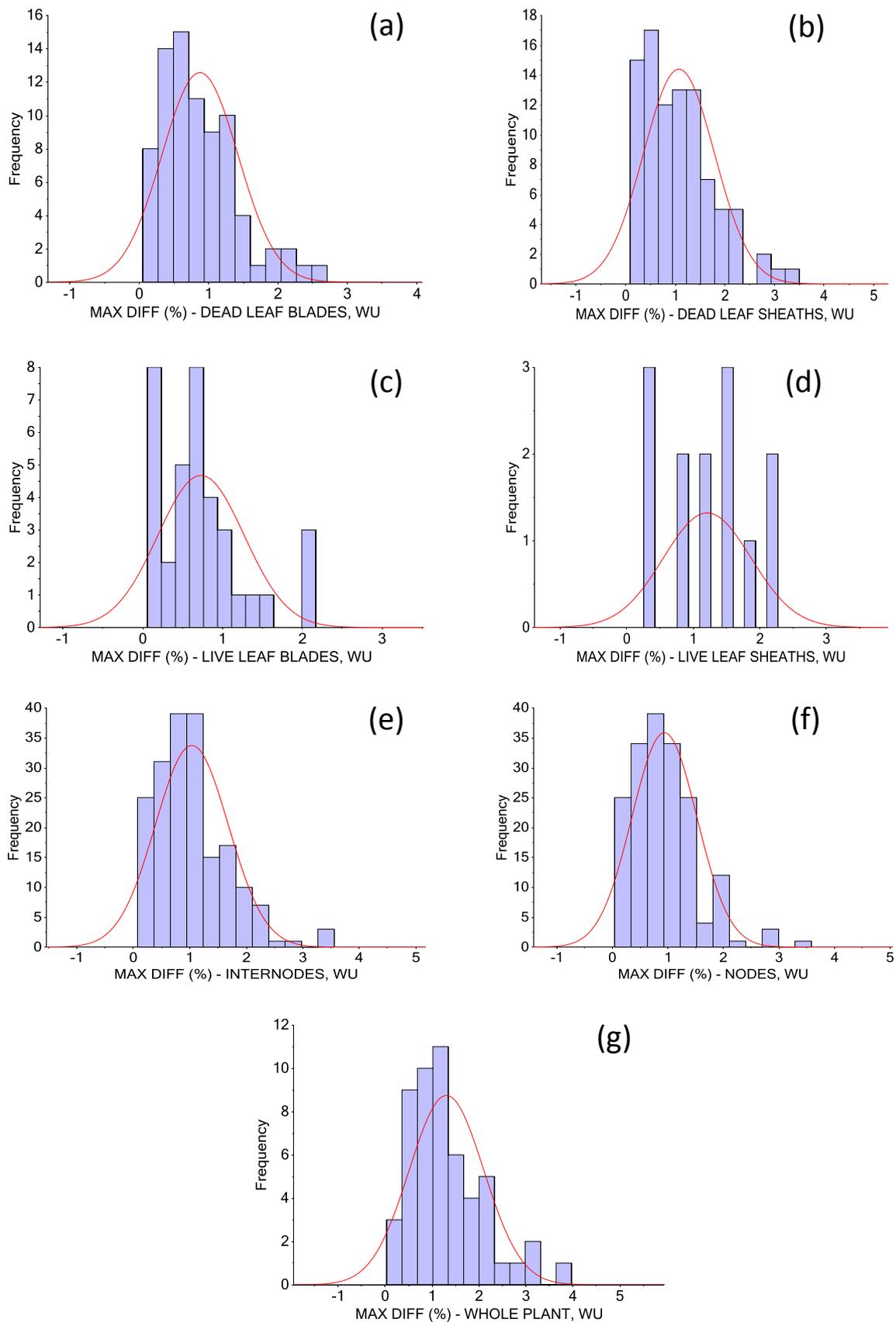


Figure F-20: Histograms for the maximum absolute difference in predicted glucose (% whole dry mass) values for the replicate WU scans of (a) dead leaf blades; (b) dead leaf sheaths; (c) live leaf blades; (d) live leaf sheaths; (e) internodes; (f) nodes; (g) whole plant (WP).

Appendix G Figures and Tables for Chapter 16: Lignocellulosic Properties

Table G-1: Compositional data for a *Miscanthus x giganteus* plant that was over 2 m in height and a plant of the same variety that was less than 1 m tall.

Plant Fraction	Extr. (%)	Ash (%)	Ara. (%)	Gal. (%)	Rha. (%)	Glu. (%)	Xyl. (%)	Man. (%)	Total Sugars	AIR (%)	KL (%)	ASL (%)	AIA (%)	TOTAL (%)	Ara :Xyl	Hc :Cel	Nitr. (%)
A plant Over 2 m High (3 Stem Sections - X1, X2, X3)																	
Live leaf blades, K	12.19	4.60	3.40	1.07	0.27	30.15	17.55	0.16	52.60	14.59	13.83	4.93	0.76	88.15	0.19	0.74	1.87
Live leaf sheaths, M	8.40	2.58*	2.88	0.86	0.11	41.30	20.70	0.21	66.06	17.64	16.66	2.17	0.98	95.86	0.14	0.60	0.43
Dead leaf blades, F	4.56	5.07	3.74	1.09	0.30	37.15	20.50	0.29	63.07	19.52	17.87	3.13	1.65	93.70	0.18	0.70	0.45
Dead leaf sheaths, H	3.93	3.00	3.03*	0.79*	0.20*	40.51*	20.59*	0.18*	66.03*	20.19	18.81	1.81	1.38	92.85	0.15	0.61	0.23
Stem, Internode 1m, X1T	6.27	2.06*	1.13	0.54	0.08	46.49	17.74	0.05	66.04	21.32	21.00	1.31	0.32	96.69	0.06	0.42	0.22*
Stem, Node 1m, X1N	6.73	1.79	1.70	0.65	0.09	41.54	19.07	0.14	63.19	22.25	22.08	1.66	0.17	95.46	0.09	0.52	0.28*
Whole Stem, 1m, X1	6.31	2.04*	1.18	0.55	0.08	46.06	17.86	0.06	65.79	21.41	21.10	1.34	0.31	96.58	0.07	0.43	0.22*
Stem, Internode 2m, X2T	7.19	3.15*	1.33	0.38	0.08	46.07	19.42	0.04	67.33	18.05	17.67	1.66	0.38	96.99	0.07	0.46	0.15
Stem, Node 2m, X2N	6.79*	2.67*	2.28*	0.91*	0.05*	39.59*	20.46*	0.27*	64.55*	18.80*	19.01*	2.23*	0.00*	94.28	0.11	0.61	0.47*
Whole Stem, 2m, X2	7.15*	3.11*	1.41*	0.43*	0.08*	45.51*	19.51*	0.06*	67.09*	18.11*	17.79*	1.71*	0.34*	96.76	0.07	0.47	0.18*
Stem, Internode 3m, X3T	6.61	7.02	2.77	0.84	0.16	38.26	21.15	0.17	63.37	14.10	13.59	4.13	0.51	94.72	0.13	0.66	1.01*
Stem, Node 3m, X3N	11.66*	10.03*	3.34*	1.30*	0.14*	34.12*	20.71*	0.38*	62.20*	16.01*	14.87*	3.44*	2.70*	99.99	0.16	0.76	0.85*
Whole Stem, 3m, X3	6.94*	7.21*	2.81*	0.87*	0.16*	37.99*	21.12*	0.19*	63.29*	14.23*	13.68*	4.09*	0.66*	95.06	0.13	0.66	1.00*
All Stem, X	6.66*	2.73*	1.36*	0.53*	0.08*	45.39*	18.66*	0.07*	66.13*	19.77*	19.45*	1.64*	0.34*	96.56	0.07	0.46	0.25*
Whole Plant, WP	7.21*	3.28*	2.06*	0.69*	0.14*	41.82*	18.85*	0.12*	63.73*	18.90*	18.30*	2.33*	0.61*	94.79	0.11	0.52	0.52*
Flower (another plant)	4.33	4.08	5.61	1.45	0.18	29.69	26.92	0.30	64.15	20.24	19.94	3.57	0.30	96.07	0.21	1.16	0.98
A plant Less Than 1 m High (1 Stem Section - X1)																	
Live leaf blades, K	12.23*	5.68*	3.57*	1.18*	0.24*	29.47*	18.12*	0.21*	51.13*	15.95*	13.39*	4.89*	2.17*	88.97*	0.13	0.61	1.69*
Dead leaf blades, F	7.06*	8.37*	3.75*	1.16*	0.47*	30.65*	18.51*	0.34*	54.07*	19.39*	16.15*	4.04*	3.96*	90.50*	0.08	0.53	0.74*
Dead leaf sheaths, H	4.90*	3.10*	2.96*	0.71*	0.05*	42.80*	22.37*	0.10*	68.52*	18.50*	17.94*	2.27*	1.66*	97.21*	0.13	0.69	0.33*
Stem, Internode 1m, X1T	5.02*	3.25*	1.62*	0.60*	0.13*	42.70*	20.11*	0.07*	65.86*	18.25*	17.36*	2.41*	0.58*	93.28*	0.09	0.55	0.24*
Stem, Node 1m, X1N	4.88*	3.72*	2.87*	1.03*	0.06*	37.24*	21.28*	0.28*	64.51*	19.70*	19.10*	2.67*	1.10*	93.14*	0.14	0.64	0.53*
Whole Stem, 1m, X1	5.00*	3.32*	1.81*	0.67*	0.12*	41.88*	20.28*	0.10*	65.66*	18.47*	17.62*	2.45*	0.66*	93.26*	0.13	0.61	0.29*
Whole Plant, WP	6.98*	4.75*	2.71*	0.88*	0.20*	37.16*	19.75*	0.17*	60.66*	18.08*	16.45*	3.27*	1.74*	92.32*	0.08	0.53	0.69*

* = Data supplied by NIRS calibration; Extr. = extractives; Ara. = arabinose; Gal. = galactose; Rha. = rhamnase; Glu. = glucose; Xyl. = xylose; Man. = mannose; AIR = acid insoluble residue; KL = Klason lignin; ASL = acid insoluble lignin; AIA = Acid insoluble ash; Hc: Cel = Hemicellulose to cellulose ratio; Nitr = nitrogen.

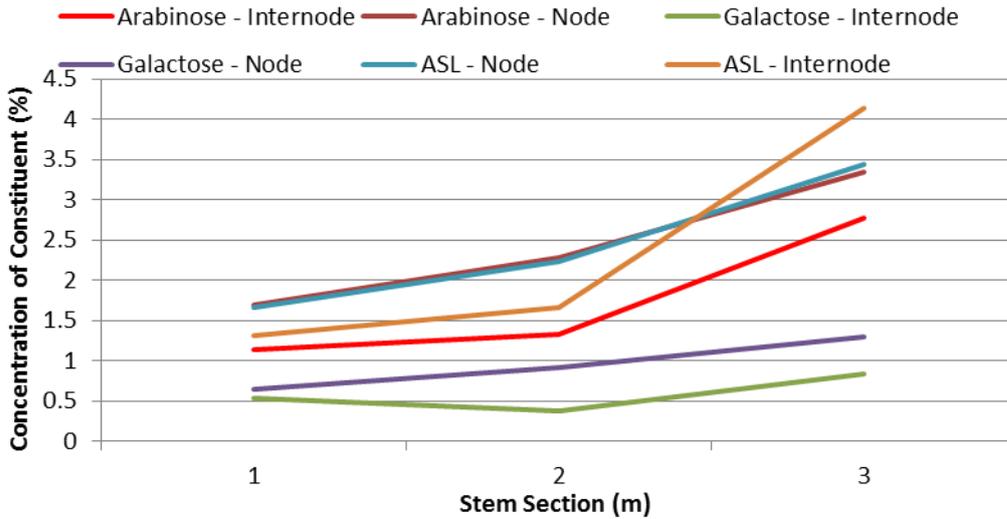


Figure G-1: Change, with stem section, in the concentrations of arabinose, galactose, and acid-soluble lignin (ASL) in the internode and node sections of the three-stem-section plant described in Table G-1.

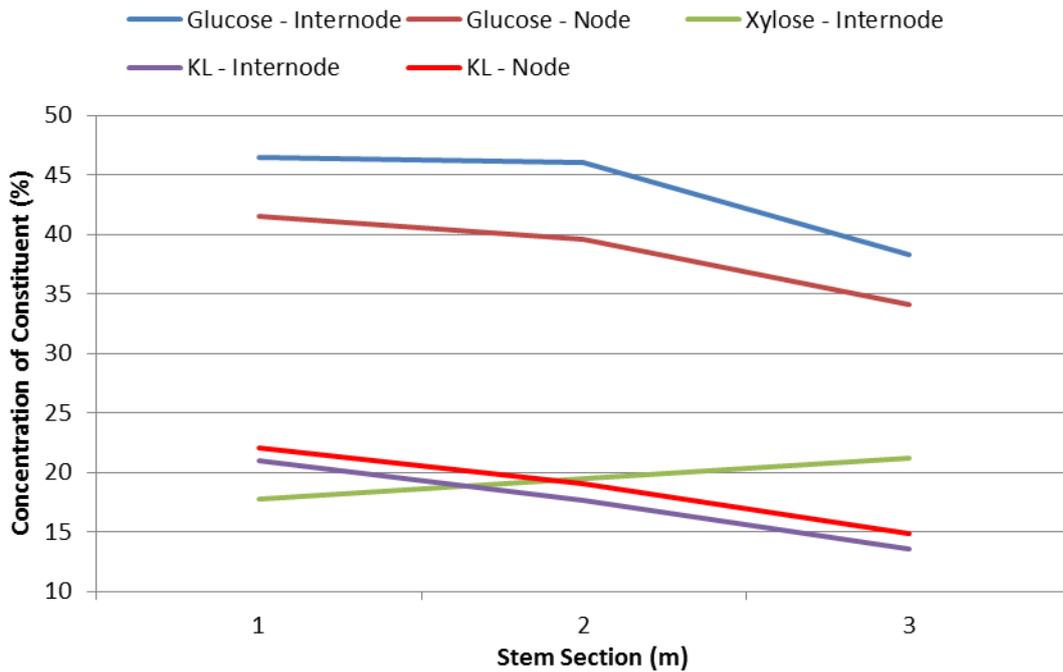


Figure G-2: Change, with stem section, in the concentrations of glucose, xylose, and Klason lignin (KL) in the internode and node sections of the three-stem-section plant described in Table G-1.

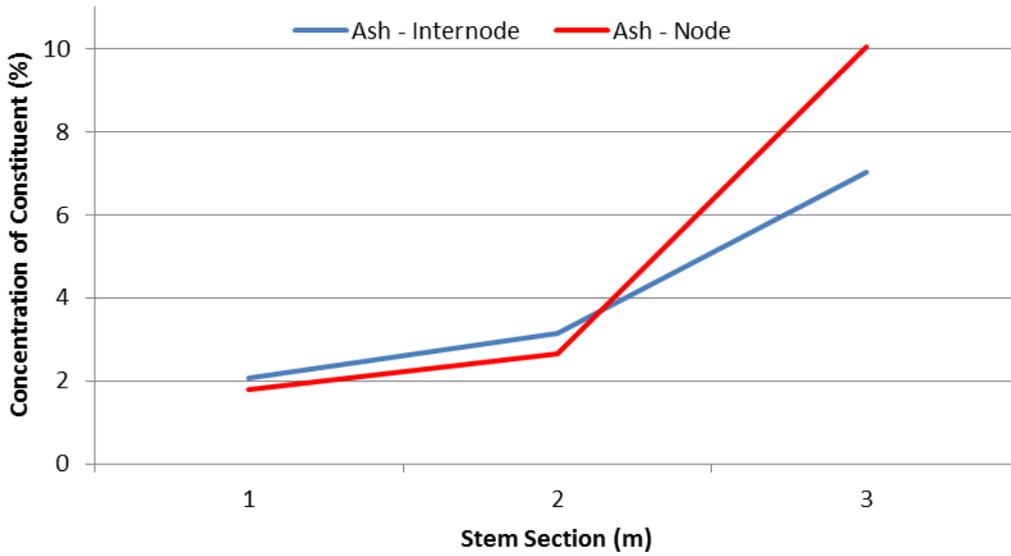


Figure G-3: Change, with stem section, in the concentration of ash in the internode and node sections of the three-stem-section plant described in Table G-1.

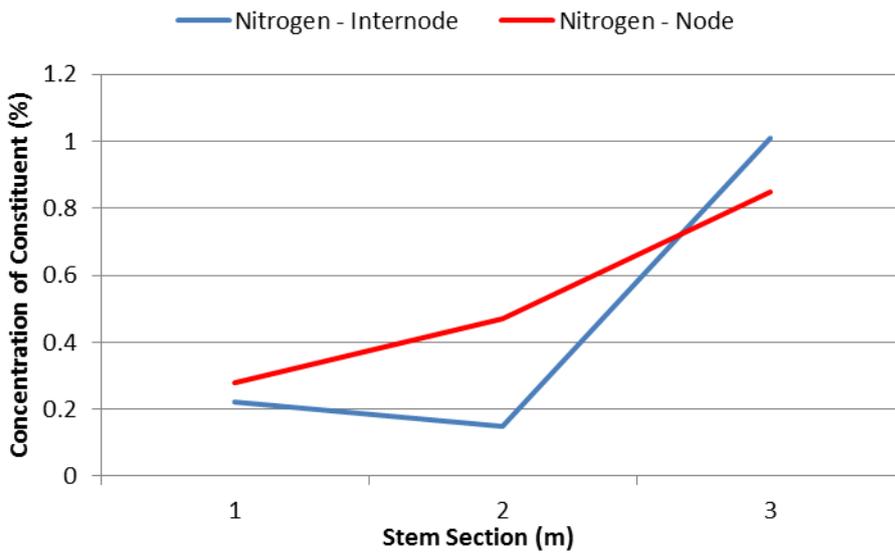


Figure G-4: Change, with stem section, in the concentration of nitrogen in the internode and node sections of the three-stem-section plant described in Table G-1.

Table G-2: Amounts (% DM) of 95% ethanol-soluble extractives, hot-water-soluble extractives, and the extractives removed after first employing a water extraction and then an ethanol extraction. Values are % of the dry mass of the sample and the numbers in the brackets represent the standard deviation of duplicates (SDD). Diff = (water + ethanol extraction) – (ethanol extraction) and % ethanol is the percentages of water + ethanol extractives represented by the ethanol extraction.

Sample	NIR #	Month	Variety	# Stem Sections	95% Ethanol Extractives (%)	Water Extractives (%)	Water + Ethanol Extr. (%)	Diff (%)	% Ethanol
Live Leaf Blade	14015	Dec	<i>Giganteus</i>	2	9.64 (0.09)	14.98 (0.05)	18.83 (0.19)	9.19	51.19%
Live Leaf Blade	14014	Nov	<i>Giganteus</i>	3	9.24 (0.22)	12.14 (0.07)	16.54 (0.01)	7.3	55.86%
Live Leaf Blade	14007	Oct	<i>Sinensis</i>	2	9.49 (0.02)	13.37 (0.16)	17.43 (0.14)	7.94	54.45%
Live Leaf Blade	14013	Oct	<i>Sinensis</i>	2	11.42 (0.20)	15.89 (0.03)	20.77 (0.34)	9.35	54.98%
Live Leaf Sheath	16011	Oct	<i>Sinensis</i>	2	7.46 (0.37)	9.96 (0.18)	11.56 (0.16)	4.1	64.53%
Dead Leaf Blade	10023	Jan	<i>Giganteus</i>	2	5.10 (0.05)	7.02 (0.09)	10.20 (0.16)	5.1	50.00%
Dead Leaf Blade	10031	Nov	<i>Giganteus</i>	3	9.45 (0.24)	13.14 (0.55)	17.13 (0.28)	7.68	55.17%
Dead leaf sheaths	12035	Nov	<i>Giganteus</i>	3	6.12 (0.24)	7.76 (0.11)	12.01 (0.09)	5.89	50.96%
Dead Leaf Sheath	12024	Jan	<i>Giganteus</i>	3	2.21 (0.14)	2.47 (0.18)	4.80 (0.01)	2.59	46.04%
X1 Node	5106	Nov	<i>Giganteus</i>	3	6.38 (0.04)	8.49 (0.19)	10.56 (0.09)	4.18	60.42%
X1 Node	5084	Feb	<i>Giganteus</i>	3	8.31 (0.16)	11.95 (0.26)	14.47 (0.32)	6.16	57.43%
X1 Internode	264	Jan	<i>Giganteus</i>	3	12.60 (0.39)	14.33 (0.01)	16.27 (0.07)	3.67	77.44%
X3 Internode	191	April	<i>Giganteus</i>	3	4.64 (0.11)	7.98 (0.07)	9.61 (0.02)	4.97	48.28%
WP (All Plant)	2555	Oct	<i>Giganteus</i>	2	12.68 (0.16)	15.14 (0.12)	17.53 (0.12)	4.85	72.33%
WP (1 st metre)	2515	Feb	<i>Giganteus</i>	3	4.14 (0.21)	4.10 (0.10)	6.34 (0.11)	2.2	65.30%

Klason Lignin Content

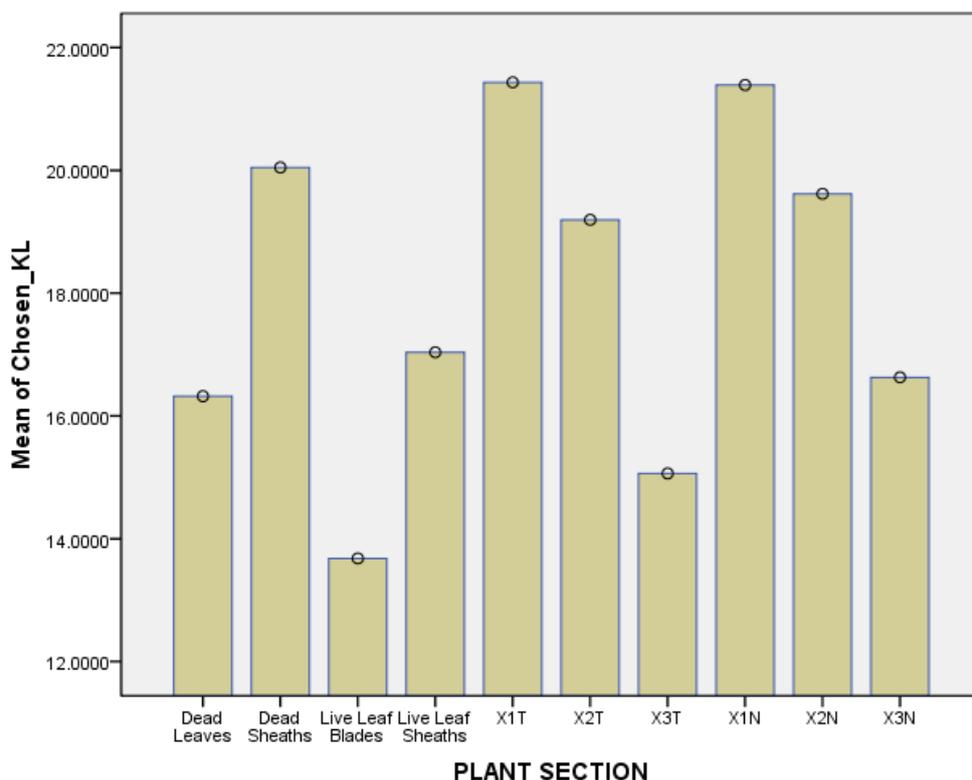


Figure G-5: Means for the Klason Lignin (KL) contents (% whole dry mass) of the plant sections used in the ANOVA tests on samples from 3-stem-section plants collected in October, November, and December 2007.

Table G-3: Significant differences between the means of the KL content of various plant fractions. ** = $P < 0.01$, * = $P < 0.05$.

	F	H	K	M	X1T	X2T	X3T	X1N	X2N	X3N
F	-	**	**		**	**		**	**	
H	**	-	**	**			**			**
K	**	**	-	**	**	**		**	**	**
M		**	**	-	**	*		**	*	
X1T	**		**	**	-	*	**		*	**
X2T	**		**	*	*	-	**	*		**
X3T		**			**	**	-	**	**	
X1N	**		**	**		*	**	-	*	**
X2N	**		**	*	*		**	*	--	**
X3N		**	**		**	**		**	**	-

F = dead leaf blades; H = dead leaf sheaths; K = live leaf blades; M = live leaf sheaths; X1T = 1st metre internode section; X2T = 2nd metre internode section; X3T = 3rd metre internode section; X1N = 1st metre node section; X2N = 2nd metre node section; X3N = 3rd metre node section

Glucose Content

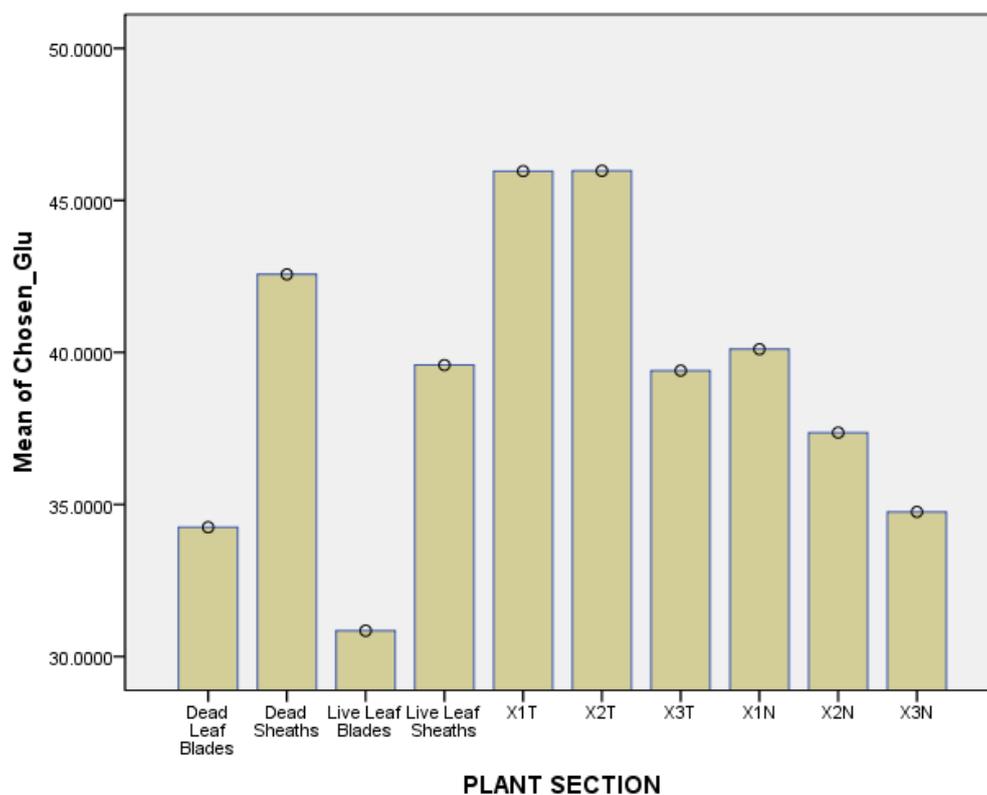


Figure G-6: Means for the glucose (Glu) contents (% whole dry mass) of the plant sections used in the One-Way ANOVA tests on samples from 3-stem-section plants collected in October, November, and December 2007.

Table G-4: Significant differences between the means of the glucose content of various plant fractions. ** = $P < 0.01$, * = $P < 0.05$.

	F	H	K	M	X1T	X2T	X3T	X1N	X2N	X3N
F	-	**	**	**	**	**	**	**	**	
H	**	-	**	**	**	**	**	**	**	**
K	**	**	-	**	**	**	**	**	**	**
M	**	**	**	-	**	**				**
X1T	**	**	**	**	-		**	**	**	**
X2T	**	**	**	**		-	**	**	**	**
X3T	**	**	**		**	**	-			**
X1N	**	**	**		**	**		-	**	**
X2N	**	**	**		**	**		**	--	**
X3N		**	**	**	**	**	**	**	**	-

Xylose Content

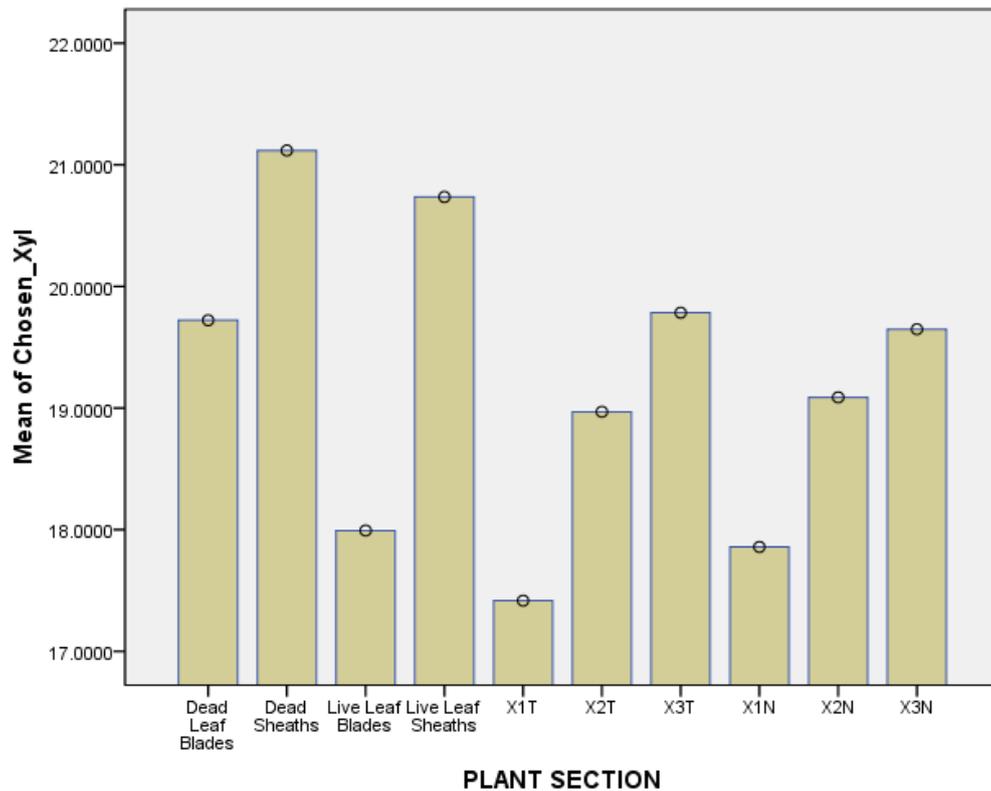


Figure G-7: Means for the xylose (Xyl) contents (% whole dry mass) of the plant sections used in the One-Way ANOVA tests on samples from 3-stem-section plants collected in October, November, and December 2007.

Table G-5: Significant differences between the means of the xylose content of various plant fractions. ** = $P < 0.01$, * = $P < 0.05$.

	F	H	K	M	X1T	X2T	X3T	X1N	X2N	X3N
F	-	**	**		**			*		
H	**	-	**		**	*		**	*	
K	**	**	-	**						*
M			**	-	**			**		
X1T	**	**		**	-		*			*
X2T		*				-				
X3T					*		-			
X1N	*	**		**				-		
X2N		*							--	
X3N			*		*					-

Arabinose Content

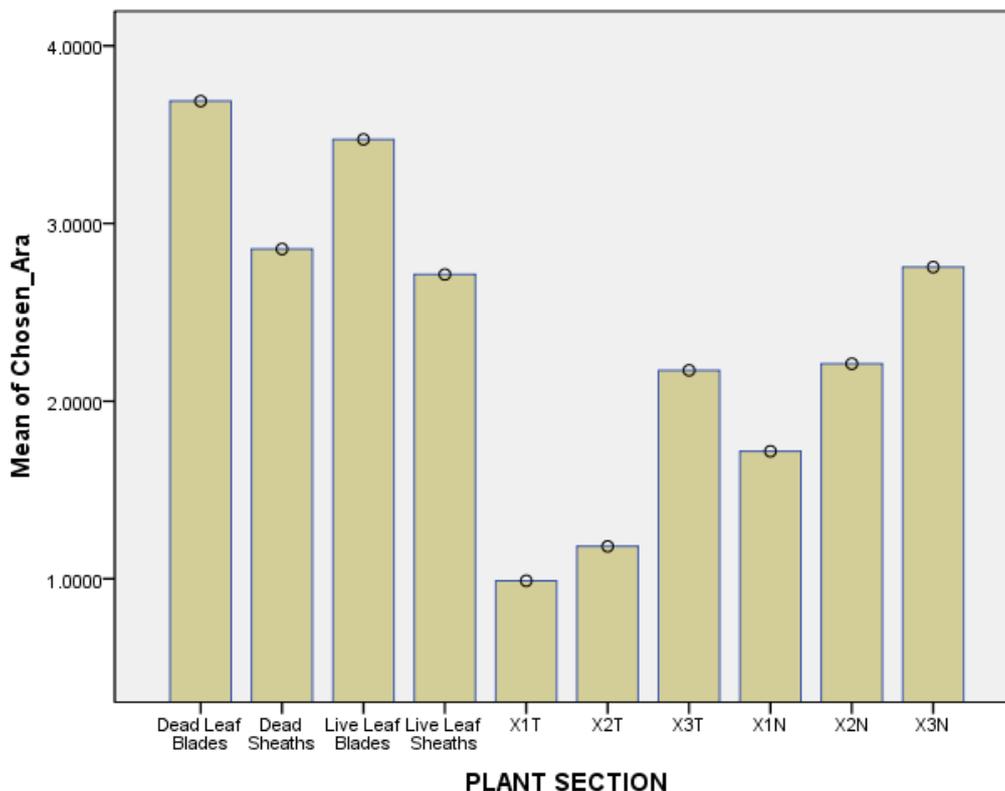


Figure G-8: Means for the arabinose (Ara) contents (% whole dry mass) of the plant sections used in the One-Way ANOVA tests on samples from 3-stem-section plants collected in October, November, and December 2007.

Table G-6: Significant differences between the means of the arabinose content of various plant fractions. ** = $P < 0.01$, * = $P < 0.05$.

	F	H	K	M	X1T	X2T	X3T	X1N	X2N	X3N
F	-	**		**	**	**	**	**	**	**
H	**	-	**		**	**		**	**	
K		**	-	**	**	**	**	**	**	**
M	**		**	-	**	**		**	**	
X1T	**	**	**	**	-		**	**	**	**
X2T	**	**	**	**		-	*	**	**	**
X3T	**		**		**	*	-			
X1N	**	**	**	**	**	**		-	**	**
X2N	**	**	**	**	**	**		**	--	*
X3N	**		**		**	**		**	*	-

Galctose Content

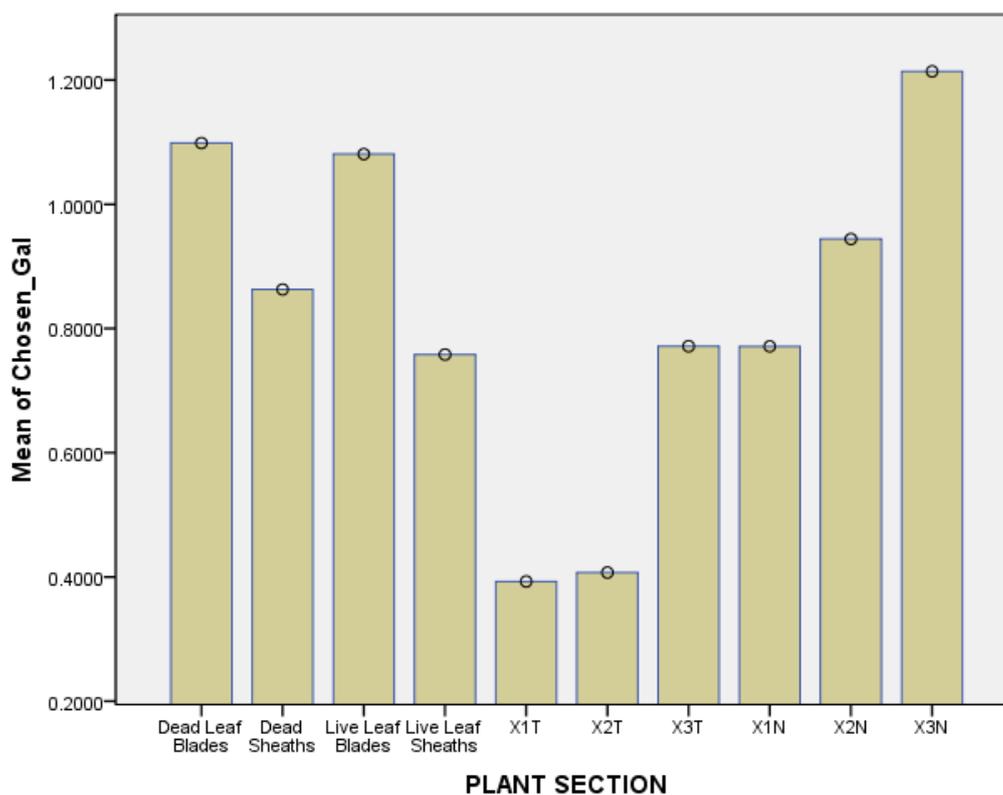


Figure G-9: Means for the galactose (Gal) contents (% whole dry mass) of the plant sections used in the One-Way ANOVA tests on samples from 3-stem-section plants collected in October, November, and December 2007.

Table G-7: Significant differences between the means of the galactose content of various plant fractions. ** = $P < 0.01$, * = $P < 0.05$.

	F	H	K	M	X1T	X2T	X3T	X1N	X2N	X3N
F	-	**		*	**	**	**	**		
H	**	-	**		**	**				**
K		**	-		**	**	**	**		
M	*			-	**	**				**
X1T	**	**	**	**	-		**	**	**	**
X2T	**	**	**	**		-	**	**	**	**
X3T	**		**		**	**	-			**
X1N	**		**		**	**		-		**
X2N					**	**			--	*
X3N		**		**	**	**	**	**	*	-

Rhamnose Content

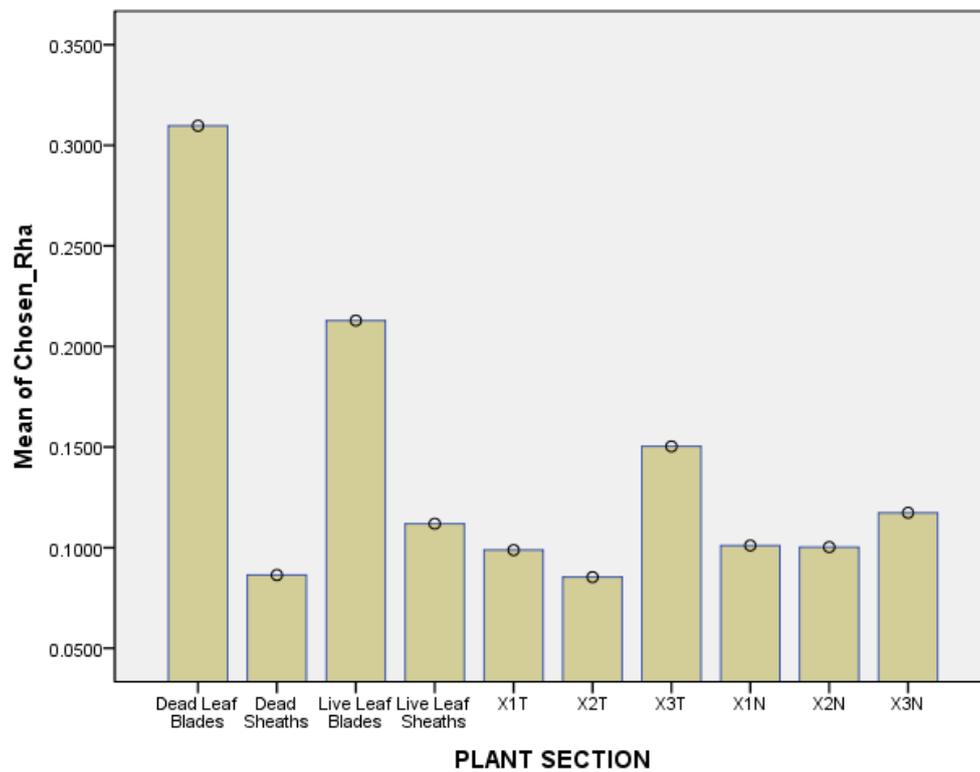


Figure G-10: Means for the rhamnose (Rha) (% whole dry mass) contents of the plant sections used in the One-Way ANOVA tests on samples from 3-stem-section plants collected in October, November, and December 2007.

Table G-8: Significant differences between the means of the rhamnose content of various plant fractions. ** = $P < 0.01$, * = $P < 0.05$.

	F	H	K	M	X1T	X2T	X3T	X1N	X2N	X3N
F	-	**		**	**	**	**	**	**	**
H	**	-	**							
K		**	-	*	**	**		**	**	*
M	**		*	-						
X1T	**		**		-					
X2T	**		**			-	*			
X3T	**					*	-			
X1N	**		**					-		
X2N	**		**						--	
X3N	**		*							-

Mannose Content

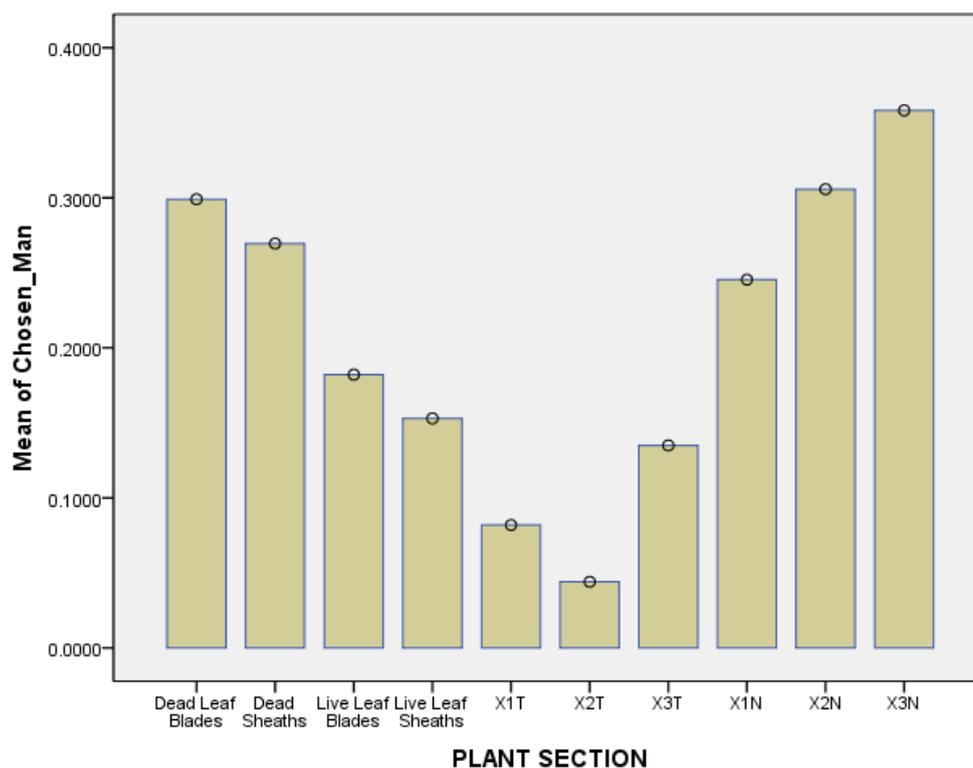


Figure G-11: Means for the mannose (Man) contents (% whole dry mass) of the plant sections used in the One-Way ANOVA tests on samples from 3-stem-section plants collected in October, November, and December 2007.

Table G-9: Significant differences between the means of the mannose content of various plant fractions. ** = $P < 0.01$, * = $P < 0.05$.

	F	H	K	M	X1T	X2T	X3T	X1N	X2N	X3N
F	-		**		**	**	**			
H		-	*		**	**	**			
K	**	*	-		**	**			**	**
M				-						
X1T	**	**	**		-			**	**	**
X2T	**	**	**			-	*	**	**	**
X3T	**	**				*	-		**	**
X1N					**	**		-		
X2N			**		**	**	**		--	
X3N			**		**	**	**			-

Hemicellulose:Cellulose Ratio

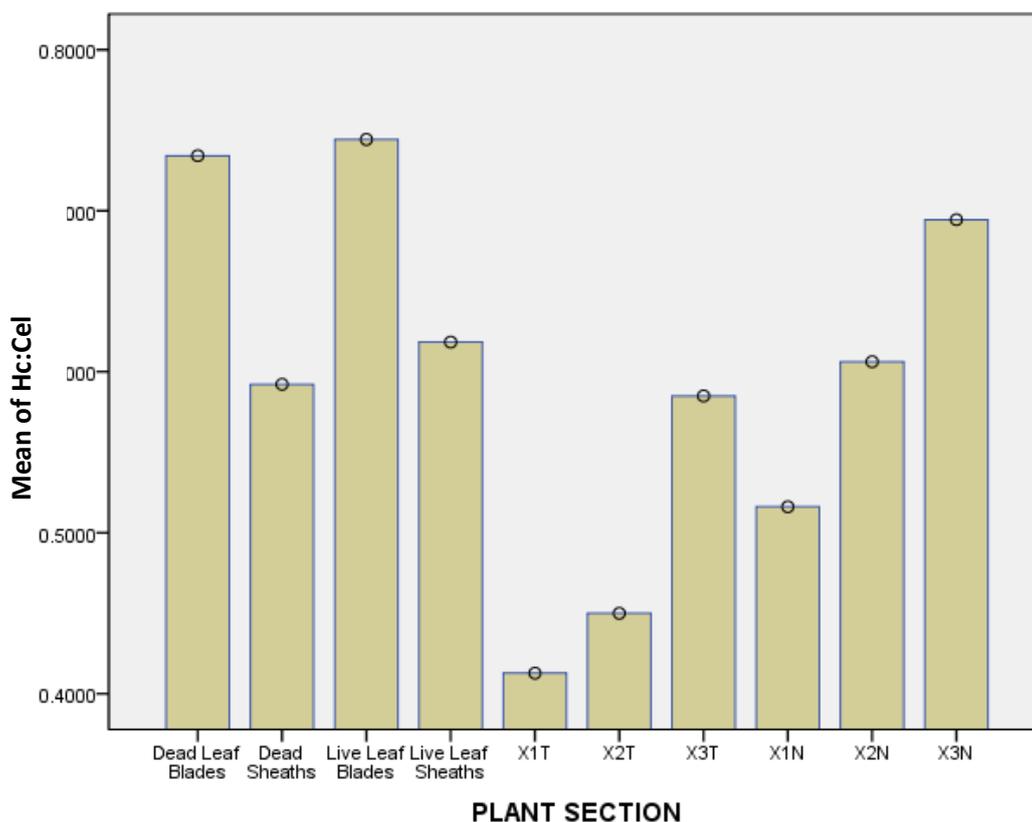


Figure G-12: Means for the hemicellulose to cellulose ratios (Hc: Cel) of the plant sections used in the One-Way ANOVA tests on samples from 3-stem-section plants collected in October, November, and December 2007.

Table G-10: Significant differences between the means of the hemicellulose to cellulose ratios of various plant fractions. ** = $P < 0.01$, * = $P < 0.05$.

	F	H	K	M	X1T	X2T	X3T	X1N	X2N	X3N
F	-	**		**	**	**	**	**	**	
H	**	-	**		**	**		**		**
K		**	-	**	**	**	**	**	**	
M	**		**	-	**	**		**		*
X1T	**	**	**	**	-		**	**	**	**
X2T	**	**	**	**		-	**	**	**	**
X3T	**		**		**	**	-			*
X1N	**	**	**	**	**	**		-	**	**
X2N	**		**		**	**		**	--	*
X3N		**		*	**	**	*	**	*	-

Extractives Content

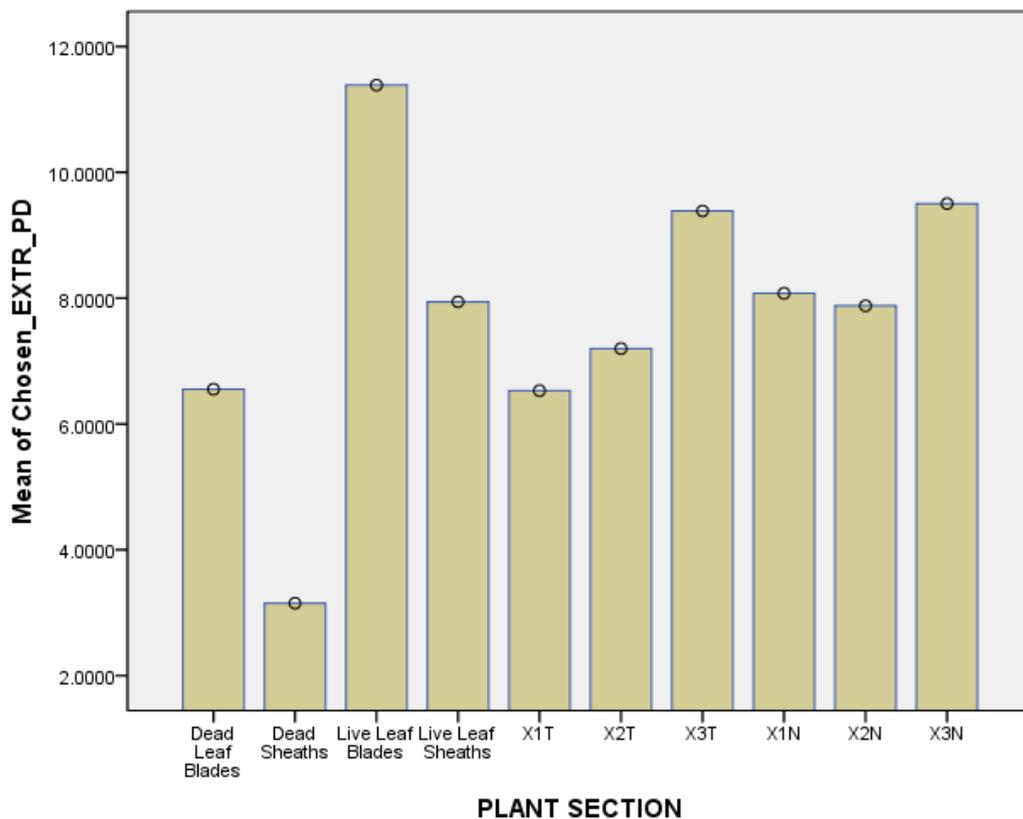


Figure G-13: Means for the extractives (EXTR_PD) contents (% whole dry mass) of the plant sections used in the One-Way ANOVA tests on samples from 3-stem-section plants collected in October, November, and December 2007.

Table G-11: Significant differences between the means of the extractives contents of various plant fractions. ** = $P < 0.01$, * = $P < 0.05$.

	F	H	K	M	X1T	X2T	X3T	X1N	X2N	X3N
F	-	**	**							*
H	**	-	**	**	**	**	**	**	**	**
K	**	**	-		**	**		*	*	
M		**		-						
X1T		**	**		-					
X2T		**	**			-				
X3T		**					-			
X1N		**	*					-		
X2N		**	*						--	
X3N	*	**								-

Ash Content

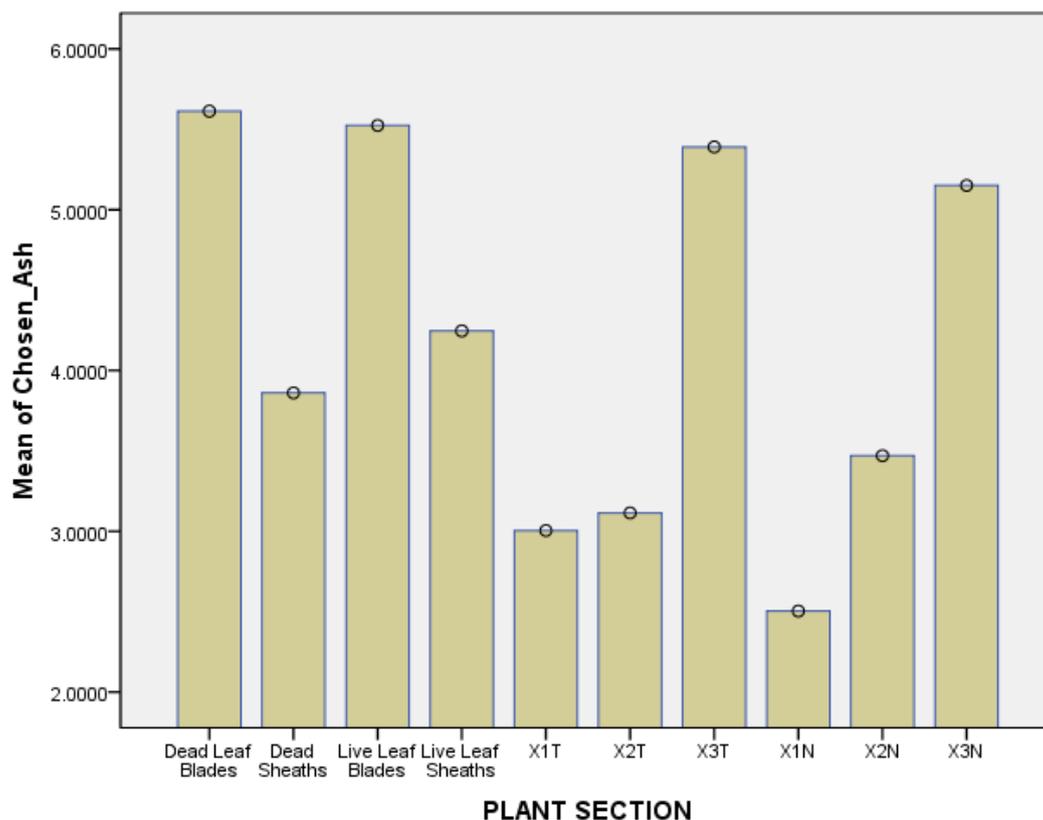


Figure G-14: Means for the ash contents (% whole dry mass) of the plant sections used in the One-Way ANOVA tests on samples from 3-stem-section plants collected in October, November, and December 2007.

Table G-12: Significant differences between the means of the ash contents of various plant fractions. ** = $P < 0.01$, * = $P < 0.05$.

	F	H	K	M	X1T	X2T	X3T	X1N	X2N	X3N
F	-	*			**	**		**	*	
H	*	-								
K			-		**	*		**		
M				-						
X1T	**		**		-		*			*
X2T	**		*			-	*			
X3T					*	*	-	**		
X1N	**		**				**	-		**
X2N	*								--	
X3N					*			**		-

Acid Soluble Lignin (ASL) Content

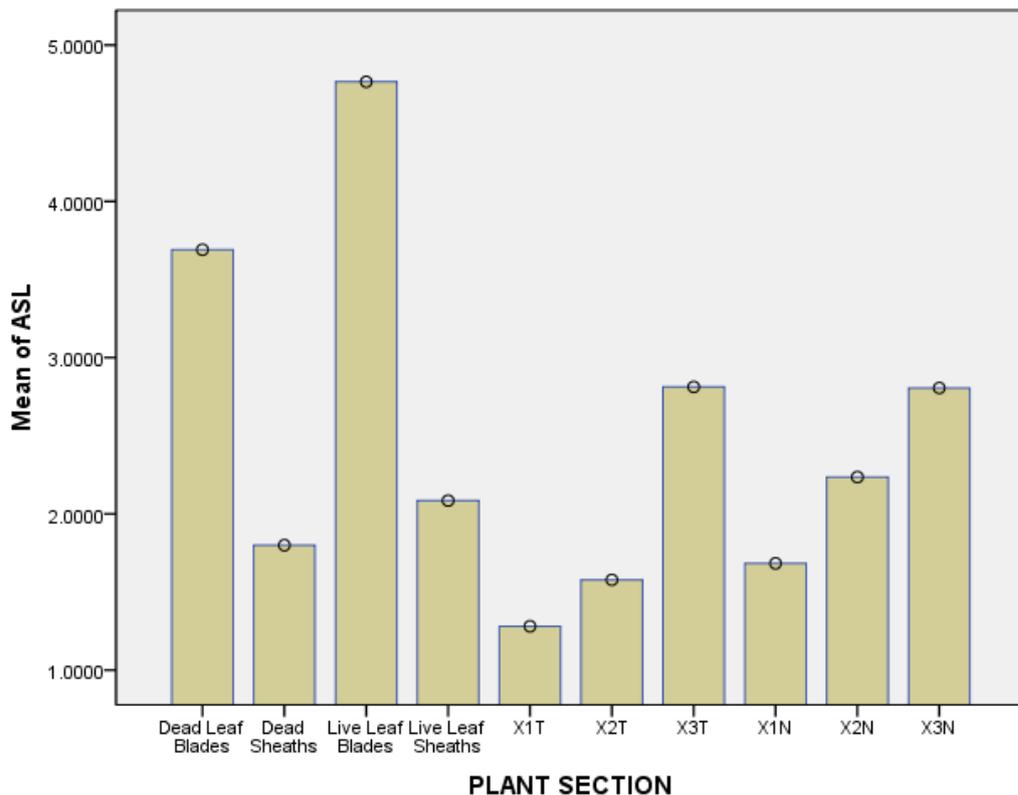


Figure G-15: Means for the ASL contents (% whole dry mass) of the plant sections used in the One-Way ANOVA tests on samples from 3-stem-section plants collected in October, November, and December 2007.

Table G-13: Significant differences between the means of the ASL contents of various plant fractions. ** = $P < 0.01$, * = $P < 0.05$.

	F	H	K	M	X1T	X2T	X3T	X1N	X2N	X3N
F	-	**	**	**	**	**		**	**	**
H	**	-	**	*	**				*	**
K	**	**	-	**	**	**	**	**	**	**
M	**	*	**	-	**	**		**		**
X1T	**	**	**	**	-	*	*	**	**	**
X2T	**		**	**	*	-			**	**
X3T			**		*		-			
X1N	**		**	**	**			-	**	**
X2N	**	*	**		**	**		**	--	*
X3N	**	**	**	**	**	**		**	*	-

Nitrogen Content

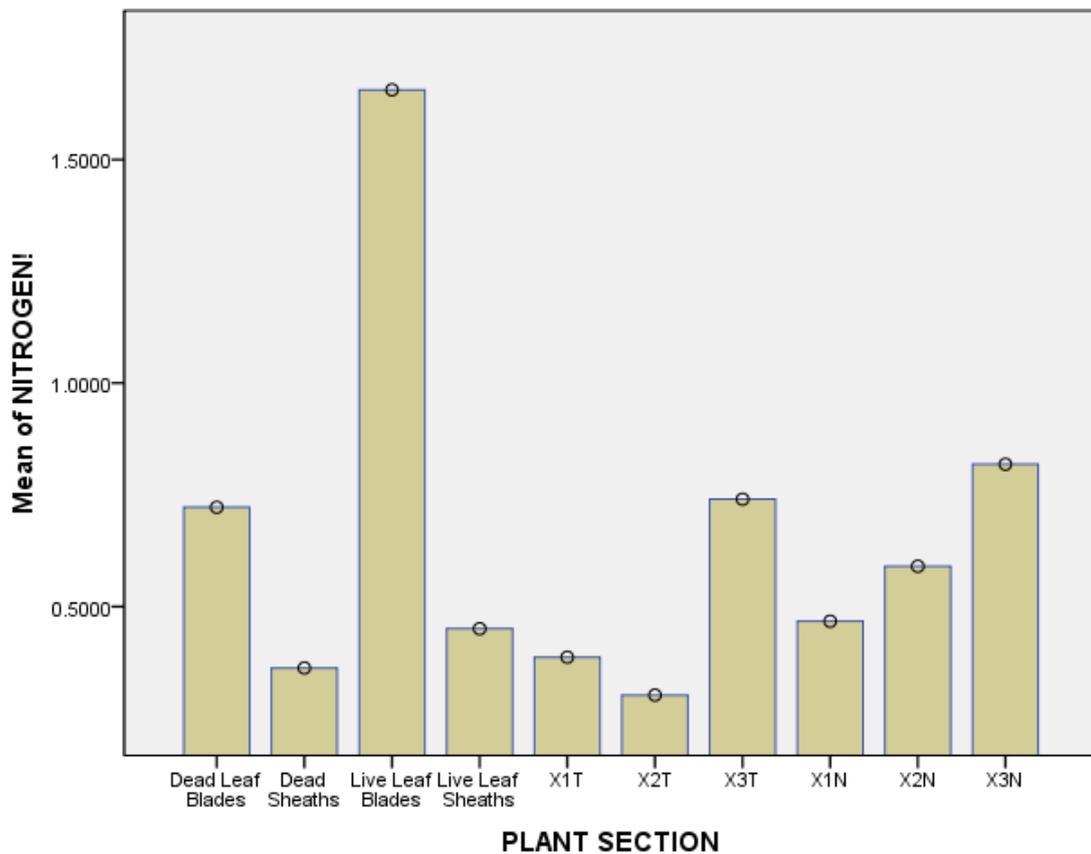


Figure G-16: Means for the nitrogen contents (% whole dry mass) of the plant sections used in the One-Way ANOVA tests on samples from 3-stem-section plants collected in October, November, and December 2007.

Table G-14: Significant differences between the means of the nitrogen contents of various plant fractions. ** = $P < 0.01$, * = $P < 0.05$.

	F	H	K	M	X1T	X2T	X3T	X1N	X2N	X3N
F	-	**	**		*	**				
H	**	-	**				**			**
K	**	**	-	**	**	**	**	**	**	**
M			**	-						
X1T	*		**		-		*			**
X2T	**		**			-	**			**
X3T		**	**		*	**	-			
X1N			**					-		*
X2N			**						--	
X3N		**	**		**	**		*		-



Figure G-17: Photographs of Miscanthus varieties. (a) *Miscanthus x sinensis*; (b) MSXY1; (c) MSXY2; (d) MSXY3.



Figure G-18: Photographs of more *Miscanthus* varieties. (a) MSXY5; (b) MSXY6; (c) MSXY7; (d) *Miscanthus x giganteus*

Table G-15: Summary statistics for the reference analytical data (% whole dry mass basis) of Early harvest samples of K (live leaf blade), M (live leaf sheath), F (dead leaf blade), H (dead leaf sheath) samples of *Miscanthus x giganteus*. Data for the separate fractions of a *sinensis* plant (MSSS5) and a whole *sinensis* plant (MSSS4) are also provided. Ara:xyl = arabinose to xylose ratio; Hc: Cel = hemicellulose to cellulose ratio. * = data supplied by NIRS model.

Plant Fraction	Extr.	Ash	Ara.	Gal.	Rha.	Glu.	Xyl.	Man.	Total Sugars	AIR	KL	ASL	AIA	TOTAL	Ara :Xyl	Hc :Cel	Nitrogen
K (<i>Giganteus</i> (15) (Av)	11.61	5.97	3.57	1.09	0.27	30.36	17.70	0.19	53.19	15.80	13.97	4.69	1.82	89.48	0.20	0.75	1.51
K (<i>Giganteus</i>) (SD)	1.79	0.97	0.23	0.06	0.06	1.10	0.70	0.05	2.03	0.91	0.51	0.41	0.60	1.26	0.01	0.01	0.35
K (<i>Giganteus</i>) (Max)	15.29	8.16	3.79	1.20	0.40	31.76	18.43	0.32	55.26	16.78	14.94	5.66	2.64	91.24	0.21	0.78	2.20
K (<i>Giganteus</i>) (Min)	9.01	4.53	3.11	0.99	0.17	27.89	16.11	0.13	48.61	13.79	12.86	4.15	0.76	87.64	0.19	0.73	0.89
M (<i>Giganteus</i> (4) (Av)	7.94	4.80	2.71	0.76	0.11	39.58	20.74	0.15	64.06	18.47	17.04	2.09	1.43	95.36	0.13	0.62	0.55
M (<i>Giganteus</i>) (SD)	2.30	0.91	0.16	0.08	0.01	1.33	0.50	0.08	1.90	1.06	0.89	0.09	0.45	0.83	0.01	0.02	0.15
M (<i>Giganteus</i>) (Max)	10.66	5.82	2.88	0.86	0.12	41.30	21.45	0.24	66.06	19.82	17.85	2.17	1.96	96.27	0.14	0.64	0.74
M (<i>Giganteus</i>) (Min)	5.08	4.09	2.57	0.67	0.11	38.14	20.39	0.07	61.95	17.62	15.97	1.98	0.98	94.65	0.13	0.60	0.32
F (<i>Giganteus</i> (18) (Av)	7.13	6.65	3.63	1.19	0.31	32.61	18.94	0.38	57.06	19.35	16.23	4.06	3.06	91.08	0.19	0.75	0.90
F (<i>Giganteus</i>) (SD)	2.40	2.03	0.28	0.12	0.07	2.00	1.20	0.16	3.28	2.30	1.16	0.54	1.60	1.53	0.02	0.02	0.29
F (<i>Giganteus</i>) (Max)	11.80	10.65	4.00	1.43	0.42	37.15	20.50	0.64	63.07	23.92	18.10	4.77	6.10	93.70	0.21	0.79	1.80
F (<i>Giganteus</i>) (Min)	4.36	3.73	3.17	1.01	0.19	28.28	16.14	0.17	49.21	15.86	14.21	3.09	1.16	88.33	0.15	0.70	0.45
H (<i>Giganteus</i> (15) (Av)	3.36	4.19	2.87	0.82	0.13	41.82	21.15	0.24	67.03	20.82	18.56	2.06	2.27	94.93	0.14	0.60	0.40
H (<i>Giganteus</i>) (SD)	1.19	1.18	0.34	0.11	0.04	1.49	0.54	0.06	1.52	1.14	1.21	0.36	0.68	1.02	0.02	0.03	0.11
H (<i>Giganteus</i>) (Max)	6.12	6.38	3.56	1.06	0.23	44.26	22.18	0.31	69.81	24.02	21.64	2.93	3.30	96.62	0.17	0.67	0.72
H (<i>Giganteus</i>) (Min)	2.10	2.64	2.33	0.65	0.10	38.62	20.38	0.11	64.33	19.42	16.78	1.45	1.32	93.29	0.11	0.57	0.20
X (<i>Giganteus</i> (17) (Av)	6.05	2.80	1.53	0.53	0.10	44.64	19.31	0.11	66.22	20.45	19.82	1.70	0.63	96.51	0.08	0.48	0.43
X (<i>Giganteus</i>) (SD)	3.11	1.43	0.44	0.26	0.03	3.08	1.46	0.09	2.83	2.14	2.13	0.40	0.30	1.25	0.02	0.08	0.24
X (<i>Giganteus</i>) (Max)	12.84	5.34	2.90	1.32	0.17	47.63	21.63	0.42	69.86	24.10	23.49	2.81	1.20	98.35	0.14	0.71	1.23
X (<i>Giganteus</i>) (Min)	2.36	0.94	1.13	0.29	0.06	36.30	16.39	0.04	60.86	15.86	15.11	1.25	0.26	92.58	0.07	0.41	0.04
Sinensis Plant (2 Stem Sections)																	
Live leaf blades	12.02	3.82	3.62	1.05	0.50	34.69	16.23	0.25	56.33	14.92	13.95	4.54	0.97	90.66	0.22	0.62	1.07
Live leaf sheaths	6.78	2.45	3.36	0.77	0.13	39.17	23.73	0.08	67.24	16.42	15.52	2.60	0.90	94.59	0.14	0.72	0.19*
Dead leaf blades	6.23	4.48	3.88	1.20	0.75	35.99	17.43	0.34	59.59	20.52	19.12	4.08	1.40	93.51	0.22	0.66	0.53*
Dead leaf sheaths	6.13	2.85	3.63	0.83	0.15	39.27	22.11	0.13	66.11	19.08*	17.85	2.64	0.99	95.57	0.16	0.68	0.16*
Stem, 1 st m	4.98	1.51	1.95	0.40	0.09	42.98	23.62	0	69.03	18.88	18.30	1.66	0.58	95.51	0.08	0.61	0.01*
Stem 2 nd m	9.28	3.51	2.99	0.81	0.11	36.80	24.06	0.05	64.81	16.04*	14.70*	2.45	1.06*	94.76	0.12	0.76	0.38*
Sinensis Whole Plant Sample (2 Stem Sections)																	
Whole Plant	4.22	1.88	2.79	0.67	0.16	39.84	24.54	0	68.01	19.19	18.51	2.16	0.68	94.94	0.11	0.71	0.22*

Table G-16: Compositional data (% whole mass dry basis) for the separate fractions of plants of different *Miscanthus* varieties. Ara:xyl = arabinose to xylose ratio; Hc: Cel = hemicellulose to cellulose ratio. Numbers in bold are greater than the maximum giganteus values for that plant fraction whilst numbers in italics are lower than the minimum giganteus values for that plant fraction. * = data supplied by NIRS model.

Plant Fraction	Extr.	Ash	Ara.	Gal.	Rha.	Glu.	Xyl.	Man.	Total Sugars	AIR	KL	ASL	AIA	TOTAL	Ara :Xyl	Hc :Cel	Nitrogen
Plant MSXY1																	
Live leaf blades	11.12	5.09	3.90	1.14	0.61	32.90	17.38	0.29	56.23	15.76	13.96	4.58	1.80	90.98	0.22	0.71	-
Live leaf sheaths	6.09	3.53	3.35	0.78	0.14	38.02	24.03	0.11	66.44	15.35	14.32	2.48	1.03	92.86	0.14	0.75	0.48*
Dead leaf blades	8.11	6.23	3.96	1.21	0.66	33.18	18.39	0.35	57.75	18.66	15.97	4.35	2.69	92.41	0.22	0.74	0.60*
Dead leaf sheaths	4.46	3.19	3.89	0.90	0.15	39.58	23.66	0.19	68.38	16.72	15.20	2.75	1.52	93.98	0.16	0.73	0.27*
Stem, 1 st m	6.36	2.14	2.09	0.58	0.10	39.86	23.91	0.09	66.63	18.45	17.90	1.68	0.55	94.71	0.09	0.67	0.10*
Stem 2 nd m	7.40	3.19	2.32	0.70	0.12	38.18	24.02	0.13	65.48	17.24	16.57	1.89	0.67	94.52	0.10	0.71	0.29*
Plant MSXY2																	
Live leaf blades	13.39	7.28*	3.55*	1.10*	0.60*	30.18*	16.22*	0.18*	52.16*	15.58*	13.84*	4.52*	2.49*	90.86	0.22	0.72	1.31
Live leaf sheaths	5.95	3.79*	2.59*	0.54*	0.06*	39.67*	23.81*	0.06*	67.30*	15.97*	15.30*	2.58*	1.13*	94.34	0.11	0.68	0.45
Dead leaf blades	7.43	7.59	3.99	1.39	0.58	31.27	18.38	0.36	55.96	20.51	16.78	2.87	3.74	90.63	0.22	0.79	0.65*
Dead leaf sheaths	3.42	5.34	3.41	0.79	0.14	38.97	23.55	0.17	67.04	19.41	16.82	2.35	2.59	94.96	0.14	0.72	0.21*
Stem, 1 st m	4.52	1.72	1.92	0.46	0.09	42.11	24.33	0.07	68.98	19.58*	19.07*	1.64	0.00	95.94	0.08	0.64	-
Stem 2 nd m	6.49	3.83	2.57*	0.62*	0.11*	38.39*	25.59*	0.12*	67.12*	16.45*	15.94*	2.21*	0.98*	95.87	0.10	0.76	0.29*
Plant MSXY3																	
Live leaf blades	11.83	6.22	3.17*	1.06*	0.41*	30.00*	16.91*	0.18*	51.57*	15.94	14.16	4.78	1.79	88.71	0.19	0.72	1.47*
Live leaf sheaths	6.31	4.86	2.64*	0.67*	0.06*	39.08*	22.80*	0.16*	65.35*	16.50*	15.91*	2.37*	1.15*	94.85	0.12	0.67	0.41
Dead leaf blades	7.53	10.50	3.41	1.25	0.34	30.78	18.55	0.30	54.62	20.69	16.15*	3.54*	4.66*	92.34	0.18	0.77	0.54*
Dead leaf sheaths	4.75	4.82	3.16*	0.68*	0.11*	40.47*	24.02*	0.14*	69.12*	17.36*	17.01*	2.13*	1.73*	97.27	0.13	0.69	0.20*
Whole Stem	8.16	1.74	1.44	0.48	0.08	43.70	20.97	0.09	66.77	17.91	17.44	1.68	0.48	95.79	0.07	0.53	0.21*
Plant MSXY5																	
Dead leaf blades	12.10	7.47	3.49	1.03	0.59	30.29	15.37	0.24	51.01	17.60	14.44	5.37	3.16	90.39	0.23	0.68	1.02*
Live leaf sheaths	5.25	4.80	2.91*	0.63*	0.14*	40.00*	23.04*	0.07*	67.69*	16.01*	15.38*	2.58*	0.81*	94.78	0.13	0.67	0.36*
Whole Stem	6.79	2.98	2.21	0.56	0.09	41.15	22.49	0.10	66.61	16.84	16.29	1.99	0.55	94.66	0.10	0.62	0.14*

Table G-17: Whole-mass data for samples of *Miscanthus* varieties. Ara:xyl = arabinose:xylose; Hc: Cel = hemicellulose:cellulose. * = data supplied by NIRS.

Plant Fraction	Extr.	Ash	Ara.	Gal.	Rha.	Glu.	Xyl.	Man.	Total Sugars	AIR	KL	ASL	AIA	TOTAL	Ara :Xyl	Hc :Cel	Nitr.
Plant MSXY6																	
Live leaf blades	10.94	8.13	3.42	1.24	0.35	28.54	17.07	0.18	50.81	17.60	14.40	5.02	3.20	89.29	0.20	0.78	1.82
Live leaf sheaths	7.83	6.86	2.61*	0.69*	<i>0.07*</i>	39.68*	<i>20.34*</i>	0.16*	63.72*	<i>17.77*</i>	17.00*	2.31*	1.71*	97.55	0.13	0.60	0.49*
Dead leaf blades	7.56	13.02	3.63	1.38	0.44	28.99	15.81	0.43	50.68	23.91	15.96	4.04	7.95	91.27	0.23	0.75	0.69*
Dead leaf sheaths	5.38	6.16	3.59	1.06	0.13	39.15	<i>19.49</i>	0.24	63.65	19.84	18.00*	3.00	2.43*	96.20	0.18	0.63	0.38*
Stem, 1 st m	8.90	2.03	1.30	0.39	0.08	43.53	17.90	0.07	63.27	21.28	20.50	1.30	0.78	96.00	0.07	0.45	0.14*
Stem 2 nd m	7.24*	5.06*	1.76*	0.62*	0.09*	41.76*	<i>19.79*</i>	0.09*	64.15*	<i>17.16*</i>	15.93*	2.08*	1.77*	94.42	0.09	0.54	0.51*
Plant MSXY7																	
All Leaves	9.90	9.47	3.33	1.13	0.50	27.47	<i>14.85</i>	0.24	47.53	21.03	16.18	5.02	4.85	88.10	0.22	0.73	1.55
All Sheaths	5.84	5.86	3.07	0.81	0.16	38.76	<i>20.30</i>	0.15	63.24	<i>17.30</i>	<i>15.24</i>	3.30	2.06	93.49	0.15	0.63	-
All Stem	5.91	3.30	1.84	0.56	0.11	42.81	21.03	0.11	66.47	16.85	16.33	2.10	0.52	94.12	0.09	0.55	0.42*
Germany Samples																	
9001	3.13	2.92	2.24	0.60	0.14	43.14	22.74	0.16	69.02	20.20	18.83	1.95	1.37	95.85	0.10	0.60	
9002	2.98	3.75	2.28	0.63	0.16	43.00	22.28	0.22	68.57	20.42	18.47	2.07	1.95	95.84	0.10	0.59	
9003	2.98	3.15	2.08	0.54	0.13	43.07	22.22	<i>0.11</i>	68.15	20.39	18.83	2.07	1.56	95.19	0.09	0.58	
9005	2.39	3.14	<i>1.98</i>	<i>0.50</i>	0.13	44.76	21.77	0.13	69.26	20.31	18.64	2.09	1.67	95.51	<i>0.09</i>	<i>0.55</i>	
9006	4.14	3.01	2.16	0.60	<i>0.12</i>	41.54	21.90	0.20	66.52	20.81	19.27	2.12	1.54	95.06	0.10	0.60	
9007	3.00	3.40	2.33	0.65	0.16	41.91	22.31	0.22	67.58	21.14	19.17	1.90	1.96	95.05	0.10	0.61	
9009	3.19	4.25	2.48	0.67	0.16	<i>41.03</i>	<i>20.78</i>	0.21	65.33	21.43	18.80	2.29	2.63	93.86	0.12	0.59	
9010	2.47	3.32	2.24	0.58	0.14	42.94	22.64	0.13	68.67	<i>19.95</i>	18.33	2.06	1.62	94.85	0.10	0.60	
9011	2.79	3.20	2.34	0.64	0.16	41.64	22.12	0.20	67.11	20.52	<i>18.27</i>	2.26	2.25	93.63	0.11	0.61	
9012	3.18	3.65	2.25	0.64	0.14	42.59	21.92	0.20	67.74	21.11	19.17	<i>1.90</i>	1.94	95.64	0.10	0.59	
9013	3.12	4.68	2.21	0.64	0.16	42.13	22.06	0.21	67.40	21.66	19.06	1.94	2.59	96.21	0.10	0.60	
9014	3.09	3.22	2.21	0.62	0.14	42.52	21.43	0.19	67.12	20.88	18.38	2.22	2.49	94.03	0.10	0.58	
9015	2.46	3.68	2.55	0.70	0.16	41.88	21.59	0.18	67.05	21.93	19.94	2.09	1.99	95.22	0.12	0.60	
9016	2.78	3.15	2.25	0.66	0.14	42.75	22.26	0.17	68.22	20.11	18.74	2.12	1.37	95.01	0.10	0.60	
Average	2.98	3.47	2.26	0.62	0.15	42.49	22.00	0.18	67.69	20.77	18.85	2.08	1.92	95.07	0.10	0.59	
Standard Deviation	0.43	0.50	0.14	0.05	0.01	0.92	0.50	0.04	1.06	0.61	0.45	0.13	0.43	0.77	0.01	0.02	
Max	4.14	4.68	2.55	0.70	0.16	44.76	22.74	0.22	69.26	21.93	19.94	2.29	2.63	96.21	0.12	0.61	
Min	2.39	2.92	1.98	0.50	0.12	41.03	20.78	0.11	65.33	19.95	18.27	1.90	<i>1.37</i>	93.63	0.09	0.55	
Summary Data for Four 3-Stem-Section WP Samples of <i>Miscanthus x giganteus</i> Samples Collected by the Author in March 2008																	
Average	4.20	2.79	1.51	0.47	0.09	44.45	19.42	0.14	65.85	22.37	21.27	1.62	0.78	95.96	0.08	0.49	
Standard Deviation	1.30	1.08	0.06	0.08	0.03	1.73	0.67	0.03	2.20	0.59	0.78	0.21	0.04	0.92	0.00	0.01	
Max	5.68	3.91	1.61	0.55	0.12	46.27	20.32	0.18	68.79	23.09	22.18	1.83	0.84	97.27	0.08	0.49	
Min	2.63	1.35	1.48	0.39	0.06	42.64	18.83	0.12	63.85	21.65	20.40	1.40	0.74	95.15	0.08	0.48	

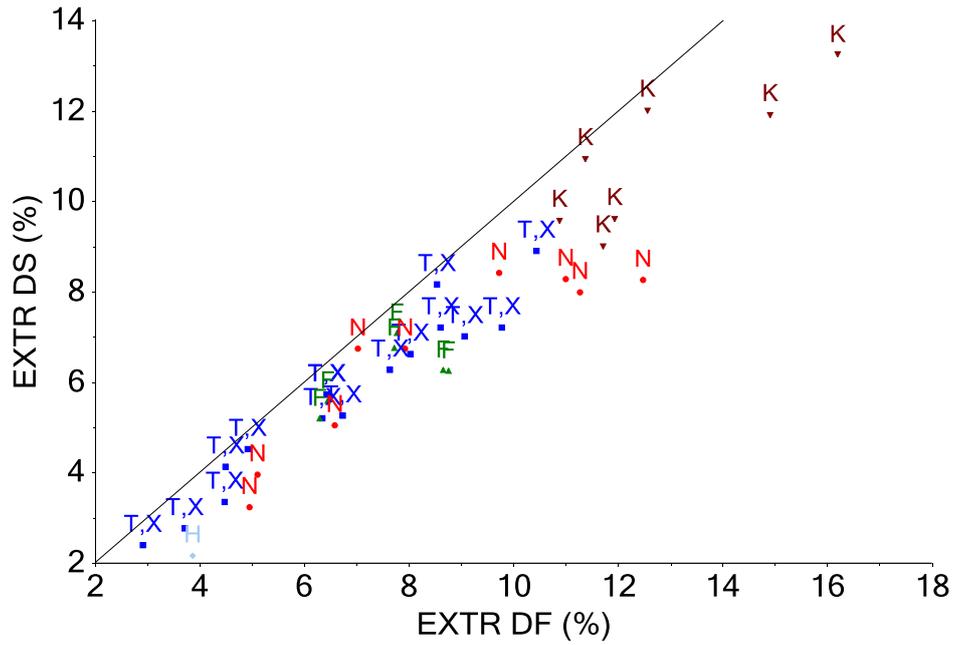


Figure G-19: 95% ethanol soluble extractives (EXTR) content for the DS samples versus the content for the DF samples. T,X = Internodes/whole stems samples, N = nodes, K = live leaf blades, F = dead leaf blades, H = dead leaf sheath.

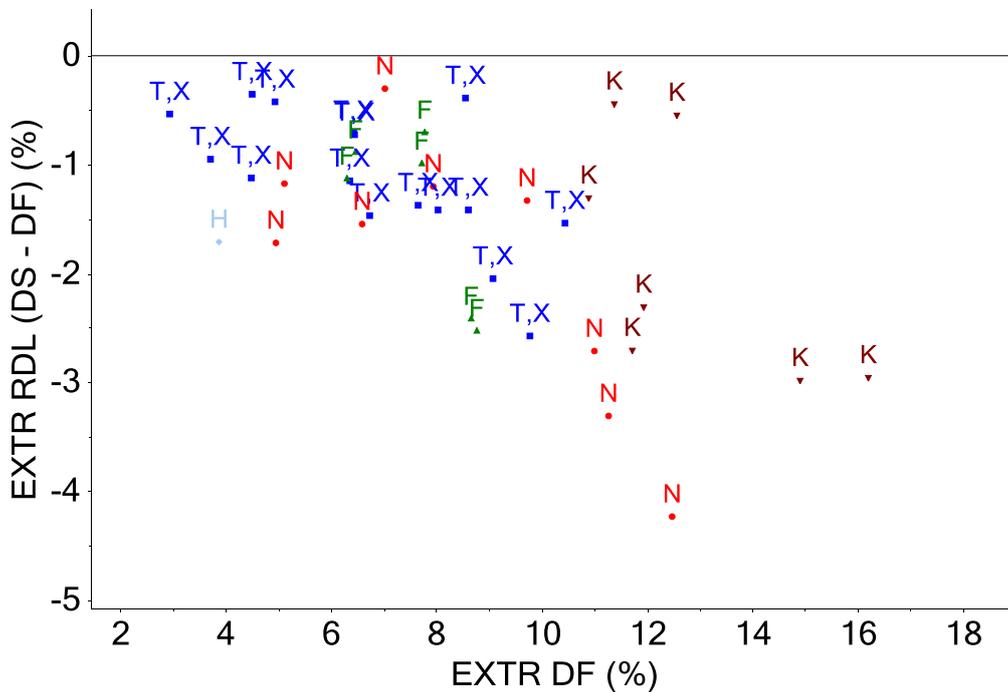


Figure G-20: The residual 95% ethanol soluble extractives (EXTR) content, determined as the DS content minus the DF content, versus the content for the DF samples. T,X = Internodes/whole stems samples, N = nodes, K = live leaf blades, F = dead leaf blades, H = dead leaf sheath.

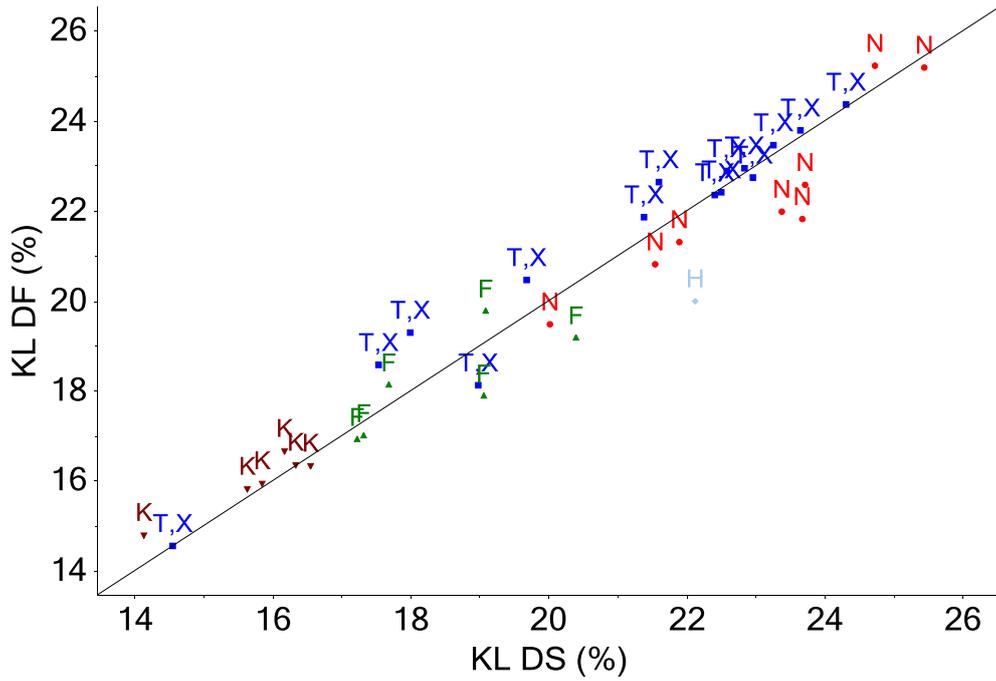


Figure G-21: Klason lignin (KL) content for the DS samples versus the content for the DF samples. T,X = Internodes/whole stems samples, N = nodes, K = live leaf blades, F = dead leaf blades, H = dead leaf sheath.

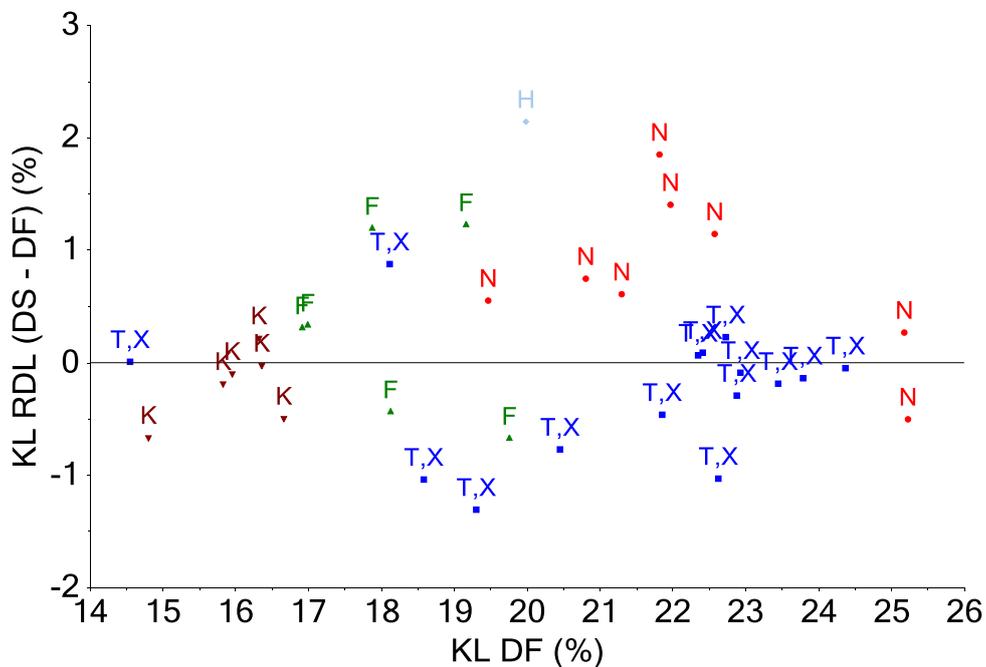


Figure G-22: The residual Klason lignin (KL) content, determined as the DS content minus the DF content, versus the content for the DF samples. T,X = Internodes/whole stems samples, N = nodes, K = live leaf blades, F = dead leaf blades, H = dead leaf sheath.

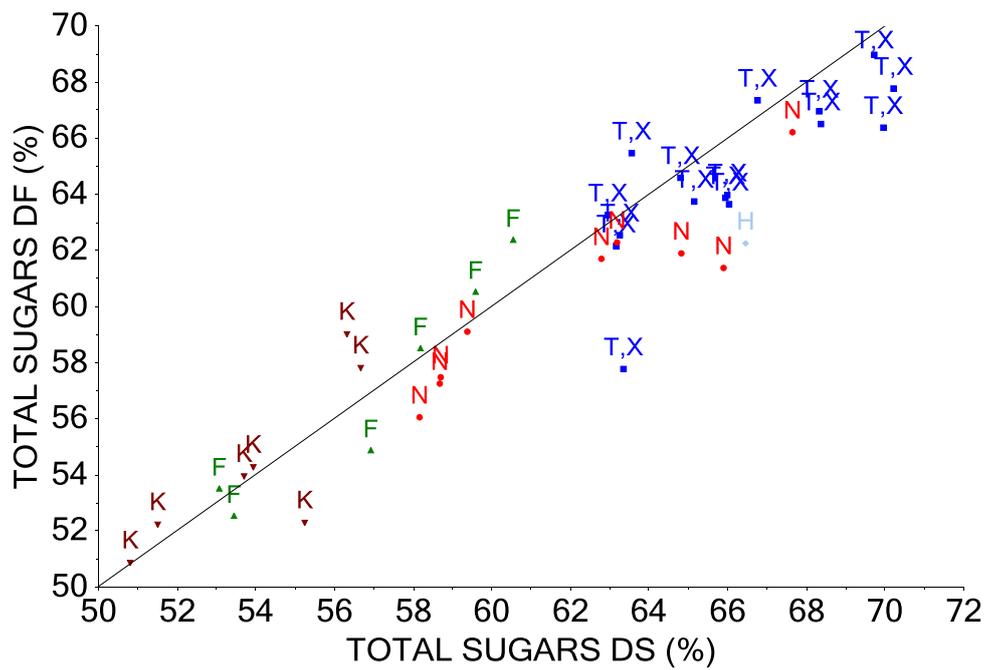


Figure G-23: Total sugars content for the DS samples versus the content for the DF samples. T,X = Internodes/whole stems samples, N = nodes, K = live leaf blades, F = dead leaf blades, H = dead leaf sheath.

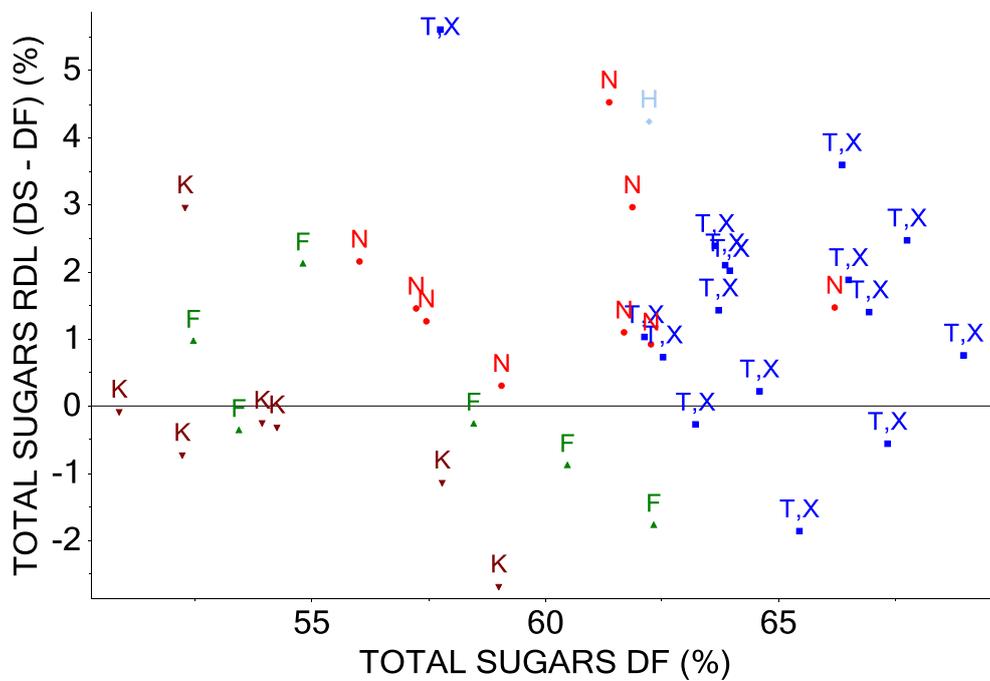


Figure G-24: The residual total sugars content, determined as the DS content minus the DF content, versus the content for the DF samples. T,X = Internodes/whole stems samples, N = nodes, K = live leaf blades, F = dead leaf blades, H = dead leaf sheath.

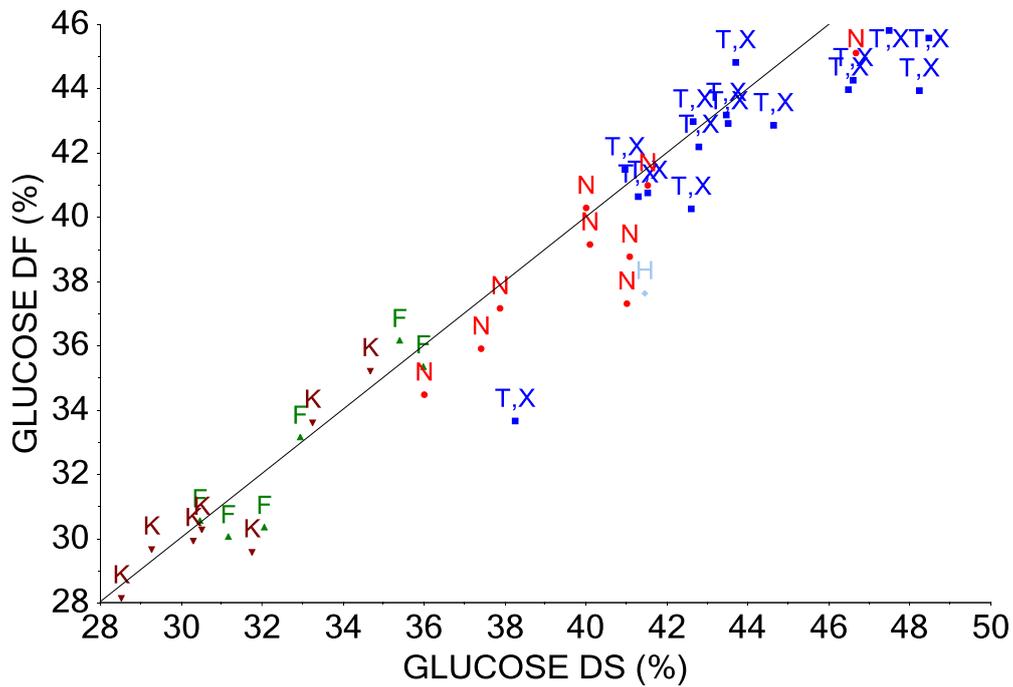


Figure G-25: Glucose content for the DS samples versus the content for the DF samples. T,X = Internodes/whole stems samples, N = nodes, K = live leaf blades, F = dead leaf blades, H = dead leaf sheath.

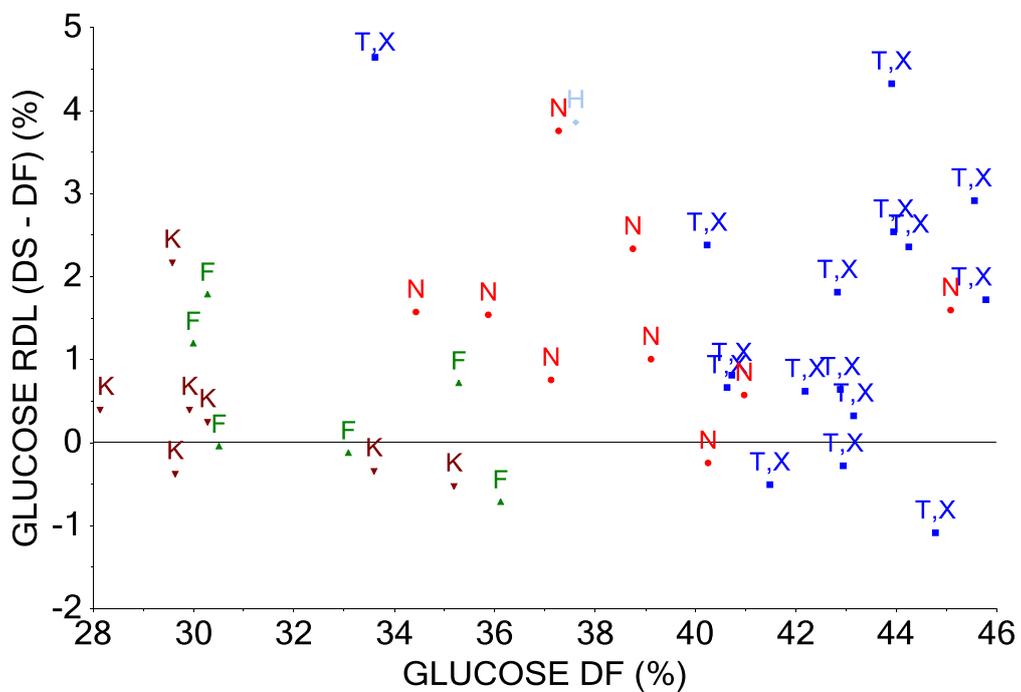


Figure G-26: The glucose residual (RDL), determined as the DS content minus the DF content, versus the content for the DF samples. T,X = Internodes/whole stems samples, N = nodes, K = live leaf blades, F = dead leaf blades, H = dead leaf sheath.

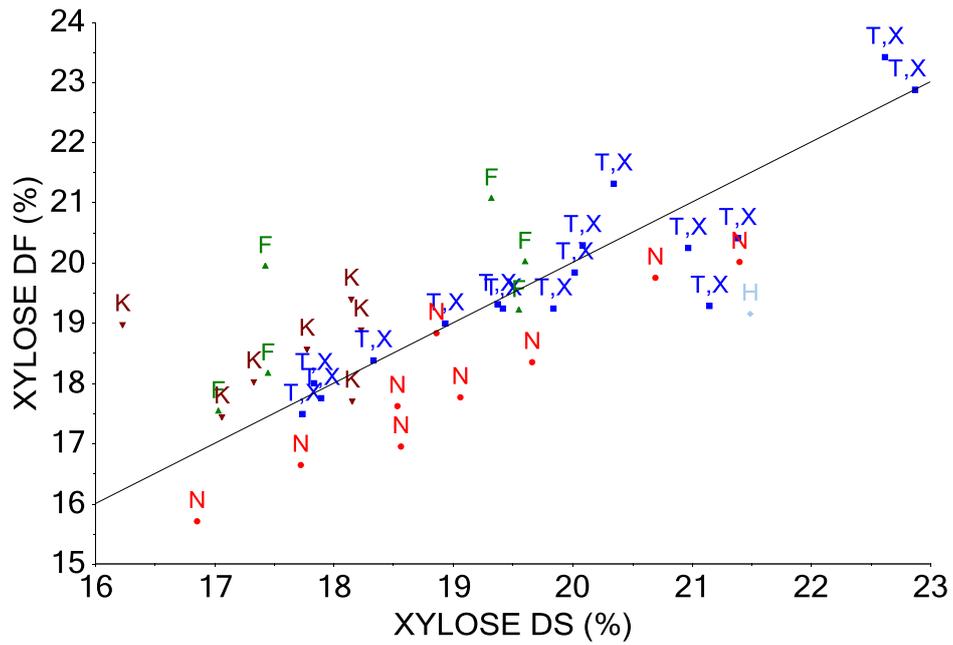


Figure G-27: Xylose content for the DS samples versus the content for the DF samples. T,X = Internodes/whole stems samples, N = nodes, K = live leaf blades, F = dead leaf blades, H = dead leaf sheath.

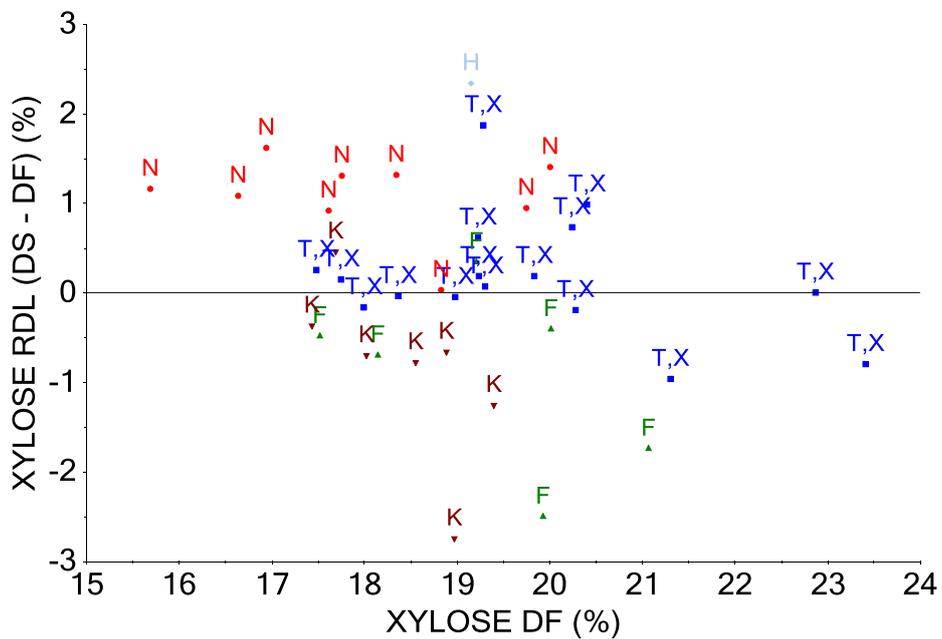


Figure G-28: The xylose residual (RDL), determined as the DS content minus the DF content, versus the content for the DF samples. T,X = Internodes/whole stems samples, N = nodes, K = live leaf blades, F = dead leaf blades, H = dead leaf sheath.

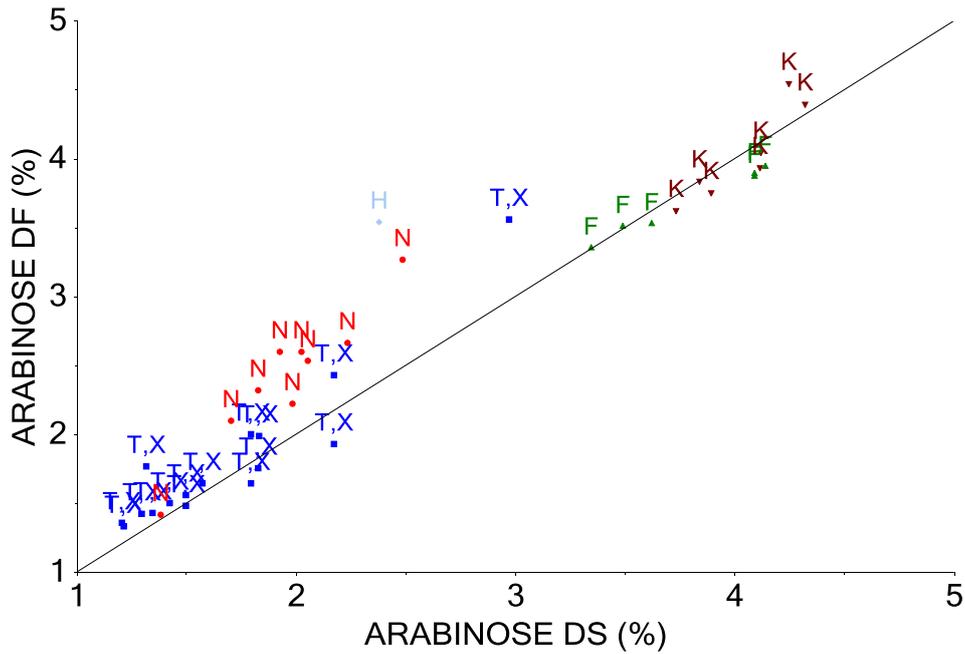


Figure G-29: Arabinose content for the DS samples versus the content for the DF samples. T,X = Internodes/whole stems samples, N = nodes, K = live leaf blades, F = dead leaf blades, H = dead leaf sheath.

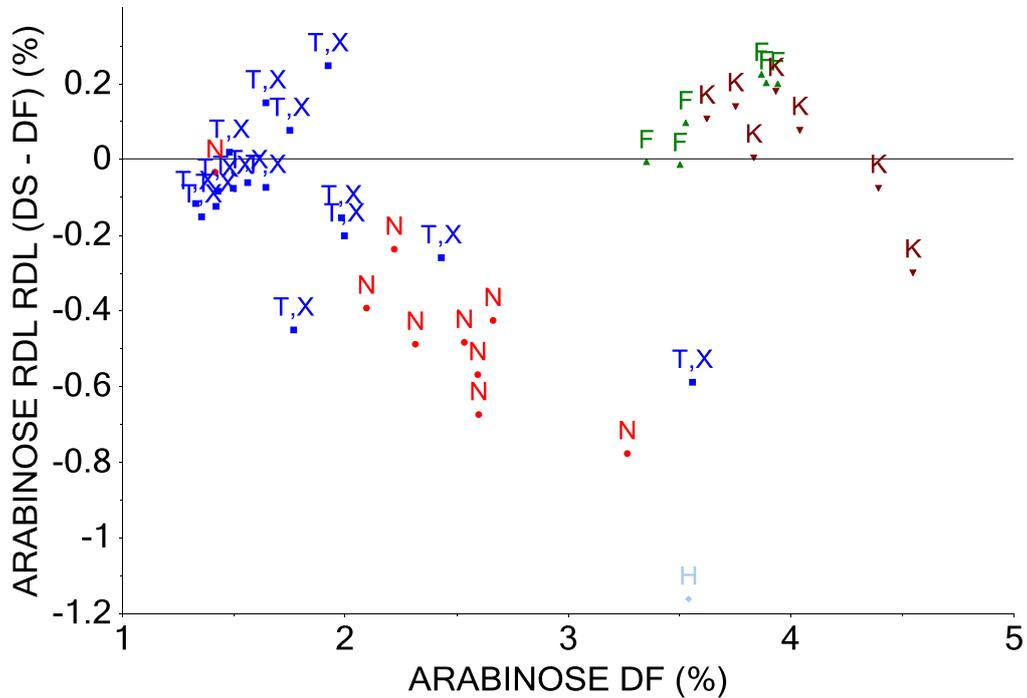


Figure G-30: The arabinose residual (RDL), determined as the DS content minus the DF content, versus the content for the DF samples. T,X = Internodes/whole stems samples, N = nodes, K = live leaf blades, F = dead leaf blades, H = dead leaf sheath.

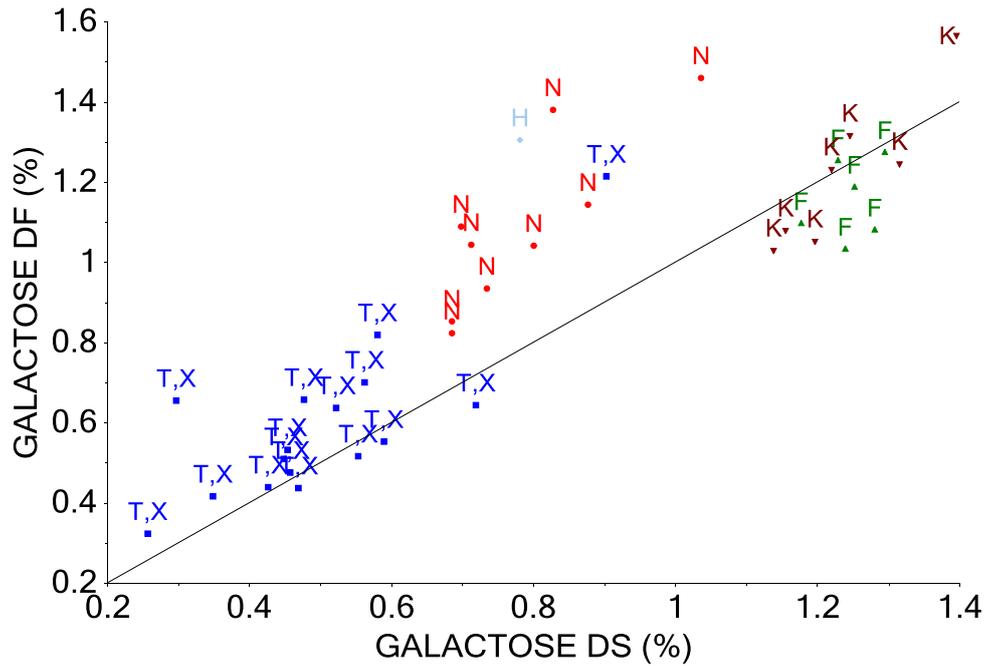


Figure G-31: Galactose content for the DS samples versus the content for the DF samples. T,X = Internodes/whole stems samples, N = nodes, K = live leaf blades, F = dead leaf blades, H = dead leaf sheath.

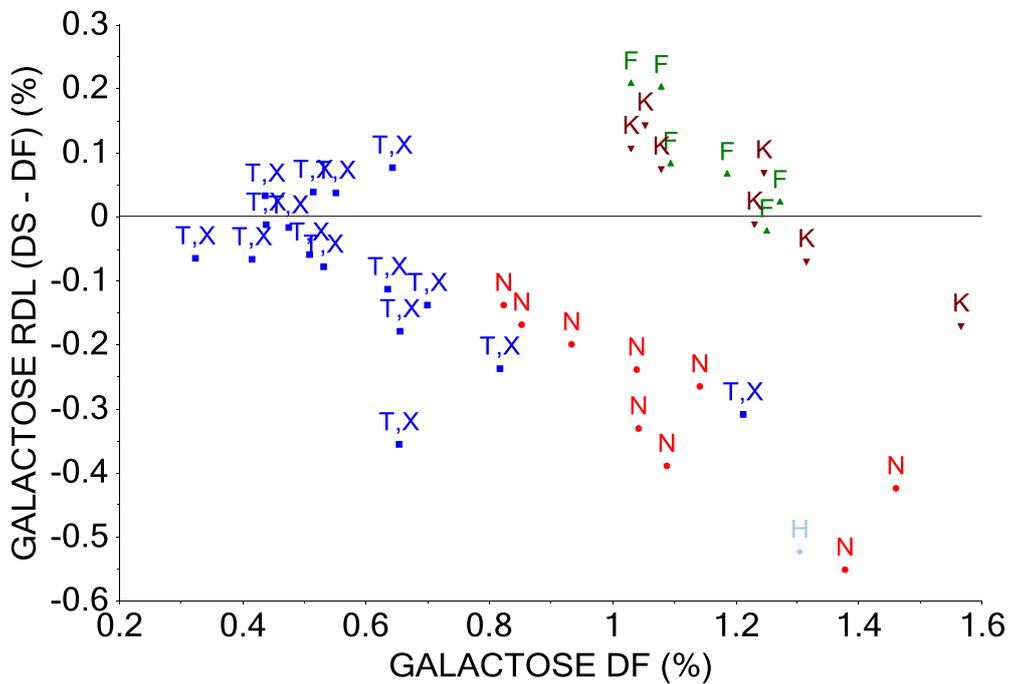


Figure G-32: The galactose residual (RDL), determined as the DS content minus the DF content, versus the content for the DF samples. T,X = Internodes/whole stems samples, N = nodes, K = live leaf blades, F = dead leaf blades, H = dead leaf sheath.

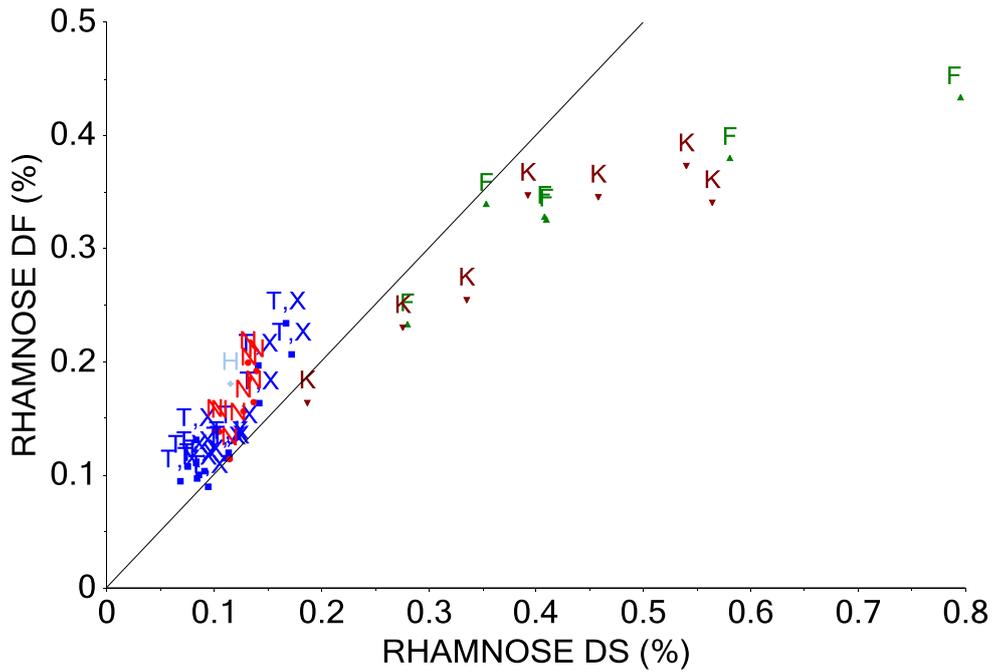


Figure G-33: Rhamnose content for the DS samples versus the content for the DF samples. T,X = Internodes/whole stems samples, N = nodes, K = live leaf blades, F = dead leaf blades, H = dead leaf sheath.

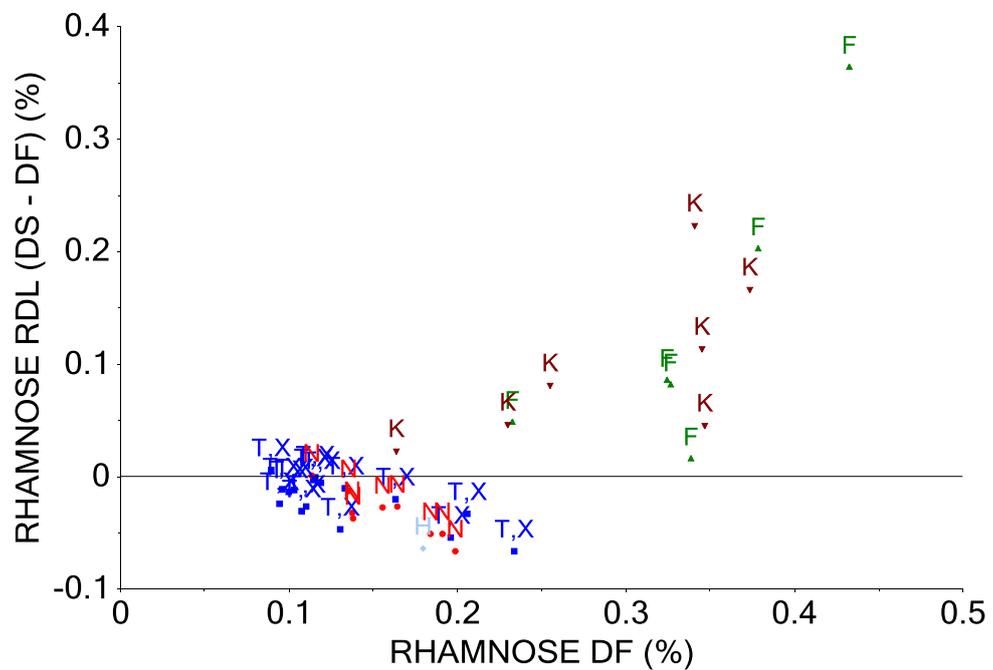


Figure G-34: The galactose residual (RDL), determined as the DS content minus the DF content, versus the content for the DF samples. T,X = Internodes/whole stems samples, N = nodes, K = live leaf blades, F = dead leaf blades, H = dead leaf sheath.

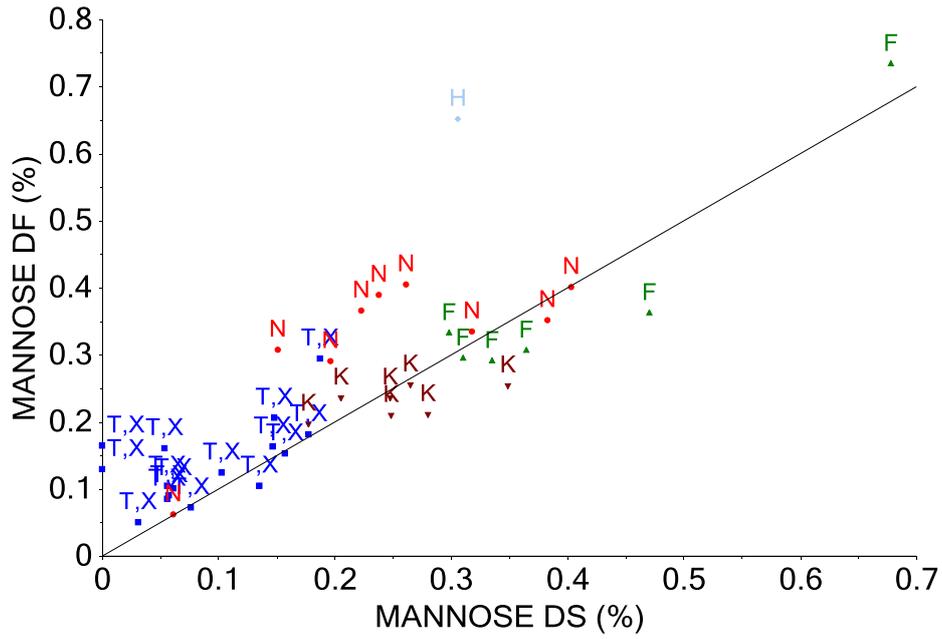


Figure G-35: Mannose content for the DS samples versus the content for the DF samples. T,X = Internodes/whole stems samples, N = nodes, K = live leaf blades, F = dead leaf blades, H = dead leaf sheath.

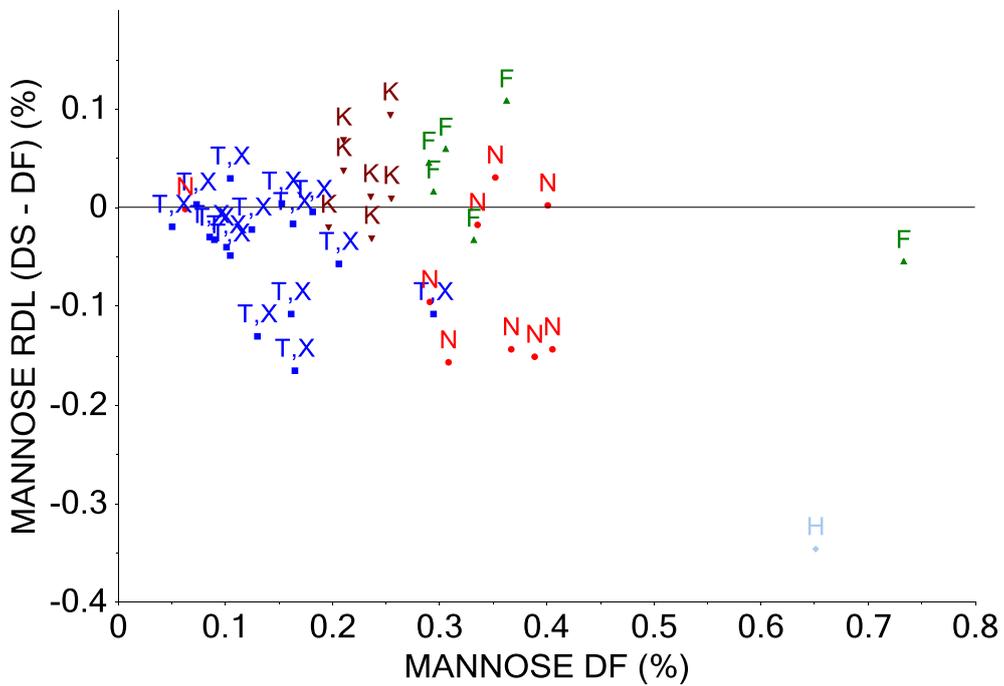


Figure G-36: The mannose residual (RDL), determined as the DS content minus the DF content, versus the content for the DF samples. T,X = Internodes/whole stems samples, N = nodes, K = live leaf blades, F = dead leaf blades, H = dead leaf sheath.

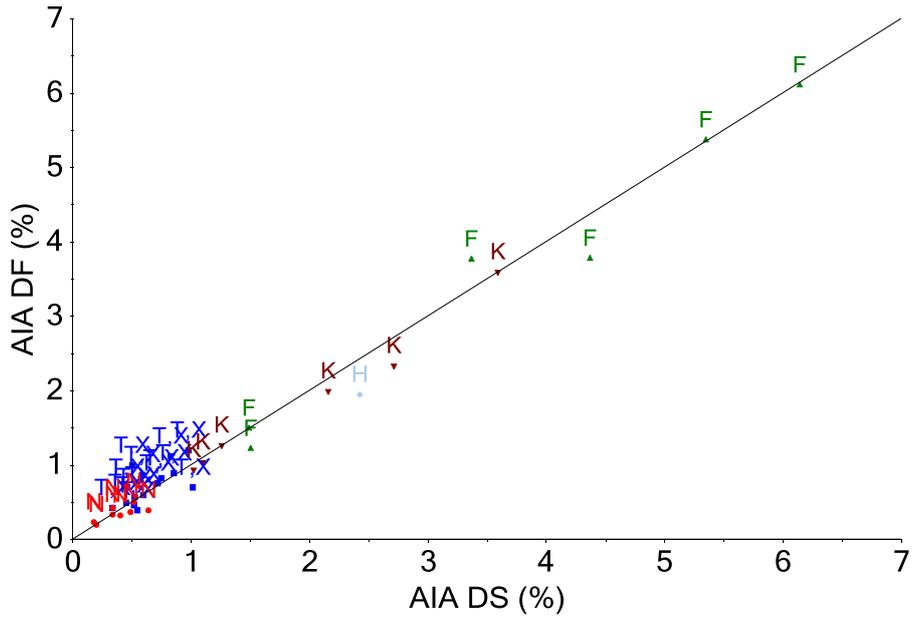


Figure G-37: Acid insoluble ash (AIA) content for the DS samples versus the content for the DF samples. T,X = Internodes/whole stems samples, N = nodes, K = live leaf blades, F = dead leaf blades, H = dead leaf sheath.

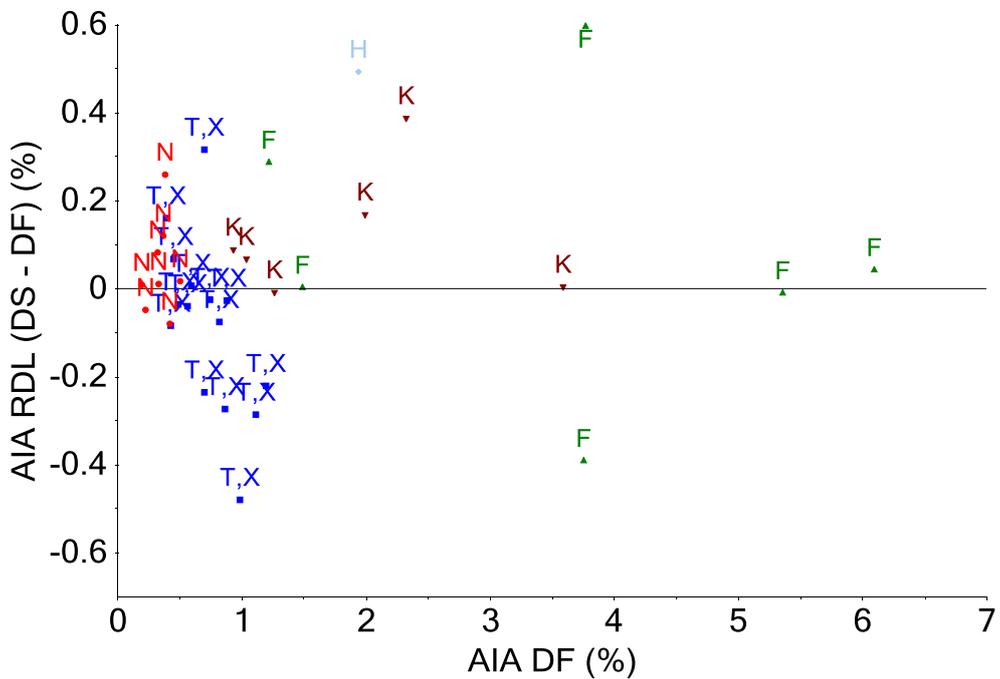


Figure G-38: The acid insoluble ash (AIA) residual (RDL), determined as the DS content minus the DF content, versus the content for the DF samples. T,X = Internodes/whole stems samples, N = nodes, K = live leaf blades, F = dead leaf blades, H = dead leaf sheath.

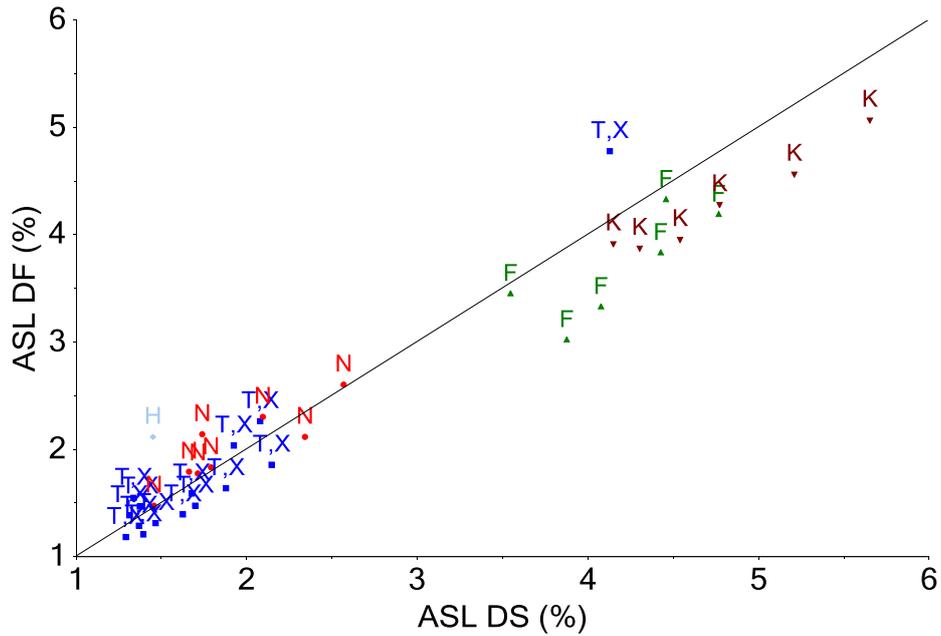


Figure G-39: Acid soluble lignin (ASL) content for the DS samples versus the content for the DF samples. T,X = Internodes/whole stems samples, N = nodes, K = live leaf blades, F = dead leaf blades, H = dead leaf sheath.

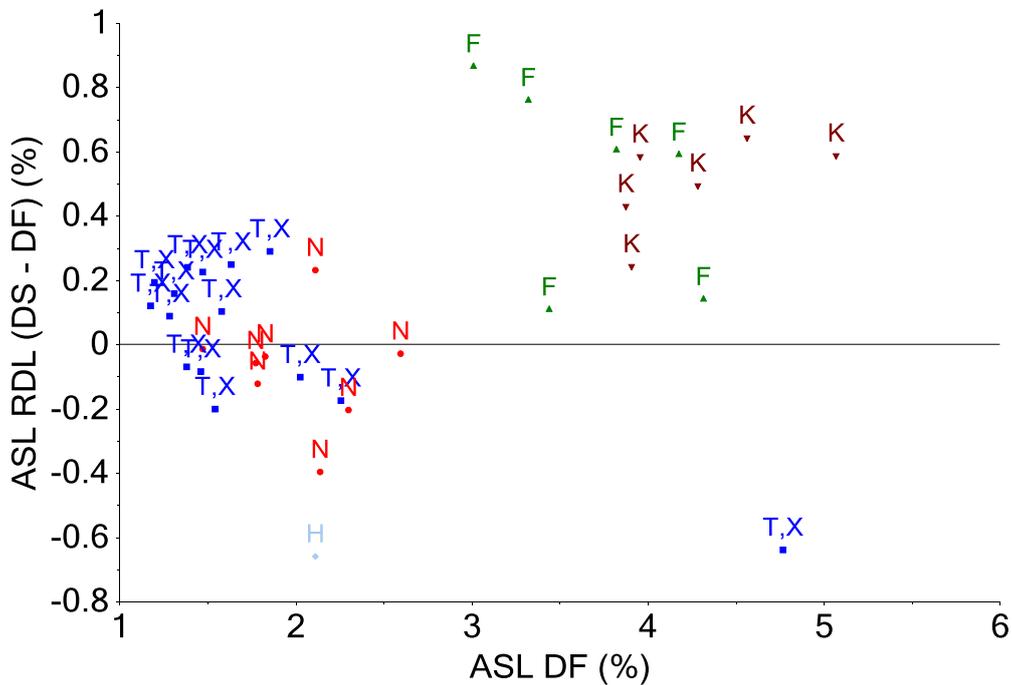


Figure G-40: The acid soluble lignin (ASL) residual (RDL), determined as the DS content minus the DF content, versus the content for the DF samples. T,X = Internodes/whole stems samples, N = nodes, K = live leaf blades, F = dead leaf blades, H = dead leaf sheath.

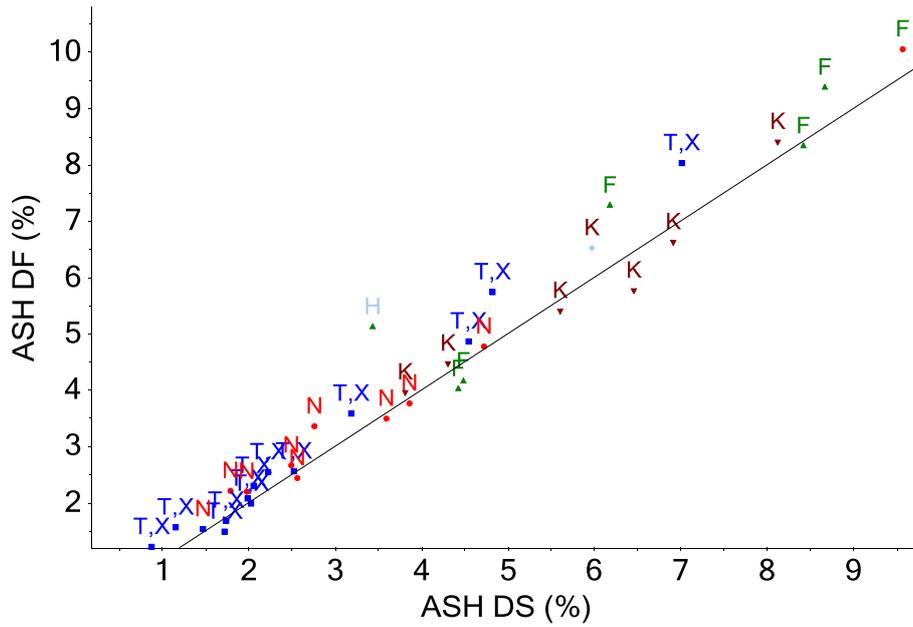


Figure G-41: Ash content for the DS samples versus the content for the DF samples. T,X = Internodes/whole stems samples, N = nodes, K = live leaf blades, F = dead leaf blades, H = dead leaf sheath.

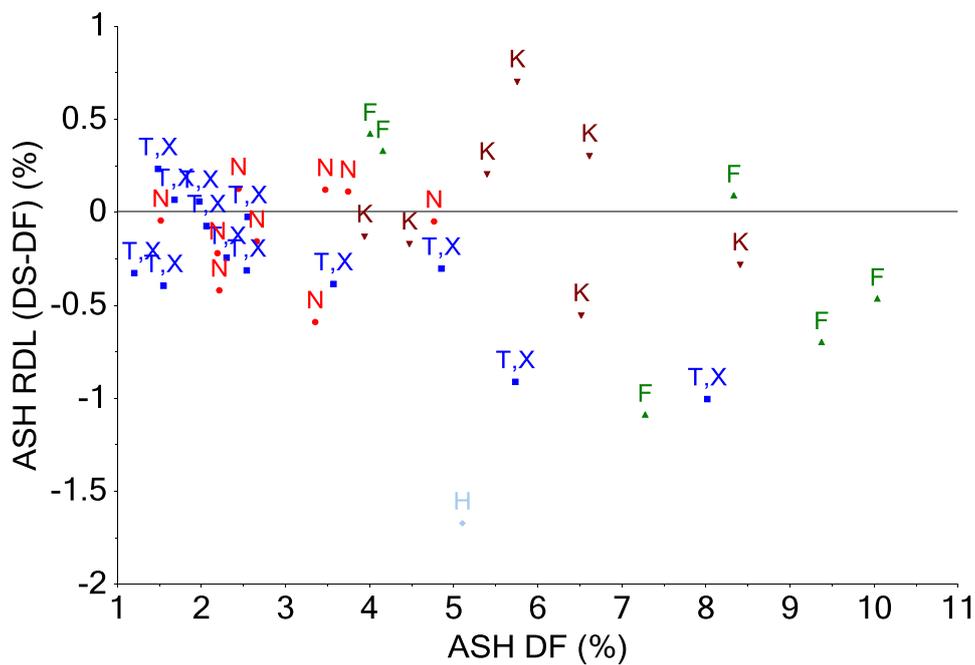


Figure G-42: The ash residual (RDL), determined as the DS content minus the DF content, versus the content for the DF samples. T,X = Internodes/whole stems samples, N = nodes, K = live leaf blades, F = dead leaf blades, H = dead leaf sheath.

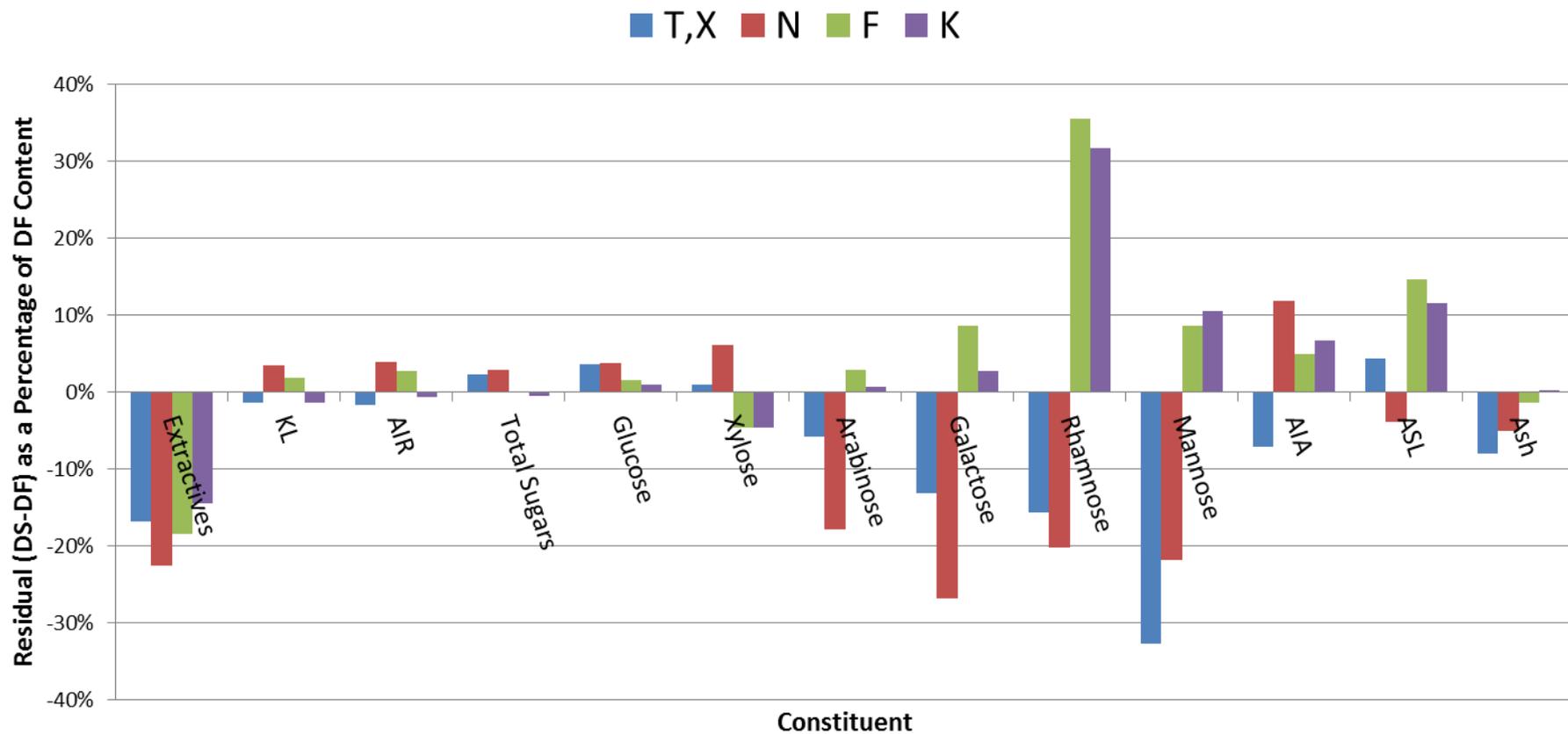
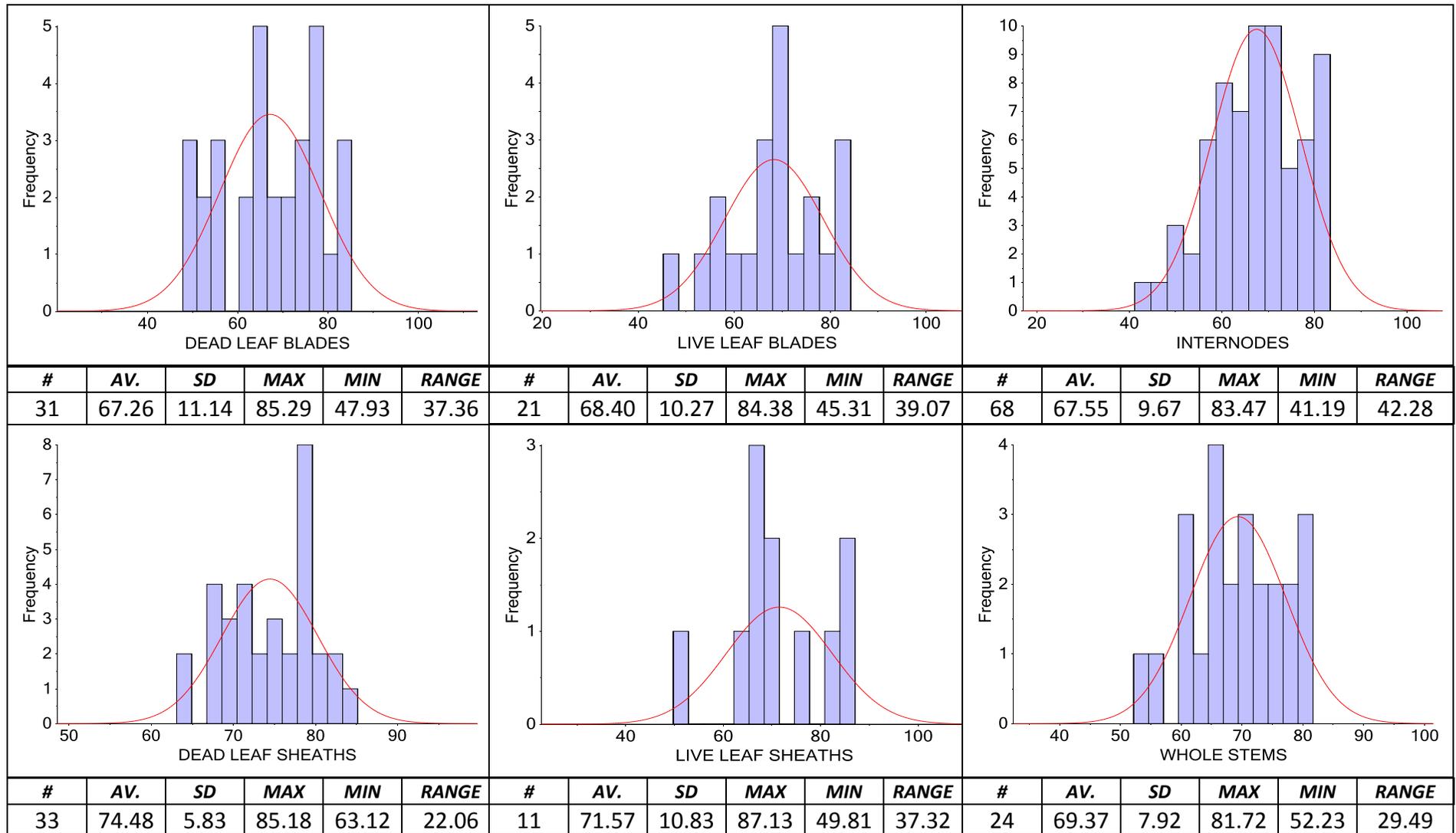
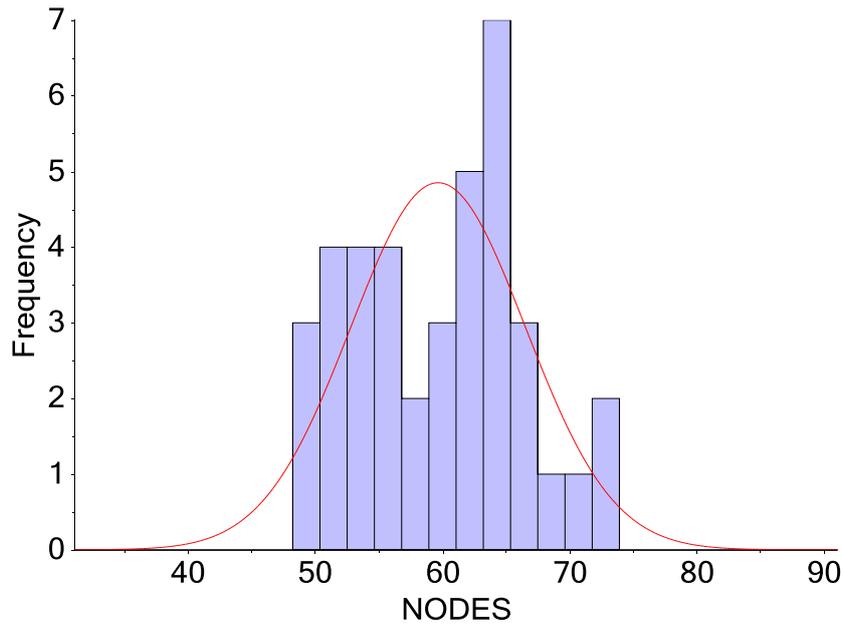


Figure G-43: The average of the residual (DS content minus DF content) expressed as a percentage of the DF content for samples of different plant fractions. T,X = Internodes/whole stems samples, N = nodes, K = live leaf blades, F = dead leaf blades

Table G-18: Histograms, with associated statistics, for the % of total DG material that was DS (the remainder was DF) for the different fractions.





#	AV.	SD	MAX	MIN	RANGE
39	59.67	6.86	73.93	48.24	25.69

Figure G-44: Histogram, with associated statistics, for % of total DG material of the nodes that was DS (the remainder was DF).

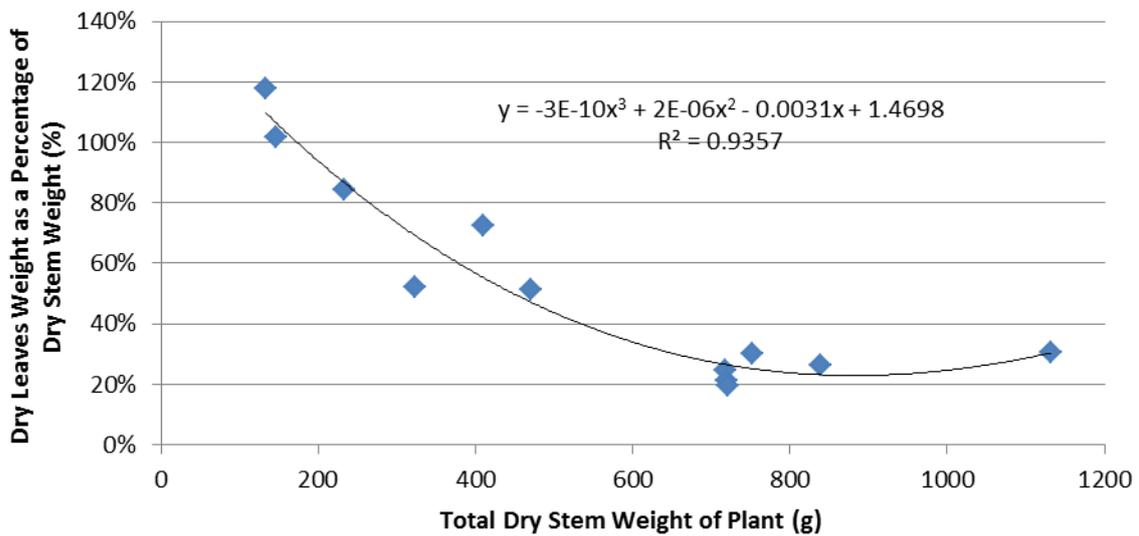


Figure G-45: The amount of total leaves (leaf blades plus leaf sheaths), expressed as a percentage of total dry stem weight, for plants of varying total dry stem weights.

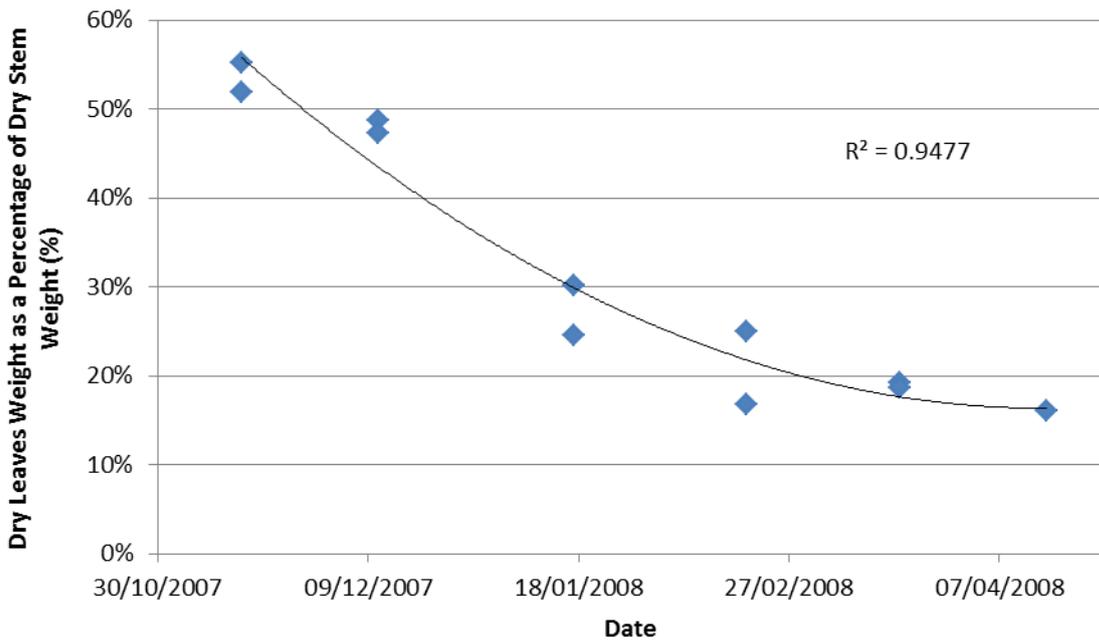


Figure G-46: The amount of total leaves (leaf blades plus leaf sheaths), expressed as a percentage of total dry stem weight, for plants with at least three stem sections that were sampled from the Shanagolden site.

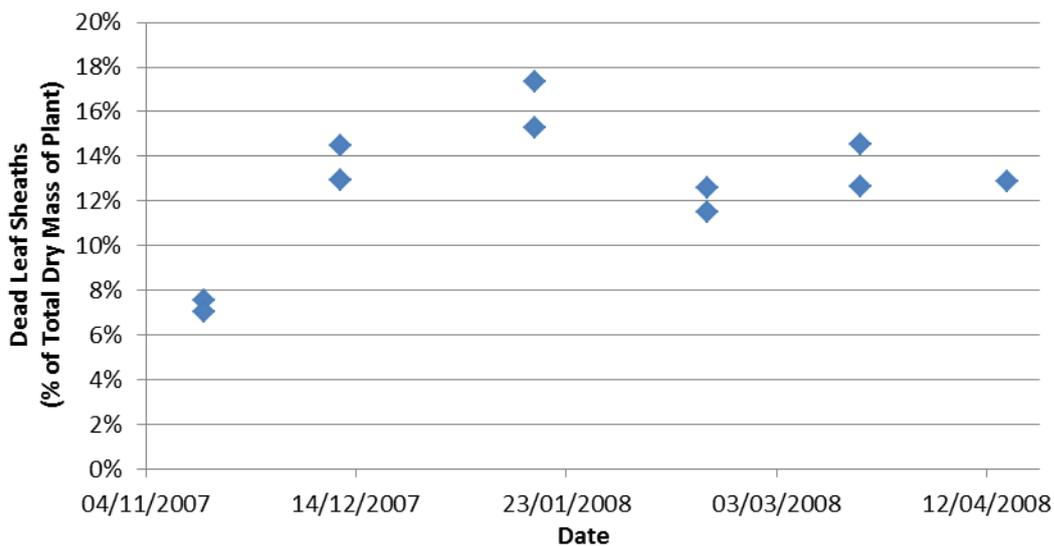


Figure G-47: The amount of dead leaf sheaths, expressed as a percentage of the total dry weight of the plant, for plants with at least three stem sections that were sampled from the Shanagolden site.

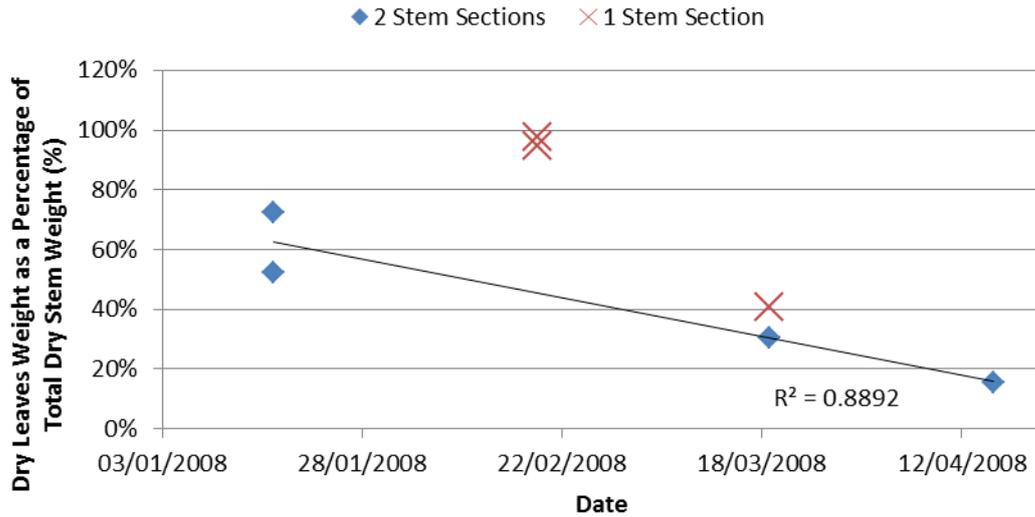


Figure G-48: The amount of total leaves (leaf blades plus leaf sheaths), expressed as a percentage of total dry stem weight, for plants with two stem sections (diamond) and one stem section (crosses), that were sampled from the Shanagolden site.

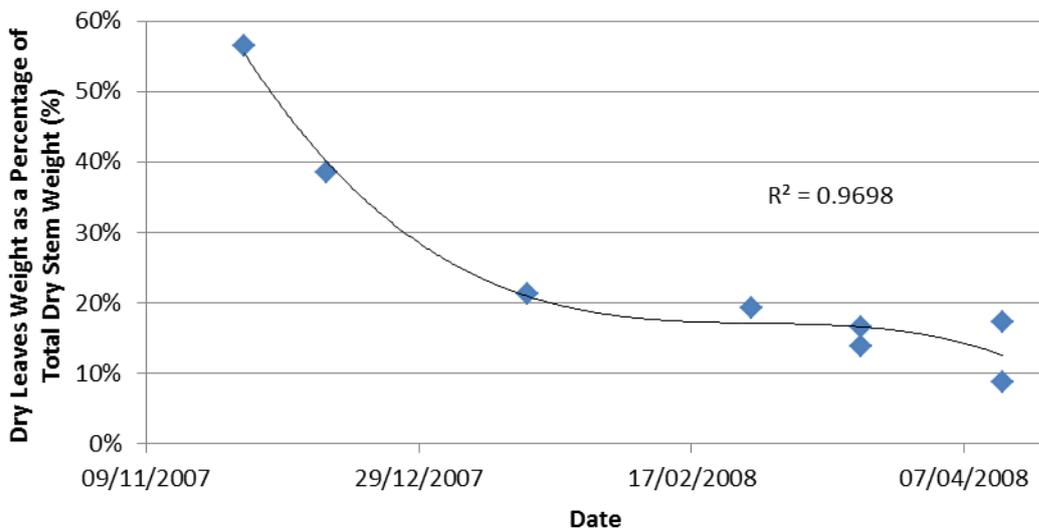


Figure G-49: The amount of total leaves (leaf blades plus leaf sheaths), expressed as a percentage of total dry stem weight, for plants with at least three stem sections that were sampled from the Adare-H site.

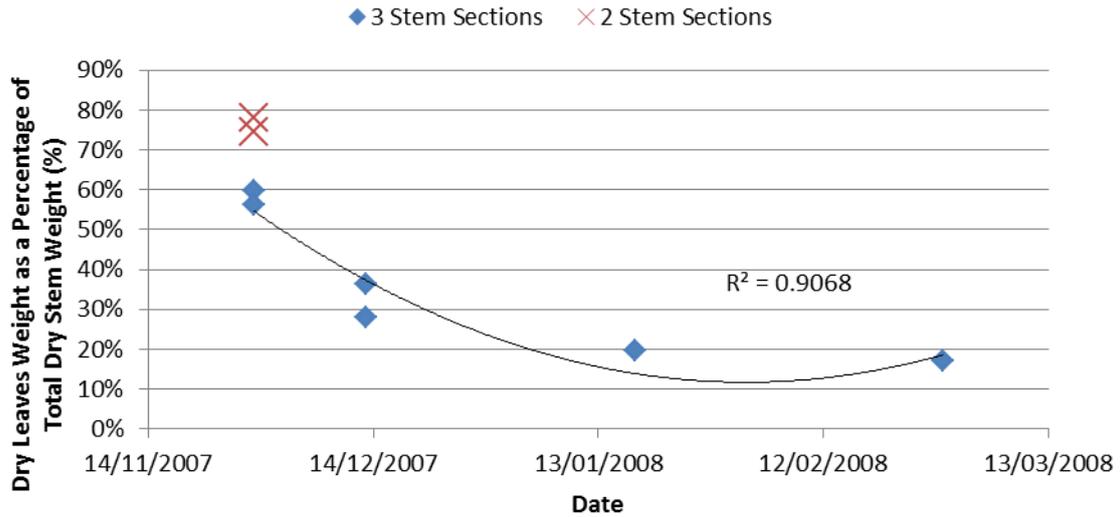


Figure G-50: The amount of total leaves (leaf blades plus leaf sheaths), expressed as a percentage of total dry stem weight, for plants with three stem sections (diamond) and two stem sections (crosses), that were sampled from the Adare-C site.

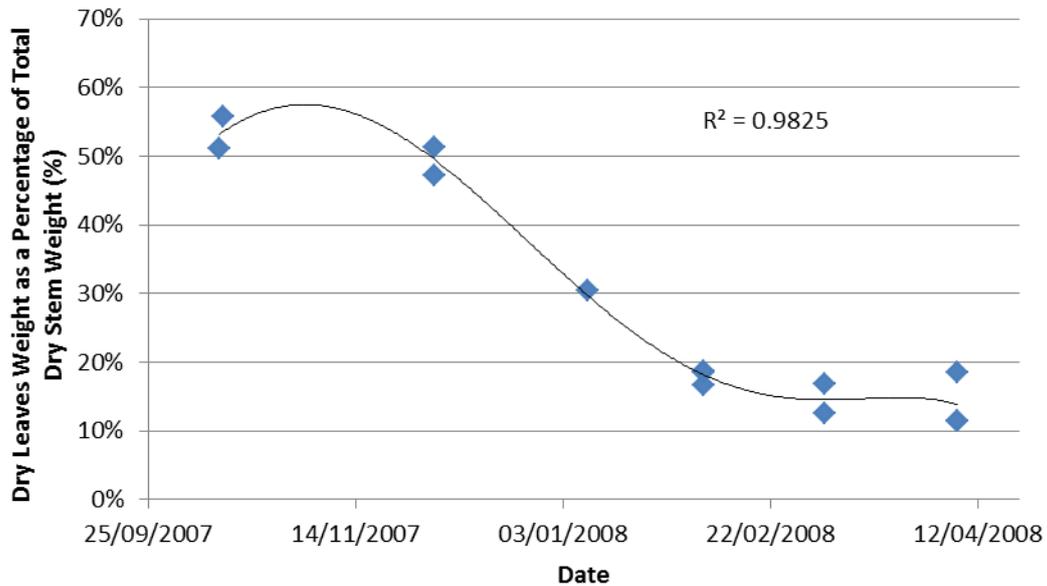


Figure G-51: The amount of total leaves (leaf blades plus leaf sheaths), expressed as a percentage of total dry stem weight for plants with three stem sections that were sampled from the Carlow-F and Carlow-G sites.

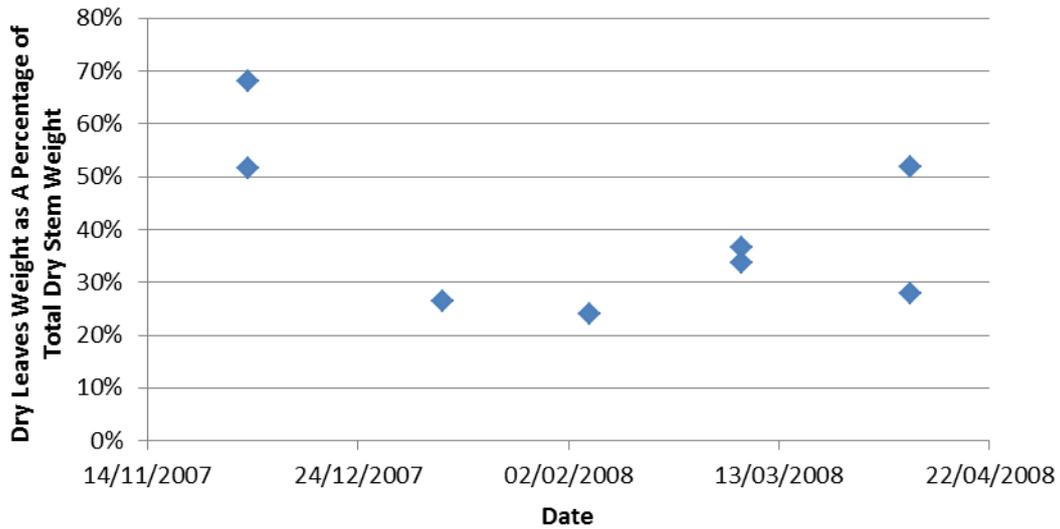


Figure G-52: The amount of total leaves (leaf blades plus leaf sheaths), expressed as a percentage of total dry stem weight, for plants with two stem sections that were sampled from the Carlow-F and Carlow-G sites.

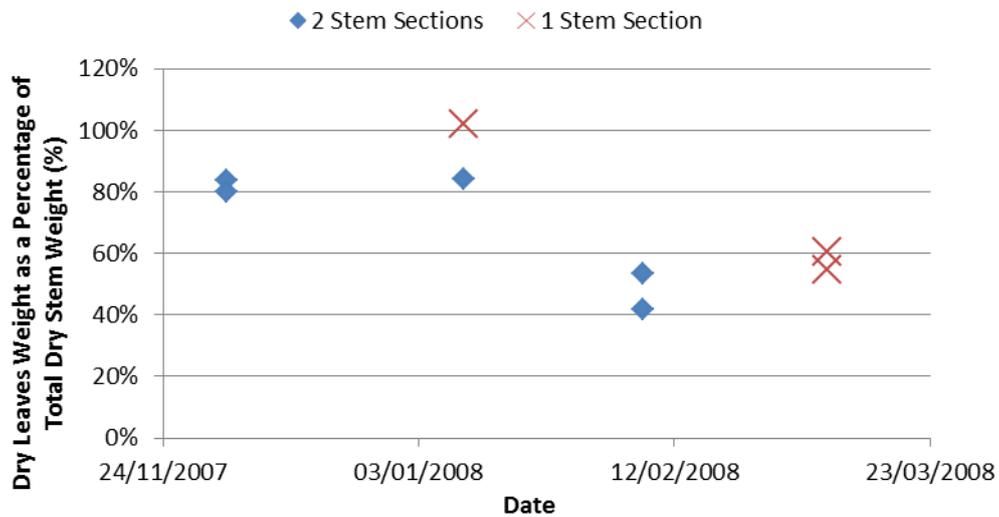


Figure G-53: The amount of total leaves (leaf blades plus leaf sheaths), expressed as a percentage of total dry stem weight, for plants with two stem sections (diamonds) and one stem section (crosses) that were sampled from the Langton site.

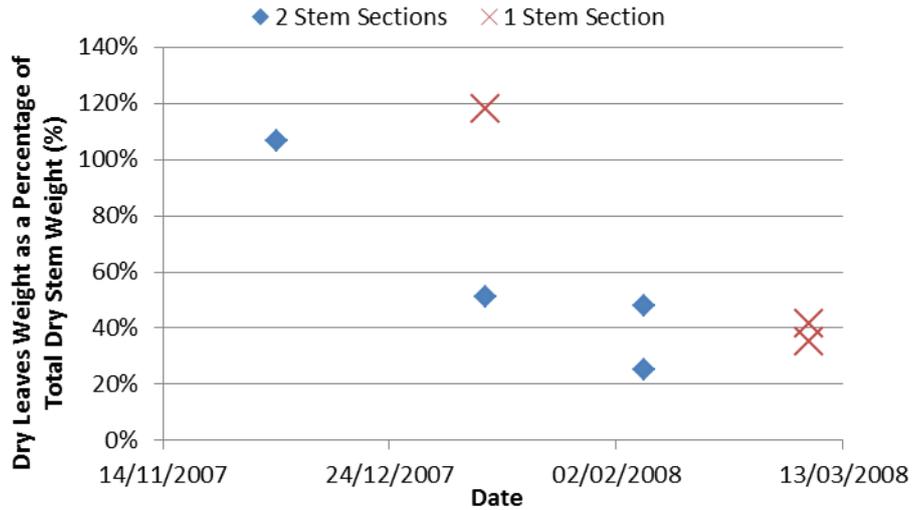


Figure G-54: The amount of total leaves (leaf blades plus leaf sheaths), expressed as a percentage of total dry stem weight, for plants with two stem sections (diamonds) and one stem section (crosses) that were sampled from the Clonmel site.

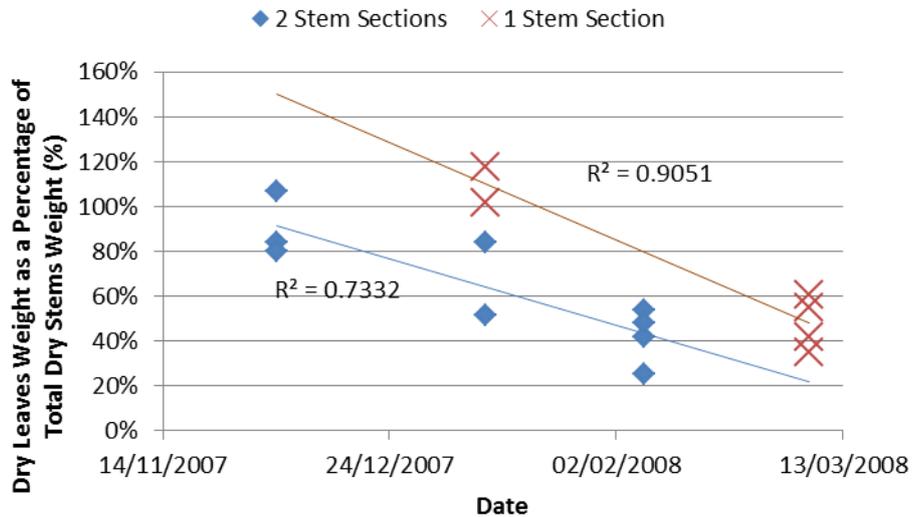


Figure G-55: The amount of total leaves (leaf blades plus leaf sheaths), expressed as a percentage of total dry stem weight, for plants with two stem sections (diamonds) and one stem section (crosses) that were sampled from the Clonmel and Langton sites.

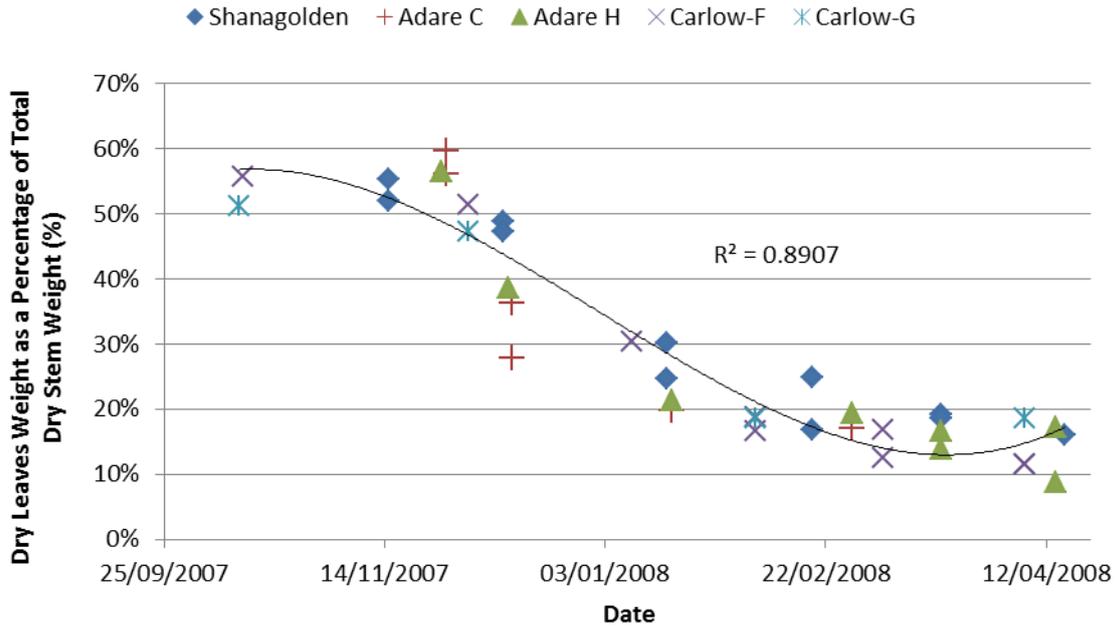


Figure G-56: The amount of total leaves (leaf blades plus leaf sheaths), expressed as a percentage of total dry stem weight, for plants with three stem sections for the Shanagolden, Adare-C, Adare-H, Carlow-F, and Carlow-G sites.

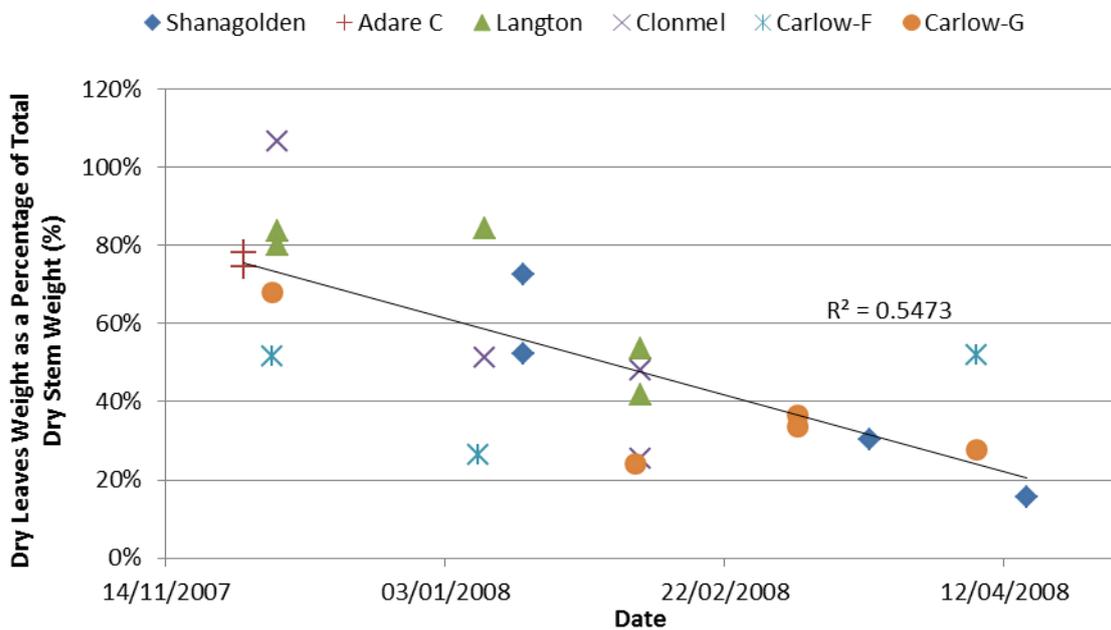


Figure G-57: The amount of total leaves (leaf blades plus leaf sheaths), expressed as a percentage of total dry stem weight, for plants with two stem sections for the Shanagolden, Adare-C, Langton, Clonmel, Carlow-F and Carlow-G sites.

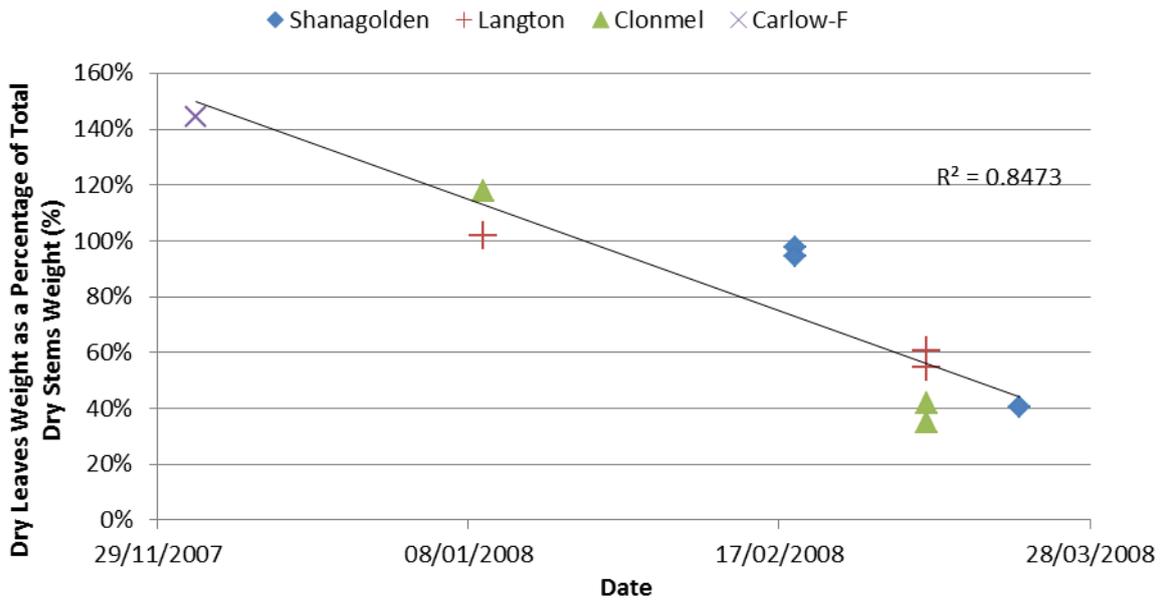


Figure G-58: The amount of total leaves (leaf blades plus leaf sheaths), expressed as a percentage of total dry stem weight, for plants with one stem section for the Shanagolden, Langton, Clonmel and Carlow-F sites.

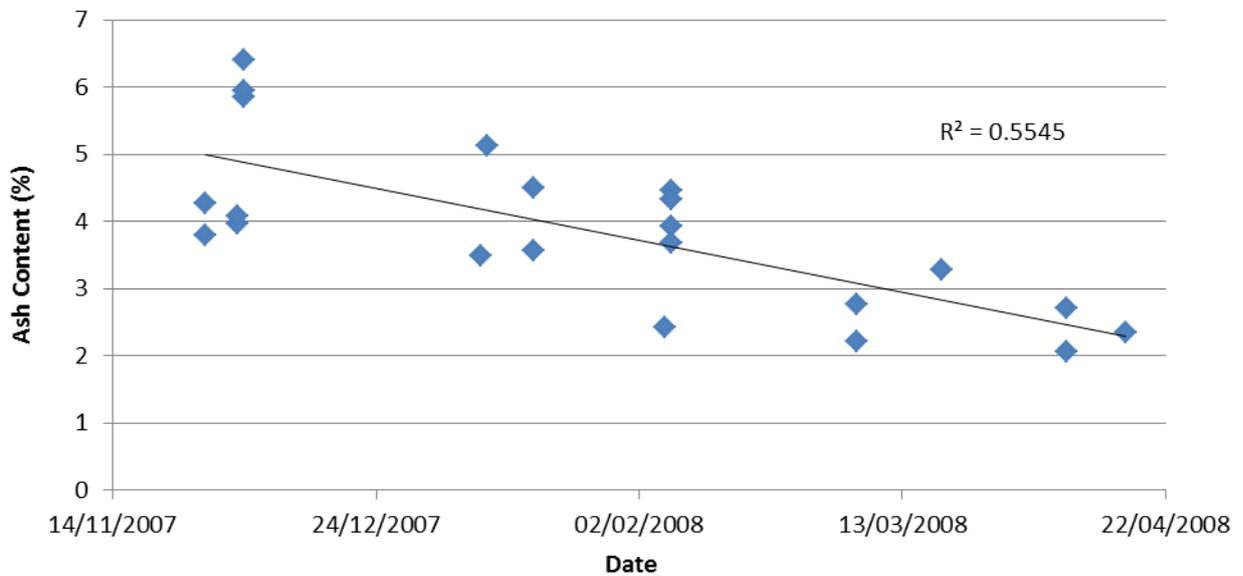


Figure G-59: Relationship between ash content of whole-plant samples and sample collection date for 2-stem-section plants.

Table G-19: R^2 and Pearson correlation coefficients for the relationship between constituent concentration and date of sample collection for Miscanthus whole-plants according to either plant location or whether the plants were less than 1 m high (one stem section), between 1 and 2 m high (2-stem section), or over 2 m high (3 stem-section).

Plant Location or Type	Carlow		Shanagolden		Adare-H		Adare-C		Adare-C and Shanagolden		All 3-Stem Section Plants		All 2-Stem Section Plants		All 1-Stem Section Plants	
	No. of Samples		11		5		6		17		35		22		9	
Constituent	R^2	r	R^2	r	R^2	r	R^2	r	R^2	r	R^2	r	R^2	r	R^2	r
Extractives	0.475	-0.689	0.504	-0.710	0.348	-0.590	0.001	-0.028	0.290	-0.538	0.238	-0.488	0.517	-0.719	0.323	-0.569
Ash	0.382	-0.618	0.555	-0.745	0.575	-0.758	0.322	-0.567	0.543	-0.737	0.417	-0.645	0.554	-0.745	0.318	-0.564
Arabinose	0.750	-0.866	0.670	-0.818	0.340	-0.583	0.185	-0.430	0.606	-0.779	0.623	-0.789	0.451	-0.672	0.206	-0.454
Galactose	0.640	-0.800	0.677	-0.823	0.975	-0.988	0.221	0.470	0.423	-0.651	0.397	-0.630	0.447	-0.668	0.312	-0.558
Rhamnose	0.660	-0.813	0.227	-0.477	0.825	-0.908	0.451	-0.671	0.202	-0.449	0.326	-0.571	0.465	-0.682	0.247	-0.497
Glucose	0.603	0.776	0.757	0.870	0.733	0.856	0.335	0.579	0.701	0.837	0.542	0.737	0.650	0.806	0.490	0.700
Xylose	0.004	0.067	0.153	0.391	0.719	0.848	0.067	-0.258	0.155	0.394	0.065	0.255	0.270	0.519	0.292	0.540
Mannose	0.004	-0.059	0.196	-0.443	0.734	-0.857	0.015	0.122	0.164	-0.405	0.084	-0.290	0.066	-0.258	0.002	-0.044
Total Sugars	0.262	0.512	0.471	0.686	0.902	0.950	0.063	0.250	0.436	0.660	0.314	0.560	0.678	0.823	0.642	0.801
AIR	0.641	0.801	0.770	0.878	0.908	0.953	0.563	0.750	0.660	0.813	0.574	0.757	0.467	0.683	0.335	0.579
Klason Lignin	0.880	0.938	0.793	0.891	0.839	0.916	0.156	0.395	0.665	0.815	0.645	0.803	0.522	0.722	0.501	0.708
ASL	0.675	-0.821	0.771	-0.878	0.856	-0.925	0.769	-0.877	0.770	-0.877	0.626	-0.791	0.587	-0.766	0.711	-0.843
AIA	0.158	-0.397	0.474	-0.688	0.536	-0.732	0.016	-0.127	0.396	-0.630	0.271	-0.521	0.451	-0.671	0.062	-0.249
Nitrogen	0.138	-0.371	0.501	-0.708	0.662	-0.814	0.025	-0.159	0.456	-0.675	0.226	-0.476	0.336	-0.579	0.697	-0.835

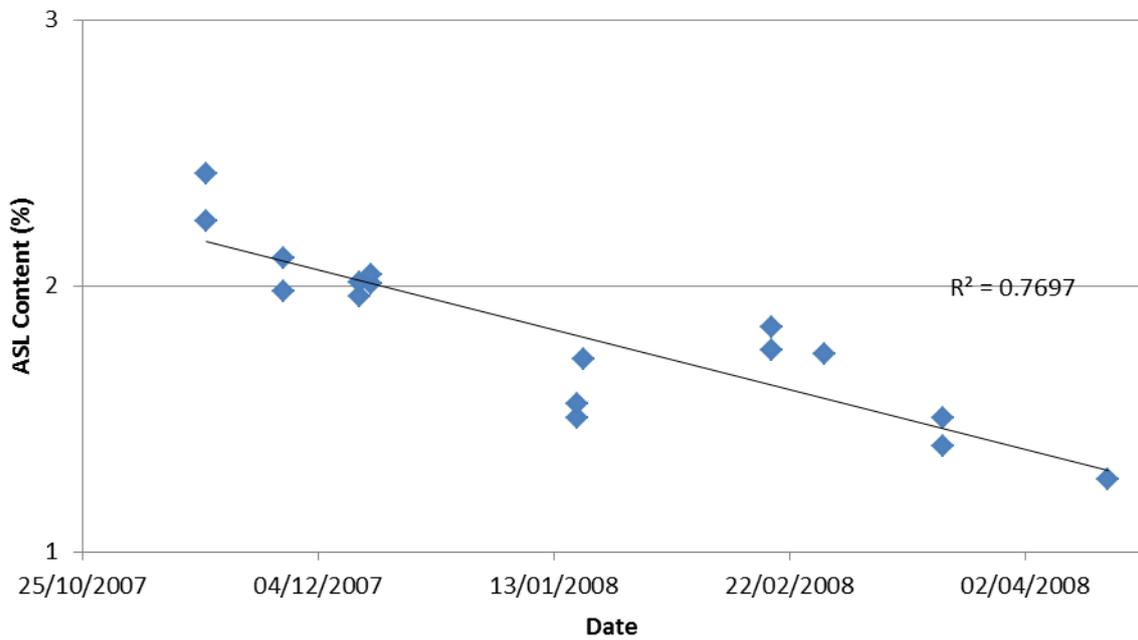


Figure G-60: Relationship between ASL content of whole-plant samples and sample collection date for 3-stem-section plants collected from the Shanagolden and Adare-C sites.

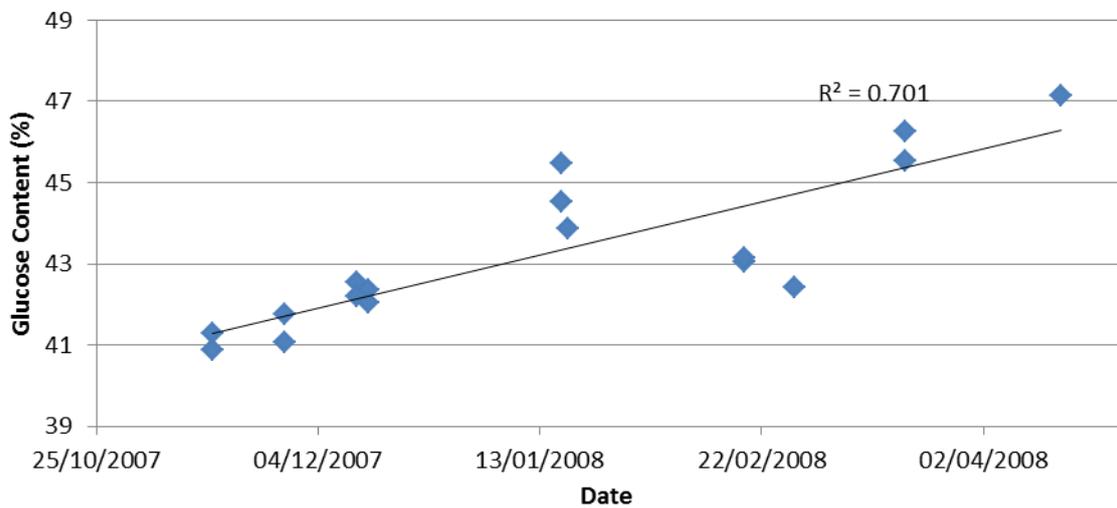


Figure G-61: Relationship between glucose content of whole-plant samples and sample collection date for 3-stem-section plants collected from the Shanagolden and Adare-C sites.

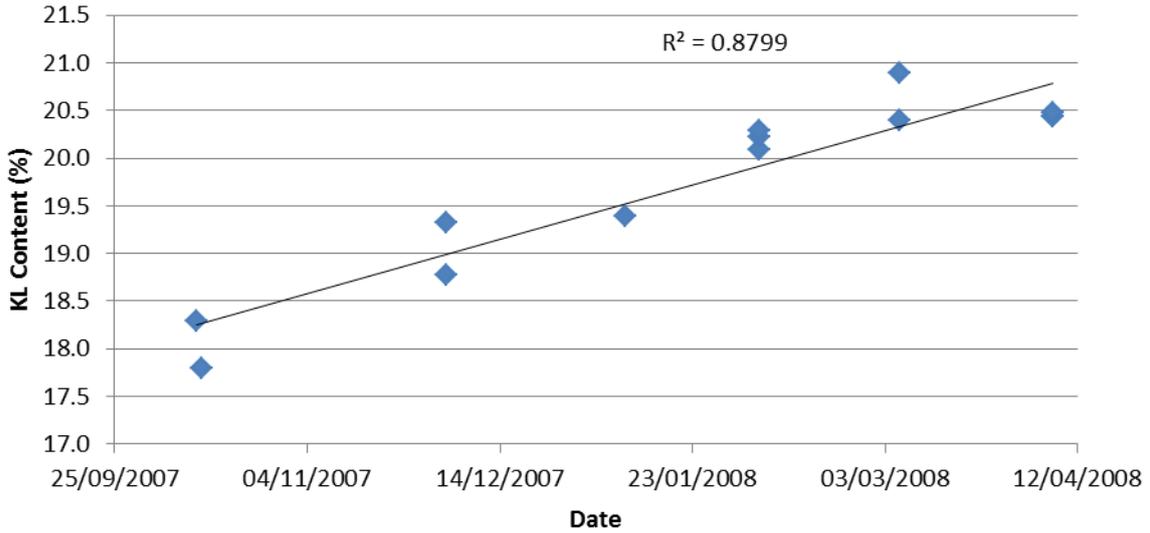


Figure G-62: Relationship between Klason lignin (KL) content of whole-plant samples and sample collection date for 3-stem-section plants collected from the Carlow sites.

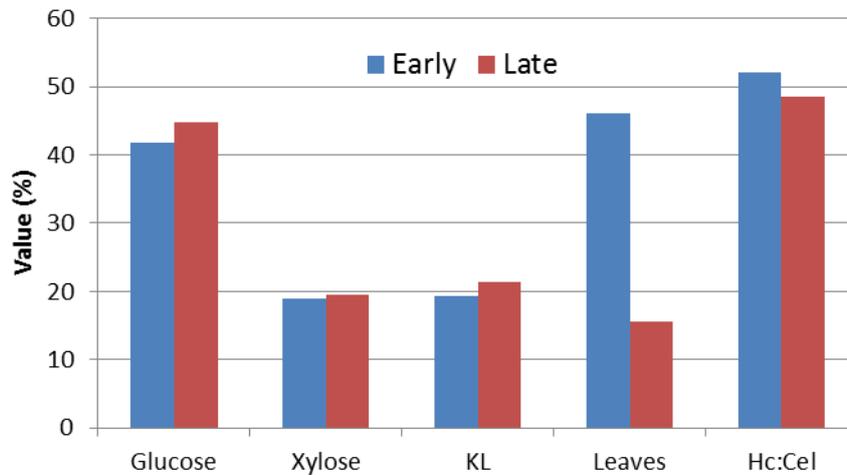


Figure G-63: Comparison of the constituent concentration means for “Early” and “Late” harvest plants. KL = Klason lignin; Leaves = leaves weight (as a percentage of stem weight); Hc:Cel = Hemicellulose content as a percentage of cellulose content. All plants were over 2 m tall (3 stem sections). Early = October, November, December; Late = March, April.

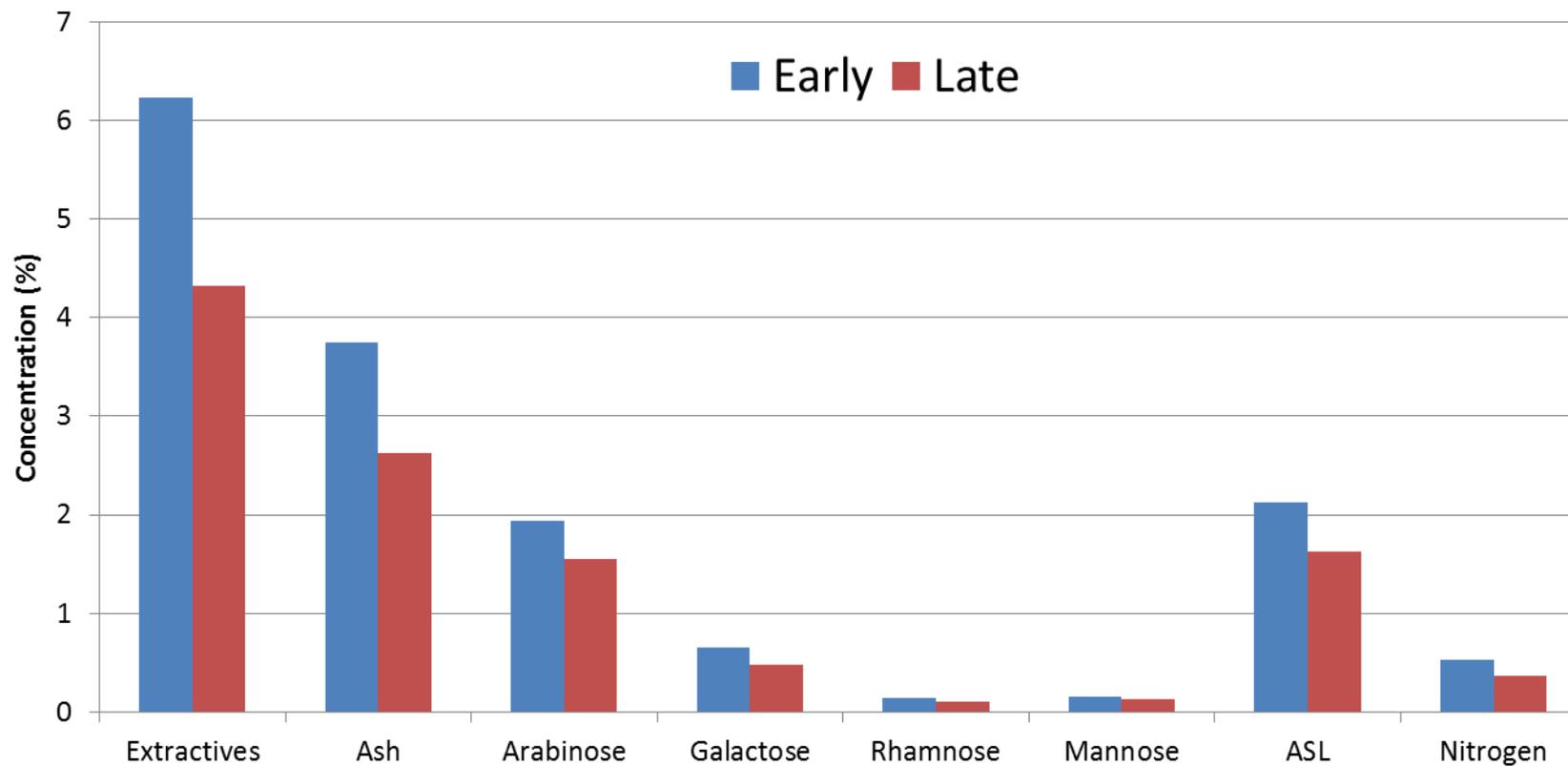


Figure G-64: Comparison of the constituent concentration means for “Early” and “Late” harvest plants. All plants were of the *Miscanthus x giganteus* variety and of a height greater than 2 m (3 stem-sections). ASL = acid soluble lignin. Early = October, November, December; Late = March, April.

Table G-20: R^2 and pearson correlation coefficients for the relationship between constituent concentration and date of sample collection for Miscanthus dead leaf blades according to either plant location or whether the plants were less than 1 m high (one stem section), between 1 and 2 m high (2-stem section), or over 2 m high (3 stem-section). All of the plants used to calculate these values for the Carlow, Shanadolden, and Adare sites had three stem sections.

Plant Location or Type	Carlow		Shanagolden		Adare-H		Adare-C		Adare-C and Shanagolden		All 3-Stem Section Plants		All 2-Stem Section Plants		All 1-Stem Section Plants	
	No. of Samples															
Constituent	R^2	r	R^2	r	R^2	r	R^2	r	R^2	r	R^2	r	R^2	r	R^2	r
Extractives	0.349	0.591	0.294	-0.542	0.457	-0.676	0.516	-0.719	0.364	-0.603	0.001	-0.035	0.089	-0.298	0.143	-0.378
Ash	0.321	-0.566	0.165	-0.406	0.640	0.800	0.070	-0.265	0.041	-0.202	0.055	-0.235	0.171	-0.414	0.097	0.311
Arabinose	0.468	0.684	0.108	0.329	0.323	0.568	0.330	0.575	0.117	0.342	0.229	0.479	0.084	0.289	0.089	0.298
Galactose	0.923	0.961	0.790	0.889	0.885	0.941	0.840	0.916	0.747	0.864	0.799	0.894	0.673	0.821	0.719	0.848
Rhamnose	0.571	-0.756	0.368	-0.606	0.756	0.870	0.369	-0.608	0.146	-0.382	0.118	-0.344	0.254	-0.504	0.576	0.759
Glucose	0.563	-0.750	0.002	0.049	0.037	-0.192	0.060	-0.245	0.002	0.047	0.198	-0.445	0.087	0.295	0.143	-0.379
Xylose	0.052	-0.228	0.000	0.018	0.161	-0.402	0.182	-0.426	0.002	-0.041	0.050	-0.225	0.170	0.412	0.537	-0.733
Mannose	0.701	0.837	0.511	0.715	0.699	0.836	0.804	0.897	0.510	0.714	0.440	0.664	0.323	0.568	0.803	0.896
Total Sugars	0.208	-0.456	0.107	0.326	0.001	0.034	0.007	0.081	0.087	0.295	0.029	-0.172	0.217	0.466	0.121	-0.348
AIR	0.261	-0.511	0.238	0.488	0.346	0.588	0.214	0.463	0.156	0.395	0.003	-0.058	0.022	-0.148	0.307	0.554
Klason Lignin	0.308	-0.555	0.104	0.322	0.272	0.522	0.122	0.350	0.054	0.231	0.027	-0.163	0.001	0.025	0.006	0.077
ASL	0.709	0.842	0.431	-0.656	0.674	-0.821	0.044	0.210	0.239	-0.489	0.002	-0.045	0.087	-0.294	0.010	-0.098
AIA	0.111	-0.333	0.305	0.552	0.607	0.779	0.124	0.353	0.207	0.455	0.012	0.110	0.063	-0.251	0.130	0.360
N	0.906	0.952	0.209	0.457	0.692	-0.832	0.446	0.668	0.043	0.206	0.219	0.468	0.011	-0.103	0.078	0.279

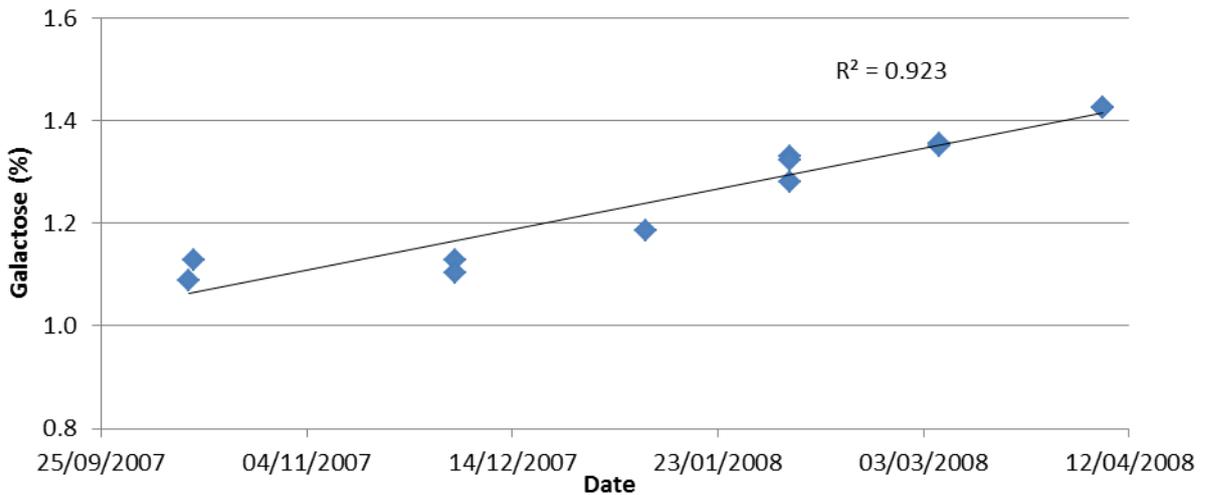


Figure G-65: Relationship between the galactose content of dead leaf blades and sample collection date for the samples of 3-stem-section plants collected from the Carlow sites.

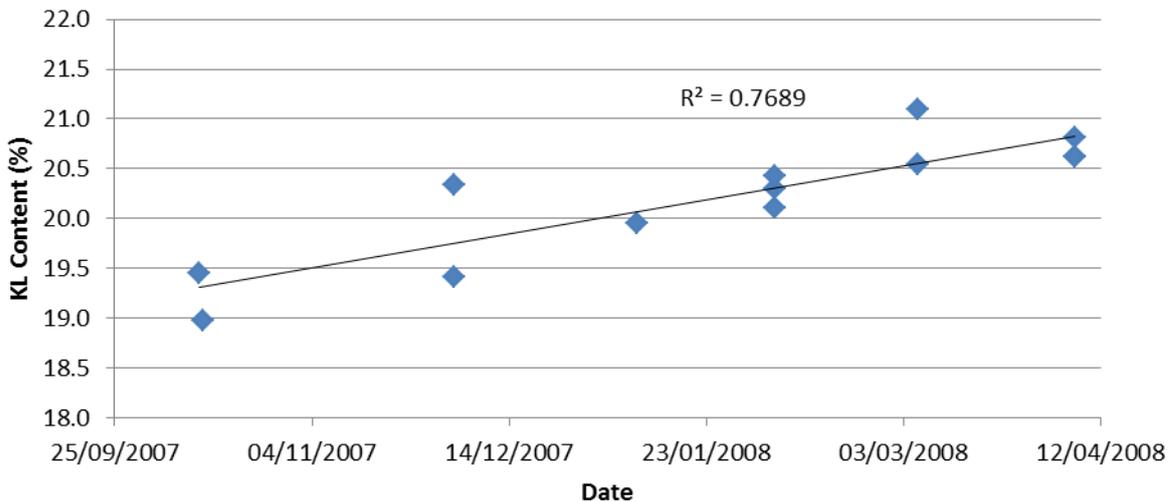


Figure G-66: Relationship between the Klason lignin (KL) content of whole-stems and sample collection date for the samples of 3-stem-section plants collected from the Carlow sites.

Table G-21: Results for ANOVA tests to determine if there are significant differences in the Klason lignin and acid insoluble residue (AIR) means of “Early” and “Late” whole stem samples for *giganteus* plants of differing heights. Note there was only one sample in the “Early” category for 1 m height plants.

Constituent	Plant Height	Mean for “Early”	Mean for “Late”	Test	Degrees of freedom	F Value	Significance of Difference
Klason Lignin	>2 m	20.25	21.72	Welch	11.916	15.885	P < 0.01
AIR	> 2 m	20.90	22.41	ANOVA	21	16.405	P < 0.01
KL	1-2 m	18.45	21.16	Welch	5.878	11.885	P < 0.05
AIR	1-2 m	19.34	21.83	Welch	6.215	13.225	P < 0.05
KL	< 1 m	17.62	20.21	ANOVA	4	6.949	
AIR	< 1 m	18.47	21.42	ANOVA	4	6.061	

Table G-22: R^2 and Pearson correlation coefficients for the relationship between constituent concentration and date of sample collection for *Miscanthus* dead leaf sheaths according to either plant location or whether the plants were less than 1 m high (one stem section), between 1 and 2 m high (2-stem section), or over 2 m high (3 stem-section). All of the plants used to calculate these values for the Carlow, Shanadolden and Adare sites had three stem sections.

Plant Location or Type	Carlow		Shanagolden		Adare-H		Adare-C		Adare-C and Shanagolden		All 3-Stem Section Plants		All 2-Stem Section Plants		All 1-Stem Section Plants	
	No. of Samples		11		5		6		17		35		22		9	
Constituent	R^2	r	R^2	r	R^2	r	R^2	r	R^2	r	R^2	r	R^2	r	R^2	r
Extractives	0.385	-0.620	0.766	-0.875	0.006	0.080	0.038	0.196	0.074	-0.271	0.130	-0.360	0.028	-0.168	0.110	-0.331
Ash	0.213	-0.461	0.420	-0.648	0.660	-0.813	0.079	0.281	0.293	-0.541	0.286	-0.535	0.168	-0.410	0.020	0.141
Arabinose	0.466	-0.682	0.105	-0.324	0.600	0.775	0.046	0.214	0.056	-0.238	0.027	-0.164	0.047	-0.217	0.020	0.142
Galactose	0.143	-0.378	0.382	-0.618	0.256	0.506	0.261	0.511	0.088	-0.297	0.022	-0.147	0.024	0.155	0.614	0.783
Rhamnose	0.031	-0.177	0.000	-0.005	0.909	0.953	0.051	-0.225	0.022	0.148	0.013	0.114	0.041	0.203	0.104	-0.323
Glucose	0.053	0.230	0.467	0.683	0.287	-0.535	0.725	-0.851	0.040	0.201	0.004	0.064	0.017	-0.130	0.014	0.119
Xylose	0.216	0.464	0.207	0.455	0.065	-0.254	0.000	-0.019	0.131	0.362	0.124	0.352	0.004	-0.067	0.007	-0.086
Mannose	0.289	0.538	0.132	-0.363	0.000	0.005	0.398	0.631	0.041	-0.202	0.006	0.079	0.127	0.356	0.270	0.520
Total Sugars	0.079	0.282	0.411	0.641	0.001	-0.029	0.337	-0.580	0.025	0.157	0.040	0.201	0.004	-0.064	0.091	0.301
AIR	0.118	0.343	0.045	0.211	0.069	-0.262	0.232	0.482	0.019	-0.136	0.006	0.075	0.044	0.209	0.003	0.059
Klason Lignin	0.146	0.382	0.367	0.606	0.007	-0.084	0.023	0.151	0.002	0.039	0.051	0.226	0.042	0.205	0.034	-0.185
ASL	0.071	0.266	0.276	-0.525	0.276	0.525	0.922	0.960	0.008	0.088	0.033	0.182	0.002	0.044	0.009	0.093
AIA	0.027	-0.164	0.343	-0.586	0.635	-0.797	0.002	0.048	0.289	-0.538	0.154	-0.392	0.126	-0.355	0.041	0.202
N	0.008	0.087	0.000	0.021	0.081	0.284	0.863	0.929	0.034	0.184	0.025	0.159	0.215	0.464	0.022	0.148

Table G-23: R^2 and Pearson correlation coefficients for the relationship between constituent concentration and date of sample collection for Miscanthus stems according to either plant location or whether the plants were less than 1 m high (one stem section), between 1 and 2 m high (2-stem section), or over 2 m high (3 stem-section). All of the plants used to calculate these values for the Carlow, Shanadolden and Adare sites had three stem sections.

Plant Location or Type	Carlow		Shanagolden		Adare-H		Adare-C		Adare-C and Shanagolden		All 3-Stem Section Plants		All 2-Stem Section Plants		All 1-Stem Section Plants	
	No. of Samples		11		5		6		17		35		22		9	
Constituent	R^2	r	R^2	r	R^2	r	R^2	r	R^2	r	R^2	r	R^2	r	R^2	r
Extractives	0.167	-0.409	0.273	-0.522	0.160	-0.400	0.004	-0.063	0.167	-0.409	0.109	-0.330	0.321	-0.567	0.072	-0.268
Ash	0.172	-0.414	0.098	-0.313	0.398	-0.631	0.025	0.158	0.053	-0.229	0.080	-0.282	0.384	-0.620	0.059	-0.243
Arabinose	0.081	0.285	0.040	-0.200	0.673	0.820	0.815	0.903	0.005	0.067	0.044	0.211	0.071	-0.267	0.001	0.027
Galactose	0.010	-0.102	0.311	-0.558	0.614	-0.783	0.661	0.813	0.087	-0.295	0.031	-0.175	0.257	-0.507	0.154	-0.392
Rhamnose	0.095	-0.309	0.090	0.300	0.880	-0.938	0.168	0.410	0.098	0.313	0.000	-0.020	0.108	-0.328	0.072	-0.268
Glucose	0.133	-0.365	0.218	0.467	0.157	0.396	0.122	-0.349	0.177	0.421	0.003	0.055	0.297	0.545	0.079	0.281
Xylose	0.000	-0.014	0.081	0.285	0.778	0.882	0.000	0.017	0.124	0.352	0.047	0.217	0.055	0.235	0.224	0.474
Mannose	0.106	0.326	0.021	-0.146	0.452	-0.672	0.128	0.357	0.006	-0.079	0.003	0.051	0.045	-0.211	0.004	0.061
Total Sugars	0.026	-0.160	0.079	0.281	0.549	0.741	0.013	-0.116	0.111	0.333	0.009	0.095	0.352	0.594	0.167	0.409
AIR	0.560	0.748	0.673	0.821	0.477	0.691	0.389	0.624	0.603	0.776	0.436	0.660	0.376	0.614	0.389	0.623
Klason Lignin	0.769	0.877	0.604	0.777	0.426	0.653	0.001	-0.023	0.435	0.659	0.359	0.599	0.357	0.598	0.391	0.625
ASL	0.114	0.337	0.185	-0.431	0.099	-0.315	0.073	0.270	0.144	-0.379	0.003	-0.052	0.172	-0.414	0.363	-0.602
AIA	0.059	-0.243	0.196	-0.443	0.318	-0.564	0.342	0.585	0.054	-0.233	0.064	-0.253	0.529	-0.728	0.040	0.201
N	0.004	0.062	0.074	-0.271	0.394	-0.628	0.113	0.336	0.075	-0.275	0.005	-0.068	0.070	-0.265	0.005	-0.073

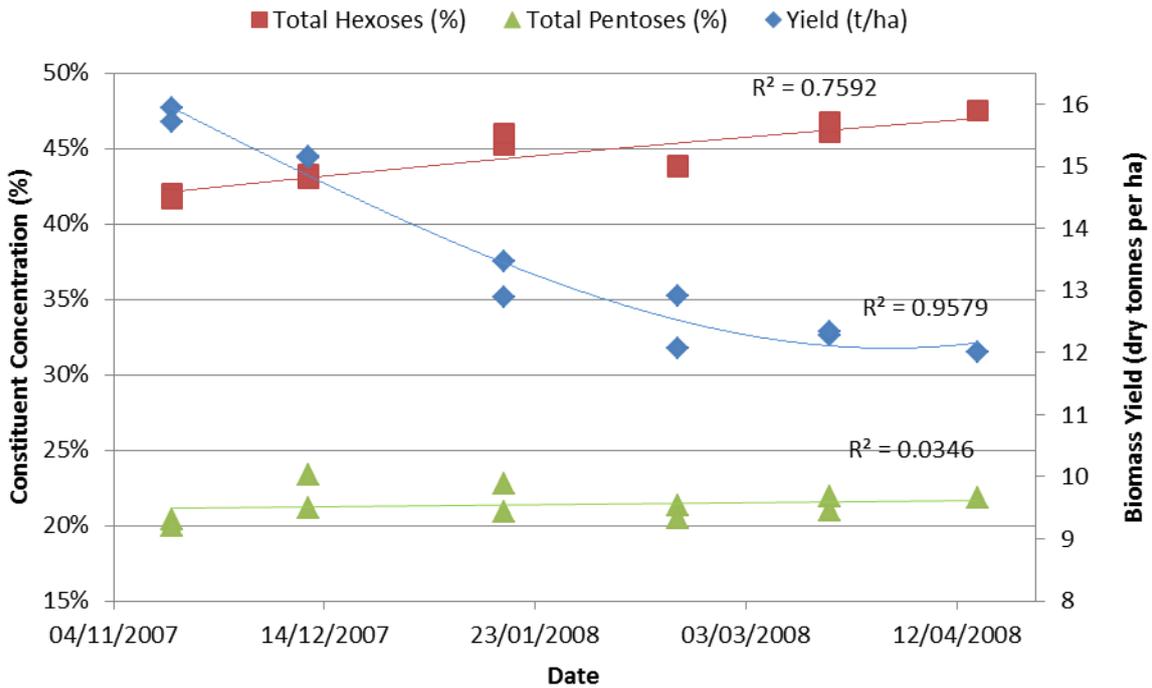


Figure G-67: Miscanthus yield (dry tonnes per hectare), and the total hexose and pentose contents of the feedstock, at various points in the harvest window on the Shanagolden site.

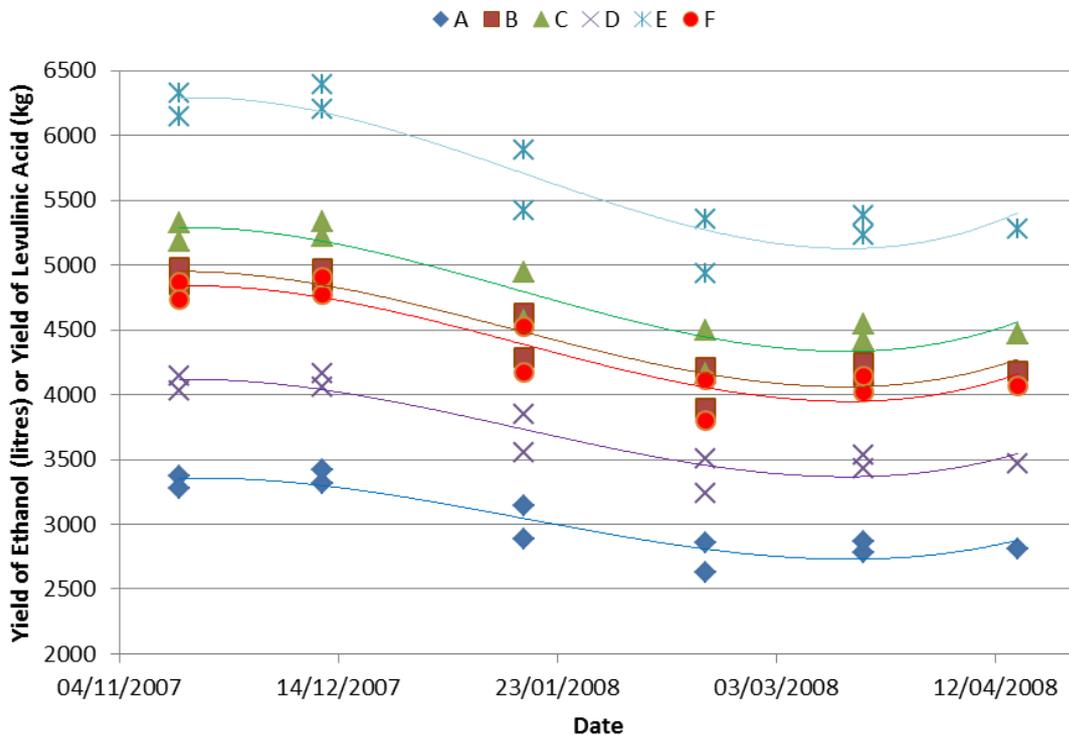


Figure G-68: Yield of ethanol (litres per hectare) and levulinic acid (kg per hectare) from processing, using technologies A to F, the projected standing stock of biomass on the Shanagolden site at various points in the harvest window. A = near-term dilute acid hydrolysis process; B = advanced dilute acid hydrolysis process; C = near-term concentrated acid hydrolysis process; D = near-term enzymatic hydrolysis process; E = advanced enzymatic hydrolysis process; F = DIBANET process projections.

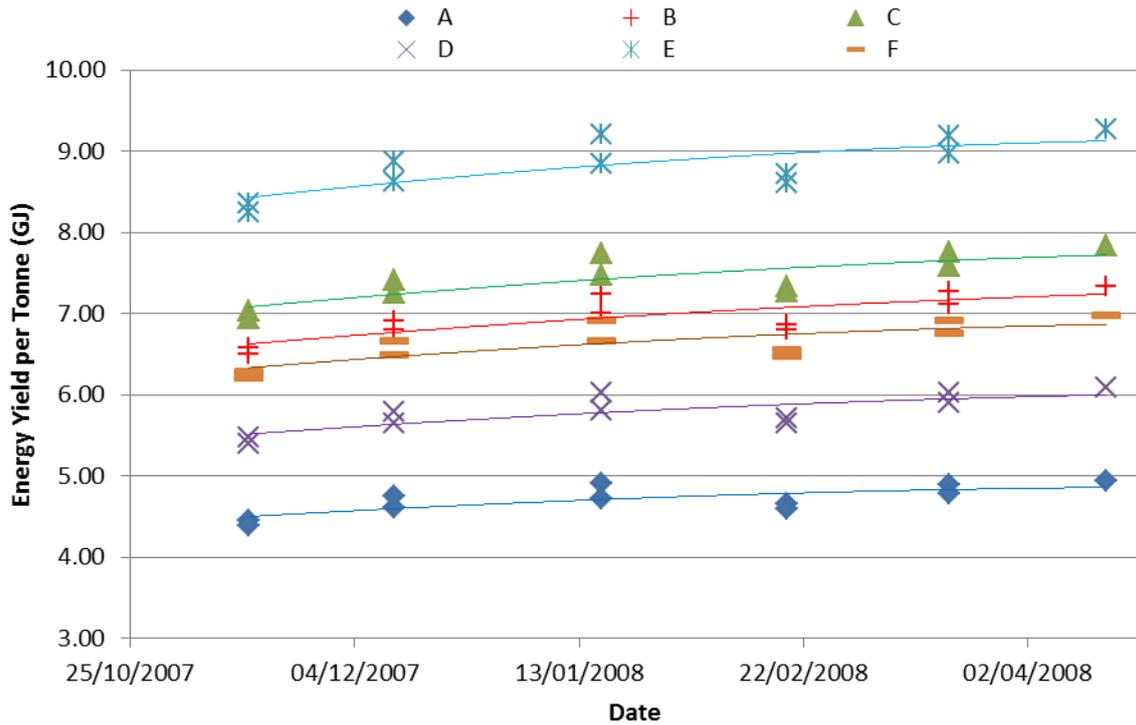


Figure G-69: Energy yield (GJ) per tonne from processing Miscanthus collected at various points in the harvest window through technologies A to F. A = near-term dilute acid hydrolysis process; B = advanced dilute acid hydrolysis process; C = near-term concentrated acid hydrolysis process; D = near-term enzymatic hydrolysis process; E = advanced enzymatic hydrolysis process; F = DIBANET process projections.

Appendix H Figures and Tables for Chapter 17: Analysis of Waste and Other Feedstocks

*Table H-1: The Universal Codes used for the straw samples and what species, variety, sampling location and date they represent. * = elemental data only for these samples. The suffixes P, B, and W for the last three samples represent the pods, branches, and leaves, respectively, of rapeseed straw.*

Code	Species	Variety	Location	Date
STSB1	Spring Barley	Cocktail	Backweston Farm	2/9/08
STSB2	Spring Barley	Quench	Backweston Farm	2/9/08
STSB3	Spring Barley	Frontier	Backweston Farm	2/9/08
STSB5	Spring Barley	Sebastian	Backweston Farm	2/9/08
STSB6*	Spring Barley	Snakebite	Backweston Farm	2/9/08
STWB3	Winter Barley	Saffron	Kildalton College	29/7/08
STWB4	Winter Barley	Volume	Backweston Farm	24/7/08
STWB6	Winter Barley	Saffron	Backweston Farm	24/7/08
STWB7	Winter Barley	Retriever	Kildalton College	29/7/08
STWB8	Winter Barley	Leibniz	Kildalton College	29/7/08
STWB9	Winter Barley	Leibniz	Backweston Farm	24/7/08
STWB10*	Winter Barley	Camion	Kildalton College	29/7/08
STWB11	Winter Barley	Spectrum	Kildalton College	29/7/08
STWB12	Winter Barley	Retriever	Backweston Farm	24/7/08
STSO2	Spring Oats	Husky	Kildalton College	30/9/08
STSO3	Spring Oats	Champion	Kildalton College	30/9/08
STSO4	Spring Oats	Nord	Kildalton College	30/9/08
STSO5	Spring Oats	Barra	Kildalton College	30/9/08
STWO2	Winter Oats	Jalna	Backweston Farm	2/9/08
STWO3	Winter Oats	Champion	Backweston Farm	2/9/08
STWO4	Winter Oats	Corrib	Backweston Farm	2/9/08
STWO5	Winter Oats	Euiat Hin	Backweston Farm	2/9/08
STWO6*	Winter Oats	Jalna Hin	Kildalton College	12/8/08
STSW2	Spring Wheat	Raffles	Kildalton College	30/9/08
STSW3*	Spring Wheat	Byron	Kildalton College	30/9/08
STSW5	Spring Wheat	Sparrow	Kildalton College	30/9/08
STSW6	Spring Wheat	Trappe	Kildalton College	30/9/08
STWW1	Winter Wheat	Timber	Kildalton College	29/7/08
STWW3	Winter Wheat	Cordiale	Kildalton College	29/7/08
STWW3	Winter Wheat	Sahara	Kildalton College	29/7/08
STWW4	Winter Wheat	Einstein	Kildalton College	29/7/08
STRS3P*	Rapeseed	Castille	Kildalton College	30/9/08
STRS3B*	Rapeseed	Castille	Kildalton College	30/9/08
STRS4W*	Rapeseed	Flash	Kildalton College	30/9/08

Table H-2: Extractives, ash, and lignocellulosic data (all in % dry matter) for various straw samples. Av = average; SD = standard deviation of duplicates; KL = Klason lignin; ASL = acid soluble lignin; AIR = acid insoluble residue; AIA = acid insoluble ash.

Sample	NIR	Extractives		Ash		KL		ASL		AIR		AIA		Arabinose		Galactose		Rhamnose		Glucose		Xylose		Mannose		Tot Sugars	
		Av	SD	Av	SD	Av	SD	Av	SD	Av	SD	Av	SD	Av	SD	Av	SD	Av	SD	Av	SD	Av	SD	Av	SD	Av	SD
STSB1	44001	6.41	(0.00)	1.29	(0.10)	19.86	(0.33)	2.11	(0.02)	19.53	(0.11)	-0.17	(0.22)	2.09	(0.05)	0.75	(0.03)	0.15	(0.00)	39.69	(0.08)	22.94	(0.01)	0.36	(0.02)	65.97	(0.16)
STSB2	44002	2.60	(0.34)	1.83	(0.07)	19.42	(0.02)	2.30	(0.12)	19.27	(0.01)	-0.14	(0.00)	2.21	(0.06)	0.73	(0.01)	0.14	(0.00)	42.23	(0.53)	23.32	(0.05)	0.40	(0.01)	69.04	(0.52)
STSB3	44003	4.38	(0.50)	1.63	(0.08)	19.01	(0.03)	2.29	(0.03)	18.84	(0.02)	-0.14	(0.05)	2.23	(0.03)	0.78	(0.01)	0.15	(0.01)	41.80	(0.12)	23.32	(0.01)	0.40	(0.00)	68.69	(0.08)
STSB5	44005	1.53	(0.64)	1.91	(0.06)	19.35	(0.18)	2.38	(0.05)	19.71	(0.28)	0.36	(0.10)	2.19	(0.05)	0.82	(0.01)	0.14	(0.00)	42.80	(0.17)	23.41	(0.29)	0.46	(0.01)	69.82	(0.53)
STWB3	40603	5.64	(0.15)	4.08	(0.03)	16.42	(0.11)	2.21	(0.04)	17.16	(0.31)	0.74	(0.41)	2.66	(0.09)	0.92	(0.02)	0.14	(0.00)	43.18	(0.15)	20.93	(0.14)	0.39	(0.04)	68.23	(0.15)
STWB4	40604	6.59	(0.94)	6.10	(0.05)	17.15	(0.06)	2.25	(0.02)	18.41	(0.06)	1.26	(0.13)	2.65	(0.02)	0.95	(0.01)	0.12	(0.00)	40.87	(0.06)	21.53	(0.14)	0.26	(0.00)	66.37	(0.24)
STWB6	40606	4.63	(0.61)	5.49	(0.10)	17.89	(0.09)	2.04	(0.02)	19.45	(0.11)	1.56	(0.02)	2.43	(0.02)	0.80	(0.00)	0.14	(0.00)	40.42	(0.06)	20.06	(0.12)	0.27	(0.02)	64.11	(0.06)
STWB7	40607	2.23	(0.00)	3.36	(0.20)	18.87	(0.30)	2.12	(0.03)	19.37	(0.19)	0.50	(0.10)	2.41	(0.06)	0.73	(0.01)	0.13	(0.00)	42.42	(0.37)	21.99	(0.09)	0.26	(0.03)	67.94	(0.35)
STWB8	40608	4.32	(0.27)	4.07	(0.06)	19.53	(0.21)	2.05	(0.00)	19.90	(0.16)	0.38	(0.05)	2.43	(0.00)	0.75	(0.00)	0.12	(0.00)	42.50	(0.01)	21.15	(0.09)	0.23	(0.00)	67.18	(0.11)
STWB9	40609	3.05	(0.07)	5.87	(0.09)	19.11	(0.13)	2.17	(0.03)	20.21	(0.03)	1.09	(0.10)	2.47	(0.03)	0.85	(0.01)	0.14	(0.00)	41.71	(0.40)	20.98	(0.08)	0.25	(0.00)	66.40	(0.44)
STWB11	40611	4.53	(0.14)	4.25	(0.13)	18.43	(0.02)	2.36	(0.01)	18.89	(0.05)	0.46	(0.03)	2.72	(0.01)	0.88	(0.01)	0.13	(0.00)	40.85	(0.00)	21.99	(0.05)	0.18	(0.06)	66.75	(0.14)
STWB12	40612	6.11	(0.08)	6.25	(0.11)	16.79	(0.09)	2.21	(0.01)	18.31	(0.11)	1.52	(0.02)	2.71	(0.08)	1.02	(0.01)	0.15	(0.00)	41.23	(0.56)	20.02	(0.11)	0.46	(0.01)	65.58	(0.59)
STSO2	45002	3.48	(0.36)	2.71	(0.14)	19.72	(0.04)	2.44	(0.01)	19.98	(0.29)	0.27	(0.33)	2.42	(0.06)	1.00	(0.05)	0.15	(0.00)	44.20	(0.12)	20.52	(0.01)	0.52	(0.02)	68.81	(0.01)
STSO3	45003	4.29	(0.34)	4.20	(0.12)	20.04	(0.17)	2.31	(0.02)	20.58	(0.19)	0.54	(0.02)	2.49	(0.03)	1.00	(0.02)	0.12	(0.00)	42.13	(0.21)	21.48	(0.16)	0.27	(0.01)	67.50	(0.31)
STSO4	45004	3.05	(0.58)	6.12	(0.12)	20.54	(0.21)	2.62	(0.03)	21.56	(0.12)	1.02	(0.10)	2.41	(0.01)	0.90	(0.02)	0.16	(0.01)	41.73	(0.22)	22.15	(0.07)	0.41	(0.03)	67.76	(0.21)
STSO5	45005	1.87	(0.07)	2.70	(0.11)	19.96	(0.08)	2.13	(0.03)	20.57	(0.08)	0.61	(0.00)	2.48	(0.05)	0.90	(0.05)	0.14	(0.00)	43.85	(0.54)	21.73	(0.14)	0.32	(0.00)	69.42	(0.58)
STWO2	41002	1.88	(0.30)	4.84	(0.04)	18.50	(0.27)	2.32	(0.04)	19.99	(0.25)	1.49	(0.02)	2.90	(0.09)	1.09	(0.05)	0.17	(0.01)	41.43	(0.03)	20.97	(0.27)	0.36	(0.04)	66.93	(0.11)
STWO3	41003	1.24	(0.16)	4.55	(0.10)	18.35	(0.22)	2.25	(0.01)	19.66	(0.13)	1.31	(0.10)	2.97	(0.04)	1.10	(0.02)	0.16	(0.00)	42.59		21.16		0.28	(0.01)	68.25	
STWO4	41004	3.43	(0.10)	4.42	(0.70)	18.20	(0.21)	2.26	(0.01)	19.10	(0.11)	0.91	(0.10)	2.90	(0.07)	1.13	(0.04)	0.17	(0.00)	40.71	(0.22)	21.12	(0.03)	0.31	(0.00)	66.34	(0.14)
STWO5	41005	2.98	(0.66)	4.78	(0.02)	17.83	(0.28)	2.32	(0.06)	18.85	(0.23)	1.02	(0.05)	2.98	(0.01)	1.14	(0.00)	0.17	(0.00)	41.06	(0.27)	21.09	(0.06)	0.30	(0.00)	66.74	(0.31)
STSW2	42002	4.77	(0.09)			19.45	(0.13)	1.94	(0.06)	21.27	(0.02)	1.82	(0.15)	2.31	(0.04)	0.83	(0.01)	0.16	(0.00)	39.56	(0.06)	22.53	(0.14)	0.55	(0.02)	65.94	(0.16)
STSW5	42005	2.06	(0.51)	5.40	(0.21)	21.59	(0.18)	2.58	(0.11)	23.44	(0.20)	1.85	(0.02)	1.93	(0.03)	0.85	(0.02)	0.13	(0.00)	38.19	(0.02)	21.62	(0.07)	0.49	(0.02)	63.21	(0.12)
STSW6	42006	4.01	(0.00)	4.47	(0.12)	20.63	(0.09)	2.73	(0.04)	22.24	(0.09)	1.61	(0.00)	2.13	(0.01)	0.78	(0.01)	0.14	(0.00)	36.61	(0.04)	21.72	(0.12)	0.30	(0.00)	61.67	(0.18)
STWW1	43001	3.39	(0.10)	3.64	(0.07)	18.84	(0.12)	2.14	(0.00)	20.72	(0.12)	1.89	(0.00)	2.58	(0.06)	0.86	(0.01)	0.16	(0.00)	38.91	(0.03)	24.29	(0.18)	0.42	(0.01)	67.21	(0.23)
STWW3	43003	4.80	(0.11)			18.67	(0.23)	2.29	(0.01)	19.81	(0.18)	1.14	(0.04)	2.41	(0.06)	0.82	(0.01)	0.15	(0.00)	39.38	(0.24)	23.20	(0.07)	0.42	(0.04)	66.39	(0.05)
STWW4	43004	4.43	(0.26)	2.99	(0.11)	18.88	(0.02)	2.31	(0.05)	20.27	(0.05)	1.40	(0.02)	2.38	(0.07)	0.85	(0.01)	0.16	(0.00)	39.73	(0.16)	23.04	(0.17)	0.54	(0.02)	66.70	(0.26)

Table H-3: Summary statistics for the average and standard deviations of the compositional values (all in % dry matter) of the different straw categories and all straw samples.

Straw Type	Extractives	Ash	KL	ASL	AIR	AIA	Arabinose	Galactose	Rhamnose	Glucose	Xylose	Mannose	Tot Sugars
Average Values													
Spring Barley, Av	3.73	1.67	19.41	2.27	19.34	-0.02	2.18	0.77	0.15	41.63	23.25	0.41	68.38
Winter Barley, Av	4.64	4.94	18.02	2.18	18.96	0.94	2.56	0.86	0.13	41.65	21.08	0.29	66.57
Spring Oats, Av	3.17	3.93	20.07	2.38	20.67	0.61	2.45	0.95	0.14	42.98	21.47	0.38	68.37
Winter Oats, Av	2.38	4.65	18.22	2.29	19.40	1.18	2.94	1.12	0.17	41.45	21.09	0.31	67.07
Spring Wheat, Av	3.61	4.94	20.56	2.42	22.32	1.76	2.12	0.82	0.14	38.12	21.96	0.45	63.61
Winter Wheat, Av	4.36	3.32	18.77	2.26	20.15	1.39	2.45	0.84	0.16	39.35	23.43	0.45	66.67
All, Av	3.80	4.04	18.95	2.27	19.89	0.94	2.48	0.89	0.15	41.08	21.91	0.36	66.86
Standard Deviation													
Spring Barley, SD	2.14	0.28	0.35	0.11	0.38	0.26	0.06	0.04	0.01	1.36	0.21	0.04	1.67
Winter Barley, SD	1.48	1.11	1.15	0.11	0.99	0.48	0.14	0.10	0.01	0.97	0.76	0.09	1.32
Spring Oats, SD	1.01	1.62	0.34	0.21	0.65	0.31	0.04	0.06	0.02	1.23	0.69	0.11	0.90
Winter Oats, SD	1.00	0.20	0.29	0.04	0.52	0.27	0.04	0.02	0.01	0.82	0.08	0.03	0.83
Spring Wheat, SD	1.40	0.66	1.07	0.42	1.09	0.13	0.19	0.04	0.02	1.48	0.50	0.13	2.16
Winter Wheat, SD	0.67	0.46	0.11	0.08	0.44	0.35	0.09	0.02	0.01	0.34	0.58	0.06	0.39
All, SD	1.48	1.48	1.16	0.18	1.27	0.62	0.27	0.12	0.02	1.76	1.12	0.10	1.82

Table H-4: Summary statistics for the maximum, minimum, and range of the compositional values (all in % dry matter) of the different straw categories, and of all straw samples.

Straw Type	Extr.	Ash	KL	ASL	AIR	AIA	Arabinose	Galactose	Rhamnose	Glucose	Xylose	Mannose	Tot Sugars
Maximum Values													
Spring Barley, Max	6.41	1.91	19.86	2.38	19.71	0.36	2.23	0.82	0.15	42.80	23.41	0.46	69.82
Winter Barley, Max	6.59	6.25	19.53	2.36	20.21	1.56	2.72	1.02	0.15	43.18	21.99	0.46	68.23
Spring Oats, Max	4.29	6.12	20.54	2.62	21.56	1.02	2.49	1.00	0.16	44.20	22.15	0.52	69.42
Winter Oats, Max	3.43	4.84	18.50	2.32	19.99	1.49	2.98	1.14	0.17	42.59	21.16	0.36	68.25
Spring Wheat, Max	4.77	5.40	21.59	2.73	23.44	1.85	2.31	0.85	0.16	39.56	22.53	0.55	65.94
Winter Wheat, Max	4.80	3.64	18.88	2.31	20.72	1.89	2.58	0.86	0.16	39.73	24.29	0.54	67.21
All, Max	6.59	6.25	21.59	2.73	23.44	1.89	2.98	1.14	0.17	44.20	24.29	0.55	69.82
Minimum Values													
Spring Barley, Min	1.53	1.29	19.01	2.11	18.84	-0.17	2.09	0.73	0.14	39.69	22.94	0.36	65.97
Winter Barley, Min	2.23	3.36	16.42	2.04	17.16	0.38	2.41	0.73	0.12	40.42	20.02	0.18	64.11
Spring Oats, Min	1.87	2.70	19.72	2.13	19.98	0.27	2.41	0.90	0.12	41.73	20.52	0.27	67.50
Winter Oats, Min	1.24	4.42	17.83	2.25	18.85	0.91	2.90	1.09	0.16	40.71	20.97	0.28	66.34
Spring Wheat, Min	2.06	4.47	19.45	1.94	21.27	1.61	1.93	0.78	0.13	36.61	21.62	0.30	61.67
Winter Wheat, Min	3.39	2.99	18.67	2.14	19.81	1.14	2.38	0.82	0.15	38.91	23.04	0.42	66.39
All, Min	1.24	1.29	16.42	1.94	17.16	-0.17	1.93	0.73	0.12	36.61	20.02	0.18	61.67
Range in Values													
Spring Barley, Range	4.88	0.62	0.85	0.27	0.87	0.53	0.14	0.09	0.01	3.11	0.47	0.10	3.85
Winter Barley, Range	4.36	2.89	3.11	0.32	3.05	1.18	0.31	0.29	0.03	2.76	1.97	0.28	4.12
Spring Oats, Range	2.42	3.42	0.82	0.49	1.58	0.75	0.08	0.10	0.04	2.47	1.63	0.25	1.92
Winter Oats, Range	2.19	0.42	0.67	0.07	1.14	0.58	0.08	0.05	0.01	1.88	0.19	0.08	1.91
Spring Wheat, Range	2.71	0.93	2.14	0.79	2.17	0.24	0.38	0.07	0.03	2.95	0.91	0.25	4.27
Winter Wheat, Range	1.41	0.65	0.21	0.17	0.91	0.75	0.20	0.04	0.01	0.82	1.25	0.12	0.82
All, Range	5.35	4.96	5.17	0.79	6.28	2.06	1.05	0.41	0.05	7.59	4.27	0.37	8.15

Table H-5: Ash and elemental data for the DS and DF fractions of selected straws. The percentage that the DS fraction contributed to the DG sample is provided allowing for the calculation of a weighted average for the ash, elemental contents, and heating values of each sample. Av = average; SD = standard deviation. Dry matter heating values are only provided for the combined sample.

Sample	DS Fraction (% Dry of Matter)										DF Fraction (% Dry of Matter)										% DS	DS and DF Combined							
	Ash		C		H		N		S		Ash		C		H		N		S			Ash (%)	C (%)	H (%)	N (%)	S (%)	HHV MJ/kg	LHV MJ/kg	
	AV	SD	Av	SD	Av	SD	Av	SD	Av	SD	AV	SD	Av	SD	Av	SD	Av	SD	Av	SD									
STSB6	1.28	(0.05)	48.48	(0.02)	6.24	(0.01)	0.56	(0.01)	0.01	(0.00)	2.21	(0.05)	48.61	(0.09)	6.28	(0.04)	0.86	(0.00)	0.03	(0.00)	73.25	1.53	48.51	6.25	0.64	0.02	19.62	18.25	
STWB10	4.40		47.51	(0.12)	6.48	(0.01)	0.48	(0.04)	0.05	(0.02)	4.95	(0.02)	47.20	(0.03)	6.51	(0.00)	0.67	(0.00)	0.04	(0.00)	67.49	4.58	47.41	6.49	0.54	0.05	19.38	17.96	
STSO2	2.71	(0.14)	48.38	(0.01)	6.36	(0.01)	0.41	(0.03)	0.01	(0.00)	4.01	(0.09)	47.91	(0.09)	6.47	(0.01)	0.79	(0.02)	0.02	(0.00)	70.26	3.10	48.24	6.39	0.52	0.01	19.59	18.19	
STWO6	4.85	(0.04)	46.84	(0.03)	6.04	(0.02)	0.40	(0.01)	0.02	(0.00)	6.14	(0.06)	46.41	(0.10)	6.03	(0.02)	0.54	(0.01)	0.02	(0.00)	68.54	5.26	46.70	6.04	0.44	0.02	18.85	17.52	
STSW3	4.49		46.83	(0.18)	6.25	(0.08)	0.77	(0.05)	0.09	(0.01)	6.30	(0.04)	46.25	(0.03)	6.19	(0.04)	1.26	(0.00)	0.10	(0.00)	60.93	5.20	46.60	6.23	0.96	0.09	18.95	17.58	
STWW4	2.99	(0.11)	47.91	(0.04)	6.52	(0.01)	0.39	(0.02)	0.00	(0.00)	3.98	(0.02)	47.70	(0.01)	6.58	(0.00)	0.54	(0.02)	0.00	(0.00)	69.41	3.29	47.85	6.54	0.44	0.00	19.57	18.13	
STRS3P	4.96	(0.08)	46.62	(0.16)	6.45	(0.04)	0.56	(0.06)	0.05	(0.01)	6.68	(0.05)	46.11	(0.03)	6.39	(0.01)	0.85	(0.00)	0.10	(0.00)	56.93	5.70	46.40	6.42	0.68	0.07	19.01	17.60	
STRS4F	4.79	(0.16)	47.16	(0.02)	6.50	(0.04)	0.50	(0.02)	0.05	(0.01)	5.14	(0.03)	47.37	(0.05)	6.52	(0.03)	0.77	(0.01)	0.06	(0.01)	65.37	4.91	47.23	6.51	0.59	0.05	19.32	17.89	
STRS3B	2.63	(0.13)	47.80	(0.06)	6.51	(0.03)	0.46	(0.05)	0.04	(0.00)	5.28	(0.01)	46.49	(0.02)	6.46	(0.01)	1.04	(0.02)	0.17	(0.01)	63.89	3.59	47.33	6.49	0.67	0.09	19.38	17.95	

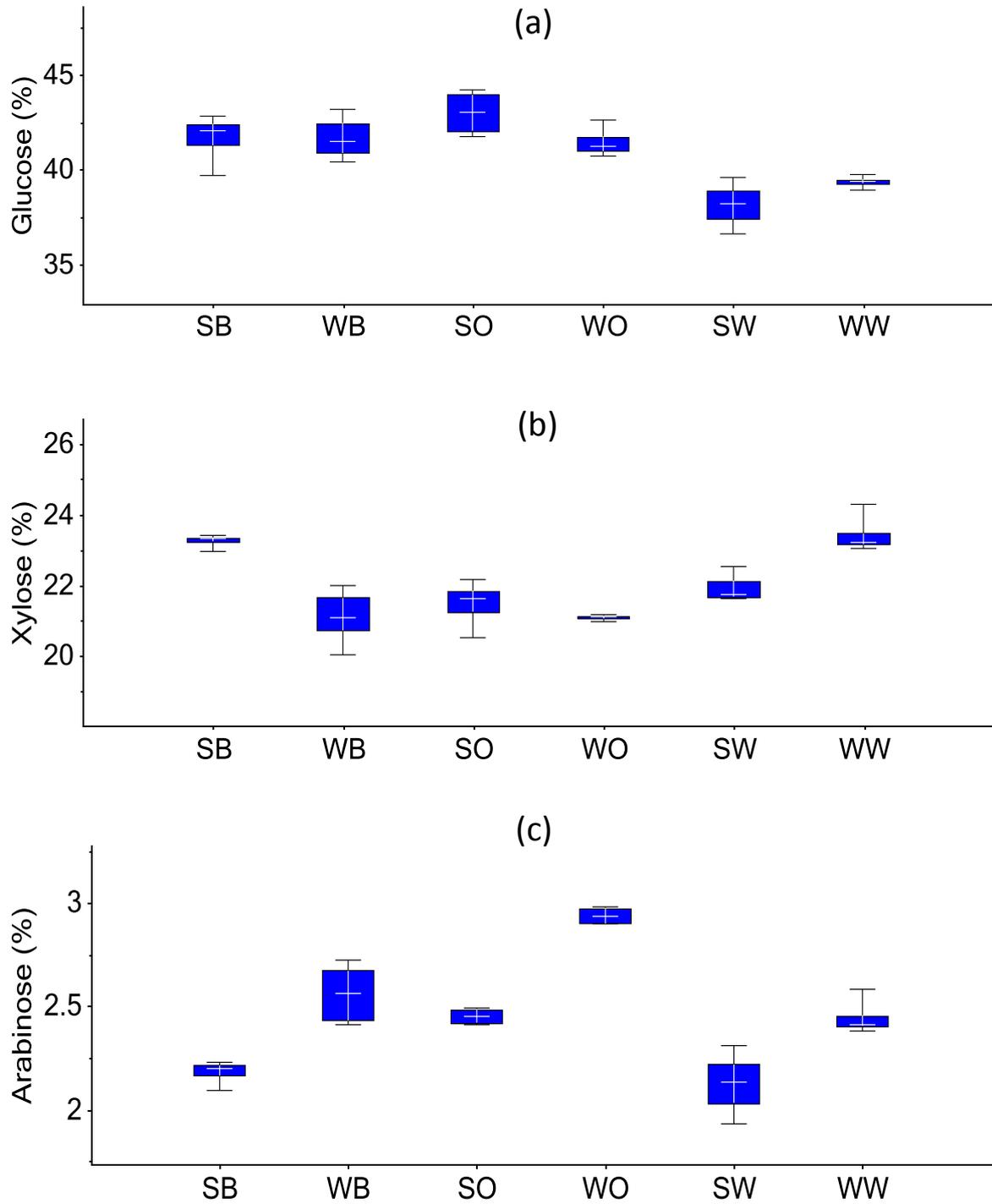


Figure H-1: Quantile plots involving spring barley (SB), winter barely (WB), spring oats (SO), winter oats (WO), spring wheat (SW), and winter wheat (WW) for: (a) glucose; (b) xylose; and (c) arabinose. All data in % of whole dry mass.

Table H-6: Estimates of the quantities of straw produced in three scenarios. (i) Total amounts of straw produced; (ii) upper estimate of practical resource available for end-use; (iii) lower estimate of practical resource available for end-use. Straw ratio = straw yield as a proportion of total grain yield. "Straw – SMC" takes away the wheat straw quantities that are estimated to be used by the mushroom compost industry (Teagasc, 2002).

Species	Area in June 2006 ('000 ha) (CSO, 2007)	Yield (t/ha) (CSO, 2007)	Straw ratio	Straw yield (odt/ha)	(i) Total Resource ('000 odt)	Straw – SMC ('000 odt)	% (total)	(ii) Upper Quantity Scenario ('000 odt)	(iii) Lower Quantity Scenario ('000 odt)
Wheat									
Winter	59.2	9.8	0.55	4.6	271	208	21.8%	71	17
Spring	28.3	7.8	0.62 ^a	4.1	116	89	9.3%	30	7
Oats									
Winter	9.3	8.0	0.86	5.8	54	54	5.7%	19	5
Spring	11.1	6.4	0.86 ^b	4.7	52	52	5.4%	18	4
Barley									
Winter	15.1	7.9	0.56	3.8	57	57	5.9%	19	5
Spring	151.9	6.7	0.55	3.1	476	476	49.8%	162	40
TOTAL					1,045	955		319	78

a = the spring wheat straw/grain ratios were not investigated in (RPS MCOS, 2004), instead the percentage difference seen (from a German study (Kaltschmitt and Hartmann, 2000)) in the straw/grain ratio compared with the winter variety is used to approximate the Irish spring variety ratio from that of the winter variety; b = winter oats straw/grain ratio is used.

Table H-7: Extractives, ash, and lignocellulosic data for animal manures. Av = average; SD = standard deviation; KL = Klason lignin; ASL = acid soluble lignin; AIR = acid insoluble residue; AIA = acid insoluble ash. All these are expressed on a % whole mass (dry matter) basis. MC = moisture content (% wet basis).

Sample	Extractives		Ash		KL		ASL		AIR		AIA		Arabinose		Galactose		Rhamnose		Glucose		Xylose		Mannose		Tot Sugars		MC			
	Av	SD	Av	SD	Av	SD	Av	SD	Av	SD	Av	SD	Av	SD	Av	SD	Av	SD	Av	SD	Av	SD	Av	SD	Av	SD	Av	SD		
Pigs																														
MNPG1	9.44	(0.18)	15.62	(0.04)	16.32	(0.15)	3.77	(0.08)	18.94	(0.01)	2.62	(0.17)	5.26	(0.06)	0.79	(0.00)	0.22	(0.00)	21.94	(0.18)	17.75	(0.14)	0.23	(0.00)	46.19	(0.37)				
MNPG2	15.38	(0.55)			16.75	(0.00)	4.08	(0.03)	19.14	(0.13)	2.39	(0.13)	4.14	(0.07)	0.74	(0.00)	0.32	(0.01)	9.70	(0.18)	8.75	(0.15)	0.17	(0.00)	23.81	(0.38)				
MNPG3	20.89	(0.04)	21.11	(0.06)	15.10	(0.09)	3.91	(0.01)	16.85	(0.04)	1.75	(0.05)	5.07	(0.02)	0.89	(0.00)	0.35	(0.00)	11.23	(0.11)	9.83	(0.06)	0.25	(0.01)	27.61	(0.18)				
MNPG4	16.39	(0.71)			19.82	(0.11)	4.38	(0.05)	21.18	(0.15)	1.36	(0.04)	6.64	(0.05)	0.92	(0.00)	0.26	(0.00)	13.17	(0.25)	12.48	(0.28)	0.24	(0.01)	33.72	(0.57)				
MNPG5	34.18	(3.80)			8.92	(0.21)	4.22	(0.10)	10.00	(0.27)	1.08	(0.06)	0.41	(0.00)	0.46	(0.00)	0.46	(0.00)	1.13	(0.03)	0.65	(0.03)	0.16	(0.01)	3.27	(0.05)				
MNPG6	22.11	(0.86)	22.94	(0.21)	15.33	(0.14)	4.77	(0.04)	16.56	(0.16)	1.23	(0.02)	4.15	(0.08)	0.75	(0.02)	0.30	(0.02)	10.20	(0.51)	9.11	(0.39)	0.26	(0.00)	24.76	(0.94)				
MNPG7	15.75	(0.07)	34.86	(0.59)	17.15	(0.05)	3.63	(0.05)	20.16	(0.23)	3.02	(0.28)	3.40	(0.08)	0.61	(0.01)	0.38	(0.01)	8.77	(0.40)	7.88	(0.34)	0.17	(0.00)	21.21	(0.80)				
MNPG8	13.91	(0.47)	24.10	(0.60)	18.83	(0.09)	3.79	(0.00)	20.62	(0.12)	1.79	(0.03)	4.45	(0.02)	0.68	(0.00)	0.22	(0.00)	16.28	(0.20)	14.18	(0.13)	0.13	(0.00)	35.96	(0.35)				
Dairy Cows																														
MNDY1	11.31	(0.50)	19.56	(0.27)	21.72	(0.04)	3.24	(0.02)	23.32	(0.09)	1.60	(0.05)	2.35	(0.02)	0.83	(0.01)	0.30	(0.01)	22.02	(0.00)	13.50	(0.04)	0.32	(0.03)	39.31	(0.04)	95.05			
MNDY2	12.88	(0.05)	21.56		22.52	(0.09)	3.72	(0.21)	24.54	(0.03)	2.02	(0.12)	2.27	(0.06)	0.86	(0.03)	0.31	(0.01)	19.95	(0.83)	12.67	(0.27)	0.24	(0.03)	36.30	(1.03)	92.90			
MNDY3	13.15	(0.71)	21.37	(0.21)	24.70	(0.32)	3.54	(0.02)	26.65	(0.40)	1.94	(0.08)	2.00	(0.01)	0.90	(0.03)	0.34	(0.01)	18.35	(0.52)	11.44	(0.29)	0.20	(0.02)	33.22	(0.74)	92.52			
MNDY4	10.55	(0.40)			24.30	(0.17)	3.79	(0.03)	28.40	(0.23)	4.10	(0.06)	1.19	(0.04)	0.74	(0.00)	0.46	(0.01)	9.54	(0.70)	6.30	(0.59)	0.30	(0.02)	18.53	(1.34)	96.99			
MNDY6	14.78	(1.14)	20.48	(0.02)	25.42	(0.19)	3.10	(0.01)	28.26	(0.18)	2.84	(0.01)	1.89	(0.01)	0.86	(0.01)	0.27	(0.00)	19.07	(0.29)	12.57	(0.13)	0.14	(0.01)	34.79	(0.41)	92.51			

Table H-8: Ash and elemental data for the DS and DF fractions of some animal manures. The percentage that the DS fraction contributed to the DG sample is provided allowing for the calculation of a weighted average for the ash and elemental contents of each sample. Av = average; SD = standard deviation.

Sample	DS Fraction (% Dry of Matter)										DF Fraction (% Dry of Matter)										% DS	DS and DF Combined							
	Ash		C		H		N		S		Ash		C		H		N		S			Ash (%)	C (%)	H (%)	N (%)	S (%)	HHV MJ/kg	LHV MJ/kg	
	AV	SD	Av	SD	Av	SD	Av	SD	Av	SD	AV	SD	Av	SD	Av	SD	Av	SD	Av	SD									
MNPG1	15.62	(0.04)	44.00	(0.33)	6.13	(0.01)	2.29	(0.06)	0.49	(0.05)	31.48	(0.93)	38.60	(0.19)	5.51	(0.36)	3.41	(0.04)	0.50	(0.23)	69.52	20.45	42.35	5.94	2.63	0.49	17.08	15.77	
MNPG6	22.94	(0.21)	43.43	(0.01)	5.82	(0.34)	3.64	(0.04)	0.59	(0.20)	30.53	(0.23)	40.57	(0.05)	5.54	(0.35)	4.38	(0.09)	0.73	(0.37)	62.90	25.76	42.37	5.72	3.91	0.64	16.76	15.50	
MNDY1	19.54		43.11	(0.14)	6.10	(0.17)	1.67	(0.01)	0.40	(0.07)	26.79		40.28	(0.09)	6.11	(0.07)	2.41	(0.01)	0.50	(0.00)	23.85	25.06	40.95	6.11	2.23	0.48	16.65	15.30	

Table H-9: Sugars content, expressed as a percentage of total sugars, for the various pig and dairy cattle manures. The ratio of total hemicellulosic sugars to cellulose is also presented.

Sample	Proportion of Total Sugars (%)						Hemicellulose: cellulose
	Arabinose	Galactose	Rhamnose	Glucose	Xylose	Mannose	
Pig Manure Samples							
MNPG1	11.39	1.71	0.48	47.50	38.43	0.50	1.11
MNPG2	17.39	3.11	1.34	40.74	36.75	0.71	1.46
MNPG3	18.36	3.22	1.27	40.67	35.60	0.91	1.46
MNPG4	19.69	2.73	0.77	39.06	37.01	0.71	1.56
MNPG5	12.54	14.07	14.07	34.56	19.88	4.89	1.89
MNPG6	16.76	3.03	1.21	41.20	36.79	1.05	1.43
MNPG7	16.03	2.88	1.79	41.35	37.15	0.80	1.42
MNPG8	12.37	1.89	0.61	45.27	39.43	0.36	1.21
Dairy Manure Samples							
MNDY1	5.98	2.11	0.76	56.02	34.34	0.81	0.79
MNDY2	6.25	2.37	0.85	54.96	34.90	0.66	0.82
MNDY3	6.02	2.71	1.02	55.24	34.44	0.60	0.81
MNDY4	6.42	3.99	2.48	51.48	34.00	1.62	0.94
MNDY6	5.43	2.47	0.78	54.81	36.13	0.40	0.82

Table H-10: Extractives, ash, and lignocellulosic data for various mushroom compost samples. Av = average; SD = standard deviation; KL = Klason lignin; ASL = acid soluble lignin; AIR = acid insoluble residue; AIA = acid insoluble ash. All these are expressed on a % whole mass (dry matter) basis. MC = moisture content (% wet basis).

Sample	Extractives		Ash		KL		ASL		AIR		AIA		Arabinose		Galactose		Rhamnose		Glucose		Xylose		Mannose		Tot Sugars		MC	
	Av	SD	Av	SD	Av	SD	Av	SD	Av	SD	Av	SD	Av	SD	Av	SD	Av	SD	Av	SD	Av	SD	Av	SD	Av	SD	Av	SD
Samples from the mushroom compost production process																												
MCPR6	3.28	(0.17)	5.47	(0.05)	19.66	(0.15)	2.34	(0.05)	19.42	(0.00)	-0.14	(0.15)	2.91	(0.03)	0.90	(0.00)	0.17	(0.00)	40.23	(0.15)	21.10	(0.01)	0.20	(0.01)	65.51	(0.11)	76.13	(1.20)
MCPR5	3.16	(0.09)	18.54	(0.99)	23.50	(0.21)	2.76	(0.01)	26.96	(0.22)	3.46	(0.01)	1.94	(0.01)	0.65	(0.03)	0.19	(0.00)	29.34	(0.32)	13.81	(0.20)	0.57	(0.04)	46.50	(0.45)	74.38	(0.16)
MCPR4	2.64	(0.12)			28.56	(0.12)	3.42	(0.00)	33.09	(0.04)	4.53	(0.16)	1.55	(0.03)	0.62	(0.01)	0.21	(0.00)	20.51	(0.32)	10.16	(0.15)	0.85	(0.04)	33.91	(0.39)	68.51	(0.22)
MCPR8	2.58	(0.14)	23.44	(0.75)	28.00	(0.02)	0.84	(0.03)	32.25	(0.07)	4.25	(0.09)	1.47	(0.03)	0.62	(0.02)	0.20	(0.00)	21.79	(0.13)	10.33	(0.16)	0.87	(0.02)	35.28	(0.28)	69.90	(0.32)
MCPR3	3.98	(0.16)			25.65	(0.19)	3.60	(0.10)	31.80	(0.08)	6.14	(0.12)	1.37	(0.01)	0.58	(0.01)	0.21	(0.00)	23.15	(0.19)	11.08	(0.07)	0.74	(0.00)	37.14	(0.23)	66.92	(0.48)
Spent Mushroom Compost Samples																												
MCSP3	2.04	(0.18)	31.33	(3.01)	32.84	(0.26)	4.36	(0.02)	38.89	(0.11)	6.06	(0.37)	0.90	(0.01)	0.83	(0.01)	0.33	(0.00)	11.69	(0.34)	4.39	(0.17)	1.22	(0.01)	19.36	(0.50)	69.66	(2.54)
MCSP6	1.77	(1.02)	27.52	(1.64)	29.82	(0.03)	4.11	(0.09)	36.40	(1.00)	6.58	(1.03)	1.06	(0.00)	0.71	(0.01)	0.26	(0.01)	15.33	(0.17)	6.84	(0.15)	0.91	(0.06)	25.11	(0.25)	67.04	(1.97)
MCSP7	4.22	(0.05)	32.09	(0.23)	31.40	(0.33)	3.45	(0.09)	41.59	(0.41)	10.19	(0.74)	0.68	(0.02)	0.88	(0.01)	0.40	(0.00)	13.63	(0.06)	4.58	(0.02)	1.08	(0.03)	21.26	(0.07)	70.10	(1.84)
MCSP8	1.82	(0.16)	38.33	(0.25)	32.75	(1.24)	3.19	(0.12)	39.17	(0.49)	6.42	(0.75)	0.75	(0.01)	0.84	(0.00)	0.31	(0.00)	11.05	(0.21)	4.25	(0.08)	0.87	(0.00)	18.06	(0.28)	68.35	(1.97)

Table H-11: Ash and elemental data for the DS and DF fractions of samples of various mushroom composts. The percentage that the DS fraction contributed to the DG sample is provided. This allows the calculation of a weighted average for the ash and elemental contents of each sample. Av = average; SD = standard deviation. Dry matter heating values are also provided for the combined sample.

Sample	DS Fraction (% Dry of Matter)										DF Fraction (% Dry of Matter)										% DS	DS and DF Combined						
	Ash		C		H		N		S		Ash		C		H		N		S			Ash (%)	C (%)	H (%)	N (%)	S (%)	HHV MJ/kg	LHV MJ/kg
	AV	SD	Av	SD	Av	SD	Av	SD	Av	SD	AV	SD	Av	SD	Av	SD	Av	SD	Av	SD								
MCPR6	5.47	(0.05)	46.54	(0.26)	6.01	(0.34)	0.58	(0.00)	0.42	(0.17)	9.38	(0.01)	44.89	(0.33)	5.89	(0.32)	1.02	(0.01)	0.96	(0.45)	52.31	7.33	45.75	5.95	0.79	0.68	18.46	17.15
MCPR5	18.54	(0.99)	41.08	(1.21)	5.15	(0.15)	2.02	(0.02)	1.77	(0.86)	29.29	(0.34)	36.96	(0.25)	4.62	(0.33)	2.27	(0.00)	3.11	(1.36)	69.36	21.83	39.82	4.99	2.10	2.18	15.68	14.58
MCPR8	23.44	(0.75)	40.34	(0.02)	4.99	(0.31)	2.75	(0.06)	2.38	(1.20)	29.05	(0.22)	37.66	(0.29)	4.66	(0.26)	2.87	(0.05)	3.30	(1.59)	69.70	25.14	39.53	4.89	2.79	2.66	15.43	14.35
MCSP3	31.33	(3.01)	36.63	(0.27)	3.91	(0.30)	2.47	(0.02)	1.12	(0.65)	33.87	(0.13)	36.51	(0.09)	3.87	(0.29)	2.33	(0.01)	1.29	(0.62)	67.48	32.16	36.59	3.90	2.42	1.18	13.71	12.85
MCSP7	32.09	(0.23)	38.57	(0.70)	3.82	(0.05)	2.08	(0.02)	0.86	(0.08)										73.53								
MCSP8	38.33	(0.25)	37.87	(1.30)	4.03	(0.51)	1.88	(0.08)	0.74	(0.33)	38.79	(0.14)	35.54	(0.17)	3.59	(0.23)	1.88	(0.06)	0.84	(0.37)	73.66	38.45	37.26	3.91	1.88	0.77	13.71	12.85
MCSP9	24.79	(0.10)	39.05	(0.85)	4.65	(0.25)	2.69	(0.03)	2.56	(1.27)	34.44	(0.27)	34.07	(0.19)	4.02	(0.24)	2.76	(0.06)	3.84	(1.81)	68.22	27.86	37.47	4.45	2.71	2.97	14.47	13.49

Table H-12: Sugars content, expressed as a percentage of total sugars, for the various compost and SMC samples. The ratio of total hemicellulosic sugars to cellulose is also presented.

Sample	Proportion of Total Sugars (%)						Hemicellulose: cellulose
	Arabinose	Galactose	Rhamnose	Glucose	Xylose	Mannose	
Mushroom Compost Samples							
MCPR6	4.44	1.37	0.26	61.41	32.21	0.31	0.63
MCPR5	4.17	1.40	0.41	63.10	29.70	1.23	0.58
MCPR4	4.57	1.83	0.62	60.50	29.97	2.51	0.65
MCPR8	4.17	1.76	0.57	61.76	29.28	2.47	0.62
MCPR3	3.69	1.56	0.57	62.35	29.84	1.99	0.60
Spent Mushroom Compost Samples							
MCSP3	4.65	4.29	1.70	60.38	22.68	6.30	0.66
MCSP6	4.22	2.83	1.04	61.05	27.24	3.62	0.64
MCSP7	3.20	4.14	1.88	64.14	21.55	5.08	0.56
MCSP8	4.15	4.65	1.72	61.15	23.52	4.81	0.64

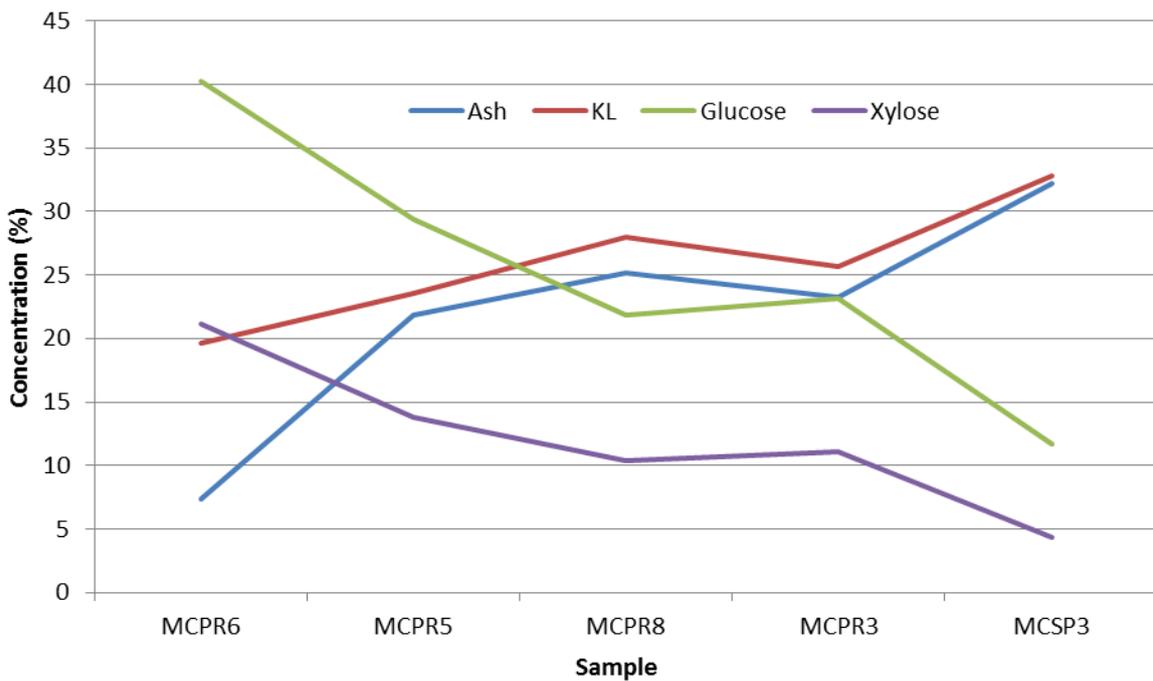


Figure H-2: The changes in constituent concentrations (% whole dry mass basis) with stages in mushroom compost, and SMC, production. MCPR6 = the mixture of straw, poultry litter and gypsum prior to composting, MCPR5 = Phase I compost; MCPR8 = Phase 2 compost after the addition of mycelium; MCPR3 = Phase III compost; MCSP3 = SMC after 2 flushes. All samples were obtained from Monaghan mushrooms.

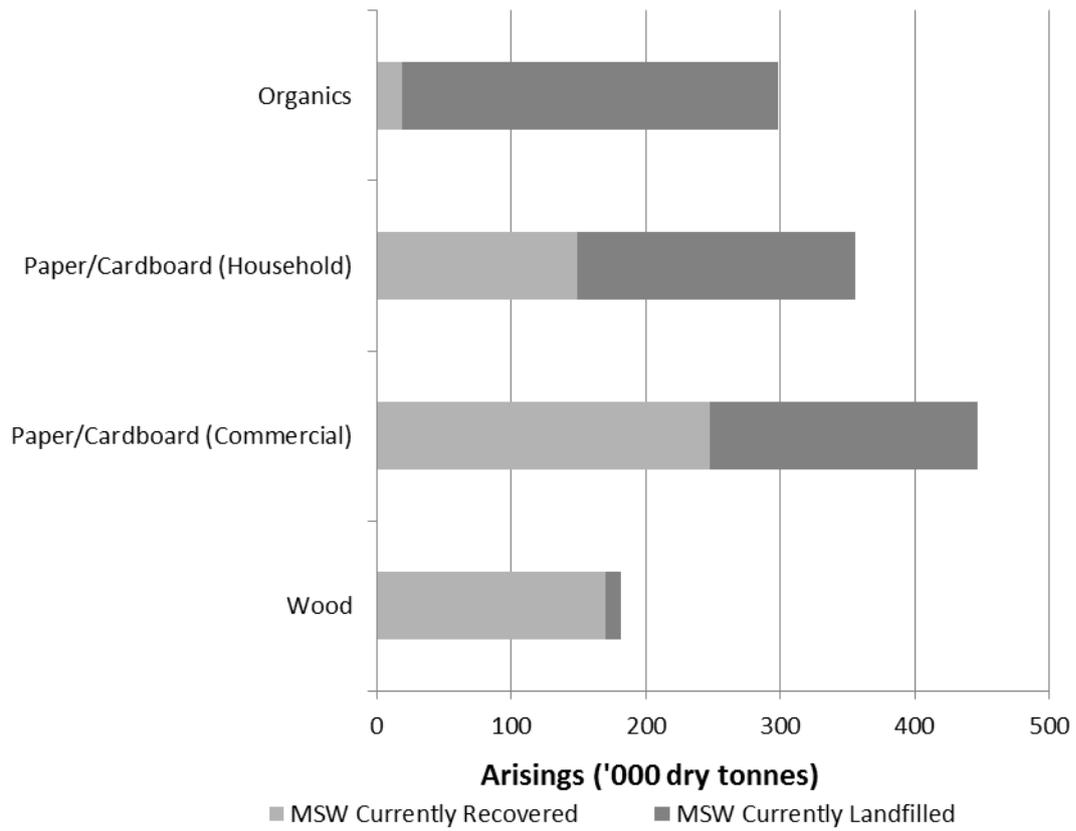


Figure H-3: The total arisings (in '000 dry tonnes) of BMW that was either recovered or landfilled in Ireland in 2005. Uses data from (EPA, 2006).

Table H-13: Extractives, ash, and lignocellulosic data for various samples of paper and cardboard wastes. Av = average; SD = standard deviation; KL = Klason lignin; ASL = acid soluble lignin; AIR = acid insoluble residue; AIA = acid insoluble ash. All these are expressed on a % whole mass (dry matter) basis.

Sample	Extractives		Ash		KL		ASL		AIR		AIA		Arabinose		Galactose		Rhamnose		Glucose		Xylose		Mannose		Tot Sugars	
	Av	SD	Av	SD	Av	SD	Av	SD	Av	SD	Av	SD	Av	SD	Av	SD	Av	SD	Av	SD	Av	SD	Av	SD	Av	SD
PCWC1, food wraps	2.34	(0.44)	17.21	(0.20)	14.79	(0.33)	0.61	(0.02)	19.87	(0.28)	5.08	(0.04)	0.49	(0.00)	0.77	(0.00)	0.09	(0.00)	49.96	(0.05)	7.73	(0.05)	6.35	(0.19)	65.39	(0.19)
PCBB1, brown bags	2.08	(0.07)	4.64	(0.16)	7.85	(0.17)	0.86	(0.01)	9.47	(0.32)	1.62	(0.15)	0.59	(0.00)	0.54	(0.01)	0.06	(0.00)	70.32	(0.38)	9.35	(0.19)	5.09	(0.02)	85.95	(0.60)
PCWB1, white bags	0.80	(0.38)			25.65	(0.63)	0.56		26.18	(0.70)	0.53	(0.07)	0.60	(0.03)	0.41	(0.36)	0.07	(0.00)	57.24	(0.67)	6.27	(0.09)	6.97	(0.22)	71.56	(1.36)
PCEN1, envelopes	2.65	(1.51)	8.93	(0.26)	2.00	(0.06)	0.74	(0.00)	2.44	(0.01)	0.44	(0.05)	0.22	(0.00)	0.17	(0.00)	0.05	(0.00)	67.06	(1.31)	13.74	(0.41)	2.30	(0.00)	83.54	(1.73)
PCGP1, glossy paper	0.98	(0.19)	19.57	(0.39)	6.29	(0.30)	0.61	(0.01)	10.21	(0.06)	3.93	(0.24)	0.28	(0.01)	0.33	(0.01)	0.06	(0.01)	58.76	(0.34)	8.37	(0.13)	3.87	(0.03)	71.67	(0.50)
PCPF1, food boxes	1.40	(0.07)	8.87	(0.41)	10.97	(0.08)	0.67		13.25	(0.15)	2.28	(0.06)	0.63	(0.01)	0.68	(0.02)	0.06	(0.00)	63.36	(0.41)	9.84	(0.05)	5.60	(0.15)	80.16	(0.28)
PCCB1, cereal boxes	1.08	(0.03)	10.01	(0.68)	12.90	(0.08)	0.64	(0.01)	14.85	(0.16)	1.95	(0.08)	0.68	(0.02)	0.97	(0.03)	0.09	(0.00)	56.88	(0.71)	9.17	(0.08)	5.81	(0.15)	73.61	(0.99)
PCBC1, cards	1.12	(0.01)	13.41	(0.03)	5.87	(0.16)	1.16	(0.04)	7.12	(0.18)	1.25	(0.02)	0.22	(0.01)	0.28	(0.01)	0.09	(0.00)	64.53	(0.32)	15.35	(0.06)	2.02	(0.08)	82.51	(0.29)
PCTR1, till receipts	2.62	(0.40)	10.60	(0.48)	4.69	(0.12)	0.54	(0.00)	5.43	(0.16)	0.74	(0.28)	0.15	(0.00)	0.10	(0.00)	0.04	(0.00)	68.36	(0.23)	11.98	(0.10)	1.25	(0.00)	81.88	(0.32)
PCPP1, printouts	0.62	(0.39)	11.43	(1.48)	1.00	(0.01)	0.72	(0.00)	1.00	(0.01)	0.00	(0.00)	0.07	(0.00)	0.13	(0.00)	0.04	(0.01)	70.64	(0.24)	13.27	(0.03)	0.98	(0.04)	85.14	(0.31)
PCNP1, newspaper	2.63	(0.16)			26.44	(0.20)	0.53	(0.00)	26.49	(0.08)	-0.03	(0.12)	1.14	(0.01)	2.05	(0.01)	0.16	(0.00)	46.41	(0.12)	6.29	(0.02)	11.86	(0.14)	67.91	(0.29)
PCTP1, Tetrapak	2.03	(0.01)	1.17	(0.00)	1.65	(0.14)	0.64	(0.00)	1.65	(0.14)	0.00	(0.00)	0.42	(0.01)	0.18	(0.00)	0.04	(0.00)	77.64	(0.52)	11.70	(0.01)	4.56	(0.05)	94.54	(0.57)
Household	1.13		9.26		7.53		0.68		8.45		0.91		0.40		0.54		0.06		65.31		11.20		3.92		81.43	
Commercial	1.55		6.77		12.80		0.64		13.58		0.75		0.61		0.95		0.09		60.14		9.80		6.15		77.74	
Export	1.29		8.32		9.51		0.67		10.38		0.85		0.48		0.70		0.07		63.36		10.67		4.75		80.04	

Table H-14: Ash and elemental data for the DS and DF fractions of samples of various paper and cardboard wastes. The percentage that the DS fraction contributed to the DG sample is provided. This allows the calculation of a weighted average for the ash and elemental contents of each sample. Av = average; SD = standard deviation. Dry matter heating values are also provided for the combined sample.

Sample	DS Fraction (% Dry of Matter)										DF Fraction (% Dry of Matter)										% DS	DS and DF Combined							
	Ash		C		H		N		S		Ash		C		H		N		S			Ash (%)	C (%)	H (%)	N (%)	S (%)	HHV MJ/kg	LHV MJ/kg	
	AV	SD	Av	SD	Av	SD	Av	SD	Av	SD	AV	SD	Av	SD	Av	SD	Av	SD	Av	SD									
PCTP1	1.17	(0.00)	56.78	(2.88)	8.70	(0.37)	0.01	(0.01)	0.00	(0.00)	2.09	(0.16)	50.70	(0.05)	7.37	(0.02)	0.05	(0.01)	0.00	(0.00)	88.65	1.27	56.09	8.55	0.02	0.00	23.30	21.42	
PCCK1	0.78	(0.06)	63.37	(0.24)	9.03	(0.03)	3.44	(0.10)	0.00	(0.00)	1.74	(0.08)	61.46	(0.04)	8.79	(0.01)	4.59	(0.39)	0.00	(0.00)	81.88	0.95	63.02	8.99	3.65	0.00	25.56	23.58	
PCPD1	5.15	(0.08)	46.71	(0.16)	5.66	(0.00)	0.07	(0.02)	0.11	(0.06)	6.32	(0.01)	47.40	(0.13)	5.54	(0.03)	0.09	(0.00)	0.04	(0.01)	74.33	5.45	46.89	5.63	0.08	0.09	18.62	17.38	
PCBC1	13.41	(0.03)	42.81	(0.06)	5.26	(0.01)	0.29	(0.01)	0.10	(0.07)	36.31	(0.07)	33.75	(0.06)	4.05	(0.02)	0.74	(0.01)	0.04	(0.00)	70.34	20.20	40.12	4.90	0.43	0.08	15.76	14.68	
PCBB1	4.64	(0.16)	44.59	(0.12)	6.45	(0.05)	0.12	(0.03)	0.02	(0.01)	9.51	(0.03)	43.57	(0.05)	6.21	(0.00)	0.16	(0.00)	0.04	(0.00)	70.01	6.10	44.28	6.38	0.13	0.03	18.37	16.97	
PCPF1	8.87	(0.41)	43.21	(0.33)	6.29	(0.06)	0.04	(0.00)	0.03	(0.00)	25.85	(0.97)	39.13	(0.29)	5.58	(0.05)	0.11	(0.00)	0.05	(0.00)	63.22	15.11	41.71	6.03	0.06	0.03	17.13	15.80	
PCEN1	8.93	(0.26)	41.48	(0.02)	6.20	(0.02)	0.07	(0.03)	0.02	(0.00)	18.91	(0.32)	38.25	(0.12)	5.55	(0.04)	0.12	(0.00)	0.03	(0.00)	67.11	12.21	40.42	5.99	0.09	0.03	16.82	15.51	
PCTR1	10.60	(0.48)	42.07	(0.33)	6.00	(0.09)	0.08	(0.01)	0.00	(0.00)	20.81	(0.05)	40.35	(0.00)	5.61	(0.00)	0.13	(0.03)	0.01	(0.00)	93.77	11.24	41.97	5.98	0.09	0.00	17.28	15.97	
PCGP1	19.57	(0.39)	38.67	(0.80)	5.48	(0.21)	0.04	(0.03)	0.01	(0.00)	52.09	(1.38)	29.21	(1.40)	3.64	(0.37)	0.11	(0.01)	0.03	(0.01)	52.94	34.87	34.22	4.61	0.07	0.02	13.44	12.42	
PCWC1	17.21	(0.20)	40.87	(0.24)	5.70	(0.05)	0.11	(0.00)	0.01	(0.00)	47.65	(0.06)	31.11	(0.07)	3.80	(0.03)	0.25	(0.02)	0.04	(0.00)	73.05	25.41	38.24	5.19	0.15	0.01	15.25	14.11	
PCPP1	11.43	(1.48)	39.76	(0.06)	5.80	(0.03)	0.03	(0.01)	0.01	(0.01)	22.01	(0.23)	37.16	(0.02)	5.30	(0.01)	0.06	(0.02)	0.01	(0.00)	73.22	14.26	39.07	5.67	0.04	0.01	16.16	14.92	
PCCB1	10.01	(0.68)	43.92	(0.05)	6.14	(0.02)	0.08	(0.02)	0.01	(0.00)	17.62	(0.72)	41.17	(2.14)	5.57	(0.52)	0.09	(0.03)	0.02	(0.01)	63.08	12.44	43.04	5.96	0.08	0.01	17.54	16.23	
Household																						13.37					16.90	15.62	
Commercial																						10.46					17.67	16.37	
Export																						12.28					17.19	15.90	

Table H-15: Extractives, ash, and lignocellulosic data for various samples of green wastes and composted green wastes. Av = average; SD = standard deviation; KL = Klason lignin; ASL = acid soluble lignin; AIR = acid insoluble residue; AIA = acid insoluble ash. All these are expressed on a % whole mass (dry matter) basis. MC = moisture content (% wet basis).

Sample	Extractives		Ash		KL		ASL		AIR		AIA		Arabinose		Galactose		Rhamnose		Glucose		Xylose		Mannose		Tot Sugars		MC	
	Av	SD	Av	SD	Av	SD	Av	SD	Av	SD	Av	SD	Av	SD	Av	SD	Av	SD	Av	SD	Av	SD	Av	SD	Av	SD	Av	SD
GRMX1	12.64	(0.05)	11.42	(0.48)	11.45	(0.10)	6.76	(0.02)	13.23	(0.11)	1.77	(0.01)	3.16	(0.04)	2.05	(0.03)	0.45	(0.00)	23.19	(0.04)	9.18	(0.04)	0.71	(0.00)	38.74	(0.15)	80.28	(0.93)
GRMX2	3.73	(0.06)	14.52	(0.37)	17.40	(0.11)			24.68	(0.04)	7.28	(0.07)	4.00	(0.05)	2.36	(0.01)	0.39	(0.00)	28.86	(0.00)	16.32	(0.12)	0.99	(0.09)	52.93	(0.09)	68.83	(6.09)
TREG1BL	5.25	(0.16)			31.69	(0.21)	0.89	(0.01)	31.69	(0.26)	0.03	(0.05)	1.55	(0.00)	4.96	(0.03)	0.28	(0.00)	32.47	(0.04)	6.94	(0.04)	6.46	(0.27)	52.67	(0.38)	52.64	(1.81)
TREG1LL	23.75	(4.36)	6.69	(0.07)	26.52	(0.03)	2.75	(0.00)	27.07	(0.08)	0.55	(0.05)	3.24	(0.08)	2.43	(0.03)	0.54	(0.01)	12.63	(0.06)	1.34	(0.04)	2.54	(0.08)	22.73	(0.03)	52.01	(0.22)
BUVR1LD	21.87	(0.10)	8.91	(0.12)	19.38	(0.00)	3.58	(0.03)	19.41	(0.00)	0.03	(0.00)	2.73	(0.01)	2.72	(0.01)	1.17	(0.01)	11.76	(0.03)	2.40	(0.01)	1.18	(0.00)	21.95	(0.04)	70.71	(0.10)
BUVR3LD	5.21	(0.02)			26.27	(0.24)	5.28	(0.06)	26.39	(0.12)	0.12	(0.12)	2.21	(0.04)	1.53	(0.00)	0.60	(0.01)	24.33	(0.23)	9.66	(0.19)	0.68	(0.01)	39.02	(0.38)		
BUVR5LD	14.41	(0.46)			20.54	(0.06)	4.58	(0.01)	20.60	(0.03)	0.12	(0.09)	1.48	(0.01)	1.86	(0.01)	1.47	(0.00)	17.76	(0.04)	2.66	(0.00)	1.60	(0.01)	26.83	(0.07)		
BUVR2TD	9.34	(0.09)	3.18	(0.27)	20.90	(0.38)	3.83	(0.03)	20.84	(0.34)	-0.03	(0.05)	2.31		1.34		0.69		30.28		13.29		1.08	(1.53)	50.08			
BUVR8TD	8.25	(0.09)	3.41	(0.21)	30.58	(0.01)	2.17	(0.01)	30.52	(0.06)	-0.03	(0.05)	2.08	(0.03)	1.83	(0.02)	0.61	(0.00)	24.50	(0.03)	12.26	(0.11)	1.14	(0.01)	42.42	(0.08)		
COMX1	8.99	(0.44)	12.01	(0.97)	27.43	(0.14)	0.70	(0.08)	31.43	(0.45)	4.00	(0.59)	2.44	(0.03)	2.23	(0.03)	0.74	(0.00)	19.85	(0.25)	6.25	(0.06)	2.03	(0.06)	33.54	(0.19)	64.14	(2.56)
COMX2	5.33	(0.06)	22.34	(1.89)	28.23	(0.57)			38.55	(0.03)	10.32	(0.54)	1.93	(0.02)	1.98	(0.03)	0.70	(0.02)	18.52	(0.80)	5.55	(0.38)	1.93	(0.04)	30.62	(1.25)	68.99	(0.33)
COMX3	2.43	(0.22)	25.77	(0.13)	31.10	(0.04)	1.95	(0.10)	39.43	(0.36)	8.33	(0.40)	0.87	(0.05)	1.81	(0.01)	0.35	(0.01)	24.27	(0.04)	8.87	(0.14)	2.70	(0.06)	38.88	(0.12)	57.84	(1.83)
COMX4	2.29	(0.31)	27.55	(3.22)	33.60	(0.17)	2.14	(0.03)	48.56	(0.57)	14.95	(0.74)	0.63	(0.02)	1.50	(0.09)	0.30	(0.01)	18.60	(0.42)	6.09	(0.24)	2.14	(0.04)	29.26	(0.72)	51.76	(1.27)
COMX5	1.63	(0.23)			34.43	(0.69)	2.51	(0.05)	55.12	(0.97)	20.69	(1.66)	0.69	(0.02)	1.01	(0.01)	0.32	(0.00)	11.73	(0.76)	3.89	(0.38)	1.27	(0.04)	18.91	(1.14)	50.43	(1.96)

Table H-16: Ash and elemental data for the DS and DF fractions of samples of green waste and composted green waste samples. The percentage that the DS fraction contributed to the DG sample is provided. This allows the calculation of a weighted average for the ash and elemental contents of each sample. Av = average; SD = standard deviation. Dry matter heating values are also provided for the combined sample.

Sample	DS Fraction (% Dry of Matter)										DF Fraction (% Dry of Matter)										% DS	DS and DF Combined							
	Ash		C		H		N		S		Ash		C		H		N		S			Ash (%)	C (%)	H (%)	N (%)	S (%)	HHV MJ/kg	LHV MJ/kg	
	AV	SD	Av	SD	Av	SD	Av	SD	Av	SD	AV	SD	Av	SD	Av	SD	Av	SD	Av	SD									
COMX1	12.01	(0.97)	48.42	(0.98)	6.43	(0.02)	1.42	(0.24)	0.08	(0.02)	15.91	(0.11)	45.71	(0.19)	6.22	(0.04)	1.53	(0.03)	0.09	(0.00)	77.15	12.76	47.80	6.38	1.45	0.08	19.15	17.74	
COMX3	25.77	(0.13)	40.80	(1.12)	5.32	(0.18)	2.23	(0.58)	0.10	(0.01)	30.56		37.99	(0.33)	4.79	(0.03)	3.98	(0.57)	0.13	(0.01)	74.18	29.34	40.07	5.18	2.68	0.11	15.72	14.58	
COMX4	27.55	(3.22)	43.03	(1.51)	6.08	(0.07)	1.39	(0.04)	0.14	(0.02)	36.96	(0.04)	35.33	(1.02)	5.11	(0.11)	1.88	(0.04)	0.20	(0.00)	71.67	30.22	40.85	5.81	1.53	0.16	16.25	14.98	
GRMX1	11.42	(0.48)	45.24	(0.09)	6.16	(0.00)	3.52	(0.03)	0.42	(0.04)	12.43	(0.46)	44.84	(0.13)	6.16	(0.01)	3.55	(0.00)	0.37	(0.00)	75.84	11.63	45.14	6.16	3.53	0.41	18.29	16.93	
GRMX2	14.52	(0.37)	42.19	(0.07)	5.67	(0.04)	1.20	(0.10)	0.05	(0.00)	34.29	(1.03)	34.16	(1.28)	4.66	(0.22)	1.41	(0.03)	0.11	(0.01)	63.10	21.75	39.23	5.30	1.28	0.07	15.72	14.55	

Table H-17: Extractives, ash, and lignocellulosic data for various brown bin waste and MSW samples and a waste wood sample. All these are expressed on a % whole mass (dry matter) basis. MC = moisture content (% wet basis).

Sample	Extractives		Ash		KL		AIR		AIA		Arabinose		Galactose		Rhamnose		Glucose		Xylose		Mannose		Tot Sugars		MC	
	Av	SD	Av	SD	Av	SD	Av	SD	Av	SD	Av	SD	Av	SD	Av	SD	Av	SD	Av	SD	Av	SD	Av	SD	Av	SD
MSW1	3.28		12.38	(0.89)			31.08				0.57	(0.01)	1.46	(0.00)	-	-	34.45	(0.89)	5.60	(0.02)	8.18	(0.43)	50.27	(1.30)		
MSW2	15.54		34.37				32.97	(1.60)			0.74	(0.04)	1.20	(0.09)	-	-	11.22	(0.83)	1.61	(0.12)	1.90	(0.20)	16.68	(1.28)		
MSW3	20.20		23.06				17.41	(1.37)			1.03	(0.10)	1.03	(0.09)	-	-	20.40	(2.80)	2.74	(0.25)	1.36	(0.16)	26.56	(3.01)		
MSW4	2.17		35.34	(1.09)			29.64	(0.12)			0.59	(0.00)	1.37	(0.02)	-	-	36.04	(0.69)	6.35	(0.12)	7.32	(0.01)	51.66	(0.86)		
MSW8	7.08		31.78	(2.54)	18.57	(0.26)	35.03	(0.26)	16.46	(0.01)	1.45	(0.02)	1.56	(0.07)	0.51	(0.11)	19.79	(0.22)	5.51	(0.06)	1.80	(0.01)	30.12	(0.05)		
MSW10	4.26		22.09		21.03	(1.32)	39.02	(2.37)	17.99	(3.69)	0.97	(0.02)	1.52	(0.09)	0.46	(0.10)	25.55	(0.33)	5.01	(0.04)	4.15	(0.01)	37.19	(0.26)		
MSW7	6.38		24.62		30.32	(0.19)	37.82	(1.83)	7.50	(2.02)	0.23	(0.00)	1.03	(0.01)	0.22	(0.04)	28.66	(1.37)	3.67	(0.14)	4.30	(0.07)	37.89	(1.58)		
MSW9	0.31		52.83		23.23	(3.08)	63.75	(5.95)	40.52	(9.03)	0.27	(0.04)	0.49	(0.11)	0.18	(0.04)	8.84	(2.12)	1.68	(0.39)	1.16	(0.34)	12.43	(3.00)		
Wood	1.44		6.23		28.42	(0.86)	29.65	(1.39)	1.22	(0.53)	0.77	(0.01)	1.56	(0.06)	0.17	(0.04)	36.94	(2.11)	6.77	(0.04)	7.87	(0.47)	53.90	(2.56)		
MSW6	6.34	(0.33)	14.60	(1.30)	19.21	(1.64)	29.15	(1.84)	9.93	(0.20)	0.45	(0.02)	0.69	(0.04)	0.09	(0.00)	40.98	(0.03)	6.87	(0.08)	3.30	(0.10)	52.38	(0.26)		
MSW11	6.63	(0.20)	19.51	(0.02)	25.78	(2.38)	30.15	(2.17)	4.37	(0.21)	0.19	(0.01)	0.37	(0.00)	0.09	(0.01)	35.94	(0.48)	4.26	(0.03)	2.77	(0.03)	43.63	(0.47)	31.23	(0.48)

Table H-18: Ash and elemental data for the DS and DF fractions of samples of brown bin waste samples. The percentage that the DS fraction contributed to the DG sample is provided. This allows the calculation of a weighted average for the ash and elemental contents of each sample. Av = average; SD = standard deviation. Dry matter heating values are also provided for the combined samples.

Sample	DS Fraction (% Dry of Matter)										DF Fraction (% Dry of Matter)										% DS	DS and DF Combined						
	Ash		C		H		N		S		Ash		C		H		N		S			Ash (%)	C (%)	H (%)	N (%)	S (%)	HHV MJ/kg	LHV MJ/kg
	AV	SD	Av	SD	Av	SD	Av	SD	Av	SD	AV	SD	Av	SD	Av	SD	Av	SD	Av	SD								
MSW1	12.38	(0.89)									34.46	(0.16)	36.10	(0.70)	3.89	(0.08)	1.96	(0.03)	2.12	(0.17)	36.17							
MSW2	34.37										59.36	(0.32)	29.06	(1.40)	2.20	(0.22)	1.81	(0.09)	2.01	(0.14)	19.17							
MSW3	23.06										27.24	(0.13)	44.67	(0.19)	4.70	(0.06)	2.75	(0.01)	2.28	(0.07)	26.66							
MSW4	35.34	(1.09)									36.34	(0.52)	34.63	(0.62)	3.80	(0.06)	1.79	(0.02)	1.97	(0.13)	28.15							
MSW8	31.78	(2.54)	34.76	(4.09)	3.72	(0.73)	2.72	(0.40)	2.48	(0.20)	33.66	(0.48)	36.05	(0.14)	3.86	(0.06)	2.54	(0.10)	2.66	(0.01)	66.6	32.41	35.19	3.76	2.66	2.54	13.22	12.39
MSW10	22.09		34.71	(9.74)	4.01	(1.31)	2.07	(0.20)	2.71	(0.01)	29.97		36.68	(0.69)	4.60	(0.14)	2.52	(0.00)	2.62	(0.03)	74.6	26.19	35.21	4.16	2.18	2.68	13.69	12.77
MSW7	19.04		43.23	(0.89)	5.19	(0.23)	1.86	(0.06)	2.50	(0.17)	24.62	(0.21)	38.56	(0.66)	4.29	(0.05)	2.78	(0.04)	2.55	(0.03)	62.1	21.16	41.46	4.85	2.21	2.52	16.07	15.01
MSW9	52.83		30.05	(2.68)	2.58	(0.66)	2.30	(0.26)	2.60	(0.15)	38.23		32.07	(0.68)	2.80	(0.10)	2.87	(0.04)	2.74	(0.17)	76.7	47.50	30.52	2.63	2.43	2.63	10.67	10.09

Table H-19: Expected biofuel yields (in litres, kg, and GJ/t) from processing various feedstocks in technologies A-H. See Table H-20 for details on these.

Feedstock	Litres of Product per Tonne of Feedstock										GJ/t									
	A	B	C	D	E	F (kg)	G	H	H (D)	H (N)	A	B	C	D	E	F	G	H	H (D)	H (N)
Spring Barley	232	334	359	281	433	332	423	168	131	37	4.89	7.03	7.55	5.91	9.12	6.83	8.90	5.66	4.50	1.16
Winter Barley	225	327	351	274	421	323	416	165	128	37	4.75	6.89	7.39	5.77	8.88	6.65	8.76	5.57	4.43	1.14
Spring Oats	232	337	361	282	433	332	421	167	130	37	4.88	7.10	7.61	5.94	9.12	6.84	8.88	5.64	4.49	1.15
Winter Oats	227	330	354	276	425	326	406	161	125	36	4.79	6.94	7.45	5.82	8.95	6.71	8.55	5.43	4.32	1.11
Spring Wheat	216	309	333	261	403	309	408	162	126	36	4.55	6.52	7.01	5.49	8.49	6.35	8.59	5.46	4.34	1.11
Winter Wheat	227	323	347	273	422	323	420	167	130	37	4.78	6.80	7.32	5.74	8.89	6.65	8.85	5.62	4.47	1.15
MNPG1	159	212		184	293	222					3.36	4.46		3.88	6.16	4.56				
COMX1							411										8.66			
GRMX2	179	256		216	334	256					3.78	5.39		4.55	7.04	5.26				
PCTP1	312	515	539	409	599	468	496	197	153	44	6.57	10.85	11.36	8.62	12.62	9.64	10.45	6.64	5.29	1.36
PCBC1	274	438	461	353	523	407					5.77	9.22	9.72	7.43	11.01	8.37				
PCBB1	283	471	492	373	544	426	393	156	121	35	5.96	9.91	10.37	7.85	11.47	8.77	8.28	5.26	4.19	1.07
PCPF1	264	436	457	347	508	397	366	145	113	32	5.57	9.19	9.62	7.30	10.70	8.17	7.71	4.90	3.90	1.00
PCEN1	277	448	470	359	529	413	359	143	111	32	5.83	9.43	9.91	7.56	11.15	8.49	7.57	4.81	3.83	0.98
PCTR1	271	442	464	353	519	405	370	147	114	33	5.70	9.32	9.78	7.44	10.93	8.33	7.79	4.95	3.94	1.01
PCGP1	236	391	410	311	454	355					4.97	8.25	8.63	6.54	9.56	7.31				
PCWC1	215	356	373	283	414	324					4.54	7.50	7.86	5.96	8.72	6.66				
PCPP1	282	458	481	366	540	421	346	137	107	30	5.93	9.65	10.14	7.72	11.37	8.66	7.28	4.63	3.68	0.94
PCCB1	243	400	419	318	466	364	376	149	116	33	5.11	8.42	8.82	6.70	9.82	7.49	7.92	5.03	4.00	1.03
Commercial Paper	269	441	462	351	516	403	379	151	117	33	5.66	9.29	9.74	7.40	10.87	8.29	7.99	5.08	4.04	1.04
Household paper	256	422	442	336	492	385	362	144	112	32	5.40	8.89	9.32	7.07	10.37	7.91	7.62	4.84	3.86	0.99
Exported Paper	264	434	455	345	507	396	368	146	114	32	5.56	9.14	9.58	7.28	10.68	8.15	7.76	4.93	3.92	1.01
Waste Wood	178	292	306	233	342	267	423	168	131	37	3.75	6.16	6.46	4.90	7.20	5.49	8.91	5.66	4.50	1.16
MSW6	173	284	297	226	331	259					3.64	5.97	6.26	4.76	6.98	5.33				
MSW11	143	240	251	189	276	216					3.01	5.05	5.28	3.99	5.81	4.45				

Table H-20: Expected total biofuel yields from processing the estimated national resources of straws and paper considered available for biorefining technologies. The yields are in million litres of ethanol for proceses A, B, C, D, E, and G, million kg of levulinic acid for process F, million litres of diesel from H (D), million litres of naphtha from H (N) and million litres of diesel and naphtha for process H. The yields are also expressed in energy terms (TJ). These total energy outputs from each technology are expressed as a percentage of total estimated petrol and diesel demand in Ireland in 2010. H (D) = FT-diesel; H (N) = FT-naphtha. A = near-term dlute acid hydrolysis process; B = advanced dilute acid hydrolysis process; C = near-term concentrated acid hydrolysis process; D = near-term enzymatic hydrolysis process; E = advanced enzymatic hydrolysis process; F = DIBANET process projections.

Feedstock	Dry Tonnes per Year	Million Litres of Product (million kg for Process F)										Terajoules (TJ)									
		A	B	C	D	E	F (kg)	G	H	H (D)	H (N)	A	B	C	D	E	F	G	H	H (D)	H (N)
Spring Barley	161,863	37.59	54.02	58.04	45.44	70.11	53.71	68.42	27.17	21.15	6.02	792	1,138	1,223	957	1,477	1,105	1,441	916	729	187
Winter Barley	19,317	4.35	6.32	6.78	5.29	8.14	6.24	8.03	3.19	2.48	0.71	92	133	143	112	171	128	169	108	86	22
Spring Oats	17,667	4.09	5.96	6.38	4.98	7.65	5.87	7.45	2.96	2.30	0.65	86	126	134	105	161	121	157	100	79	20
Winter Oats	18,502	4.21	6.10	6.54	5.11	7.86	6.03	7.51	2.98	2.32	0.66	89	128	138	108	166	124	158	101	80	21
Spring Wheat	30,318	6.55	9.38	10.09	7.91	12.22	9.35	12.36	4.91	3.82	1.09	138	198	213	167	257	192	260	165	132	34
Winter Wheat	70,830	16.06	22.86	24.60	19.31	29.90	22.87	29.74	11.81	9.19	2.62	338	482	518	407	630	471	627	398	317	81
Exported Paper	387,000	102.20	167.85	175.94	133.67	196.17	153.20	142.54	56.59	44.06	12.54	2,153	3,536	3,706	2,816	4,133	3,153	3,003	1,908	1,519	390
TOTAL	705,496	175	272	288	222	332	257	276	110	85	24	3,688	5,740	6,075	4,671	6,995	5,295	5,815	3,696	2,941	755
% of 2010 Transport Petrol and Diesel Fuels Energy Demand												1.76%	2.74%	2.90%	2.23%	3.34%	2.53%	2.78%	1.76%		
% of 2010 Transport Petrol and Diesel Fuels Energy Demand (Only Exported Paper is Biorefined)												1.03%	1.69%	1.77%	1.34%	1.97%	1.51%	1.43%	0.91%		

Table H-21: Extractives, ash, and lignocellulosic data for samples of switchgrass, willow short rotation coppice (SRC), a Miscanthus whole plant (WP) sampled in September (Misc-Sept), and a Miscanthus WP sampled in February (Misc-Feb). All data on a % whole mass (dry matter) basis.

Sample	Extractives		Ash		KL		ASL		AIR		AIA		Arabinose		Galactose		Rhamnose		Glucose		Xylose		Mannose		Tot Sugars	
	Av	SD	Av	SD	Av	SD	Av	SD	Av	SD	Av	SD	Av	SD	Av	SD	Av	SD	Av	SD	Av	SD	Av	SD	Av	SD
Switchgrass	12.64	(0.05)	11.42	(0.48)	11.45	(0.10)	6.76	(0.02)	13.23	(0.11)	1.77	(0.01)	3.16	(0.04)	2.05	(0.03)	0.45	(0.00)	23.19	(0.04)	9.18	(0.04)	0.71	(0.00)	38.74	(0.15)
Willow SRC	7.36	(0.15)	1.67	(0.09)	21.84	(0.19)	2.23	(0.02)	21.90	(0.22)	0.05	(0.02)	0.90	(0.06)	1.32	(0.01)	0.49	(0.01)	39.68	(0.43)	12.44	(0.01)	1.91	(0.05)	56.74	(0.39)
Misc-Sept	11.22	(0.11)	3.68	(0.08)	16.80	(0.00)	2.36	(0.02)	18.03	(0.01)	1.22	(0.01)	2.26	(0.05)	0.66	(0.02)	0.16	(0.00)	38.31	(0.21)	18.22	(0.06)	0.12	(0.00)	59.74	(0.09)
Misc-Feb	3.82	(0.08)	2.34	(0.10)	18.14	(0.07)	2.24	(0.02)	19.23	(0.10)	1.08	(0.03)	2.63	(0.08)	0.70	(0.03)	0.17	(0.00)	41.47	(0.45)	22.03	(0.04)	0.21	(0.01)	67.20	(0.29)

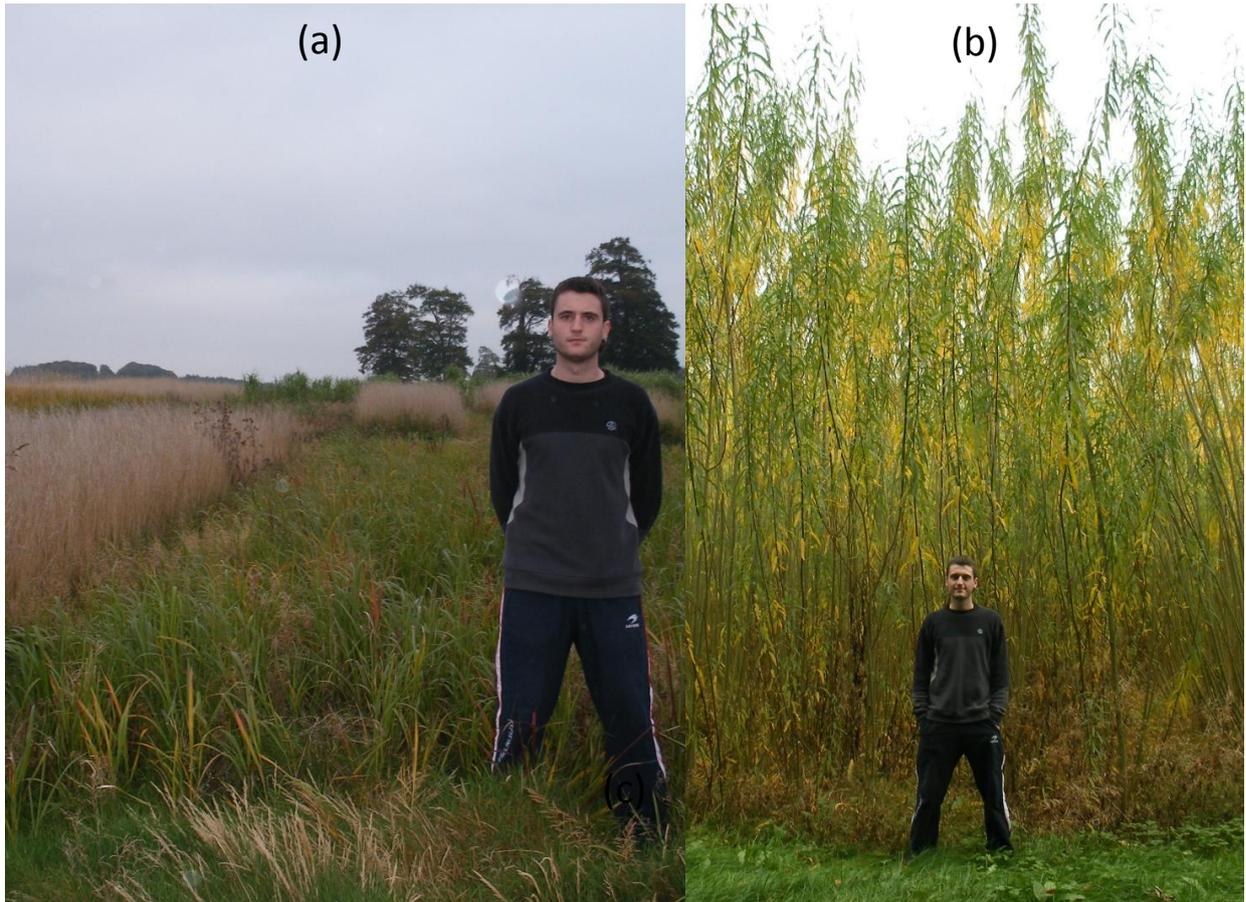


Figure H-4: Photographs of energy crop plantations in the Teagasc Oak Park Research Centre in Carlow. (a) Switchgrass (Shawnee variety); (b) short rotation coppice willow plantation.

Appendix I Figures and Tables for Chapter 18: The DIBANET Project

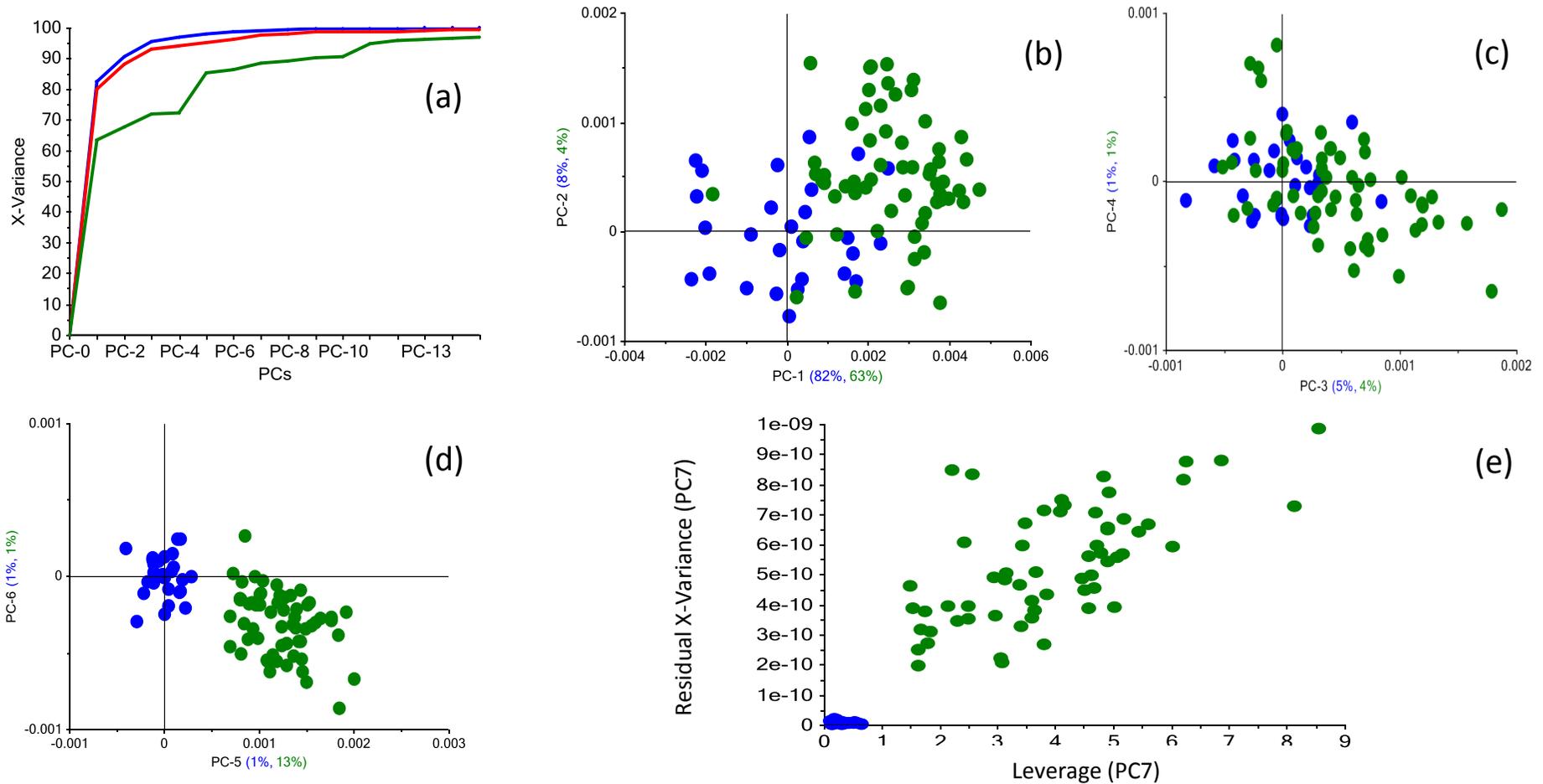


Figure I-1: Plots for the projection of the DS scans of the CTC sugarcane bagasse samples onto the BSES DS sugarcane bagasse model. (a) Explained X-variance plot (blue line = BSES samples in calibration, red line = BSES samples in cross-validation, green line = CTC samples); (b) PC1 vs. PC2 scores plot (blue dots = BSES samples, green dots = CTC samples); (c) PC3 vs. PC 4 scores plot; (d) PC5 vs. PC6 scores plot; (e) influence plot for a 7 PC model.

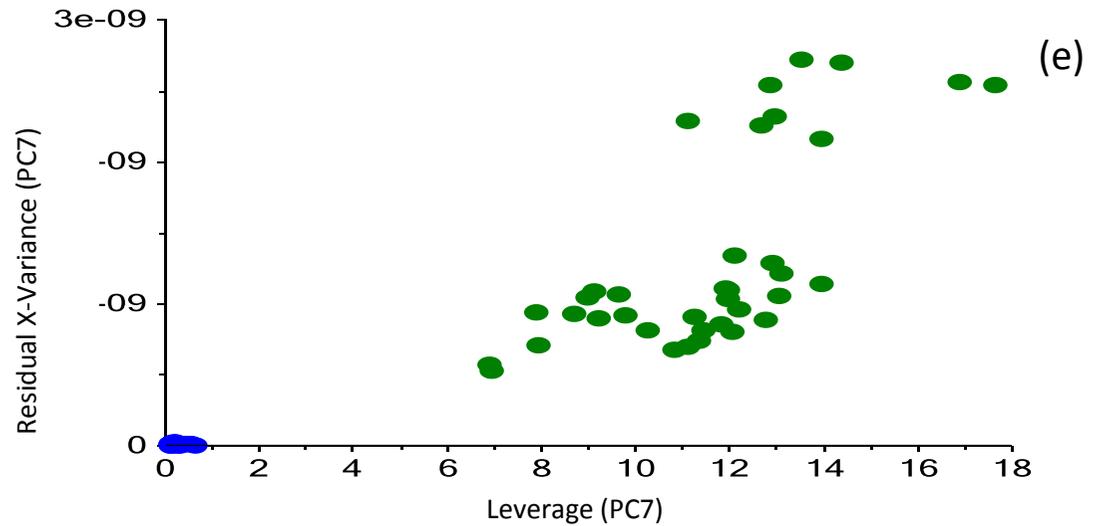
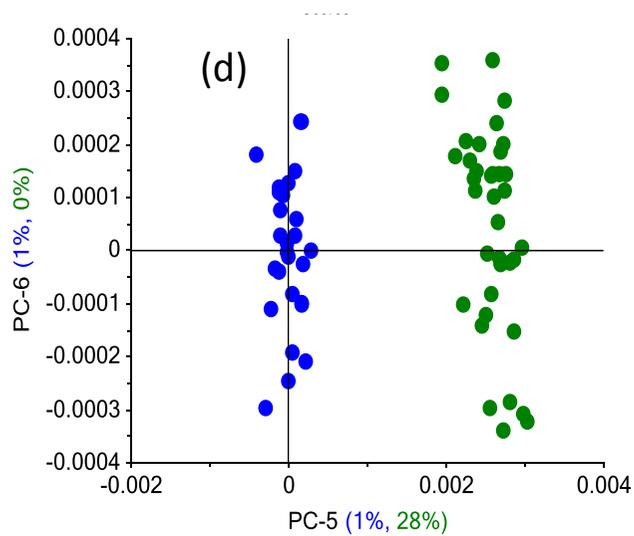
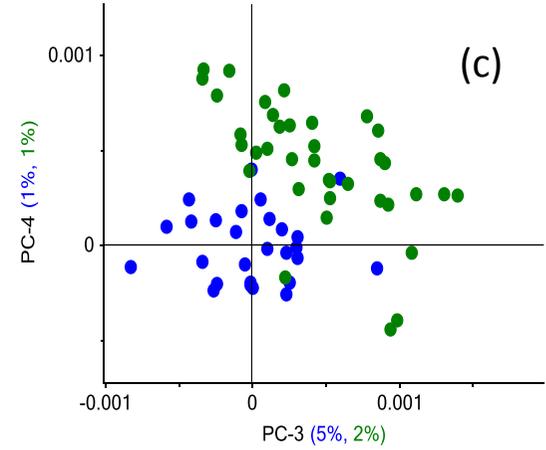
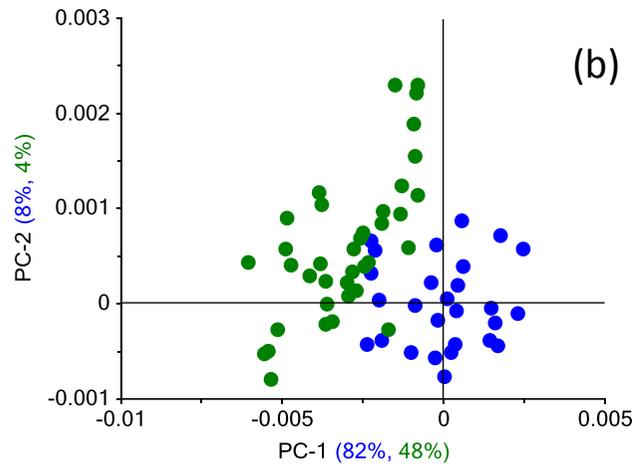
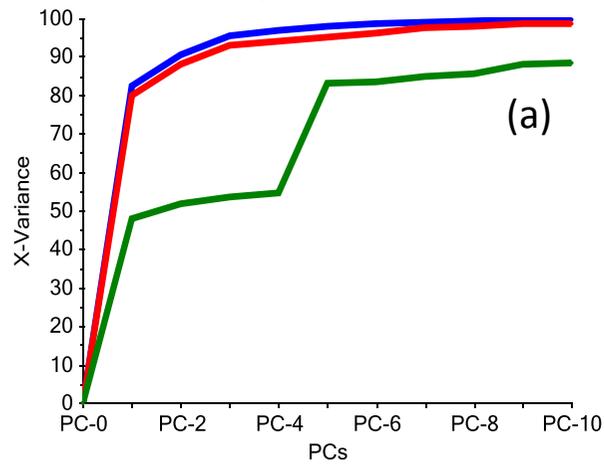


Figure I-2: Plots for the projection of the DS scans of the CTC sugarcane “trash” samples onto the BSES DS sugarcane bagasse model. (a) Explained X-variance plot (blue line = BSES samples in calibration, red line = BSES samples in cross-validation, green line = CTC samples); (b) PC1 vs. PC2 scores plot (blue dots = BSES samples, green dots = CTC samples); (c) PC3 vs. PC 4 scores plot; (d) PC5 vs. PC6 scores plot; (e) influence plot for a 7 PC model.

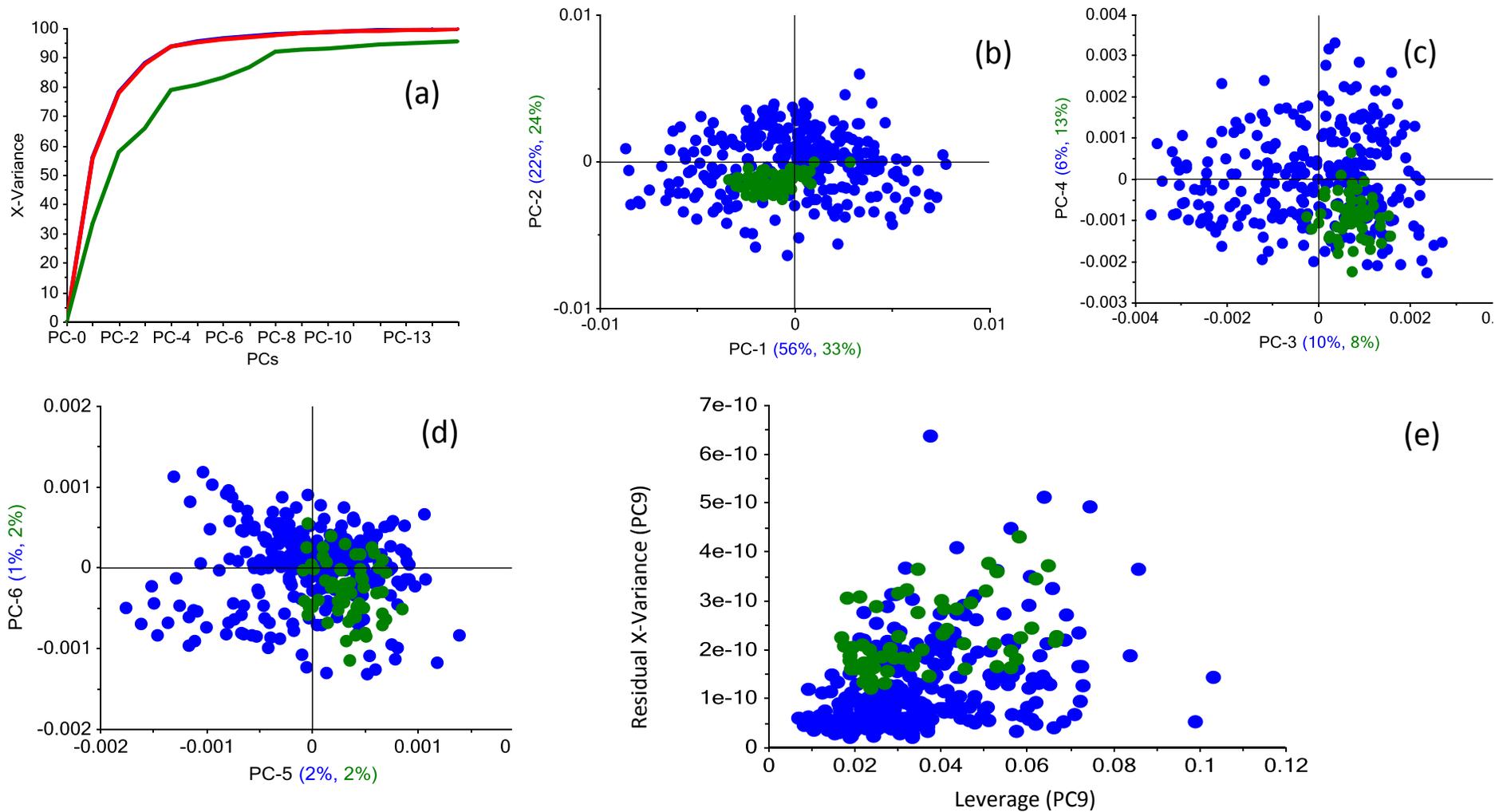


Figure I-3: Plots for the projection of the DS scans of the CTC sugarcane bagasse samples onto the UL Miscanthus DS model. (a) Explained X-variance plot (blue line = UL samples in calibration, red line = UL samples in cross-validation, green line = CTC samples); (b) PC1 vs. PC2 scores plot (blue dots = UL samples, green dots = CTC samples); (c) PC3 vs. PC 4 scores plot; (d) PC5 vs. PC6 scores plot; (e) influence plot for a 9 PC model.

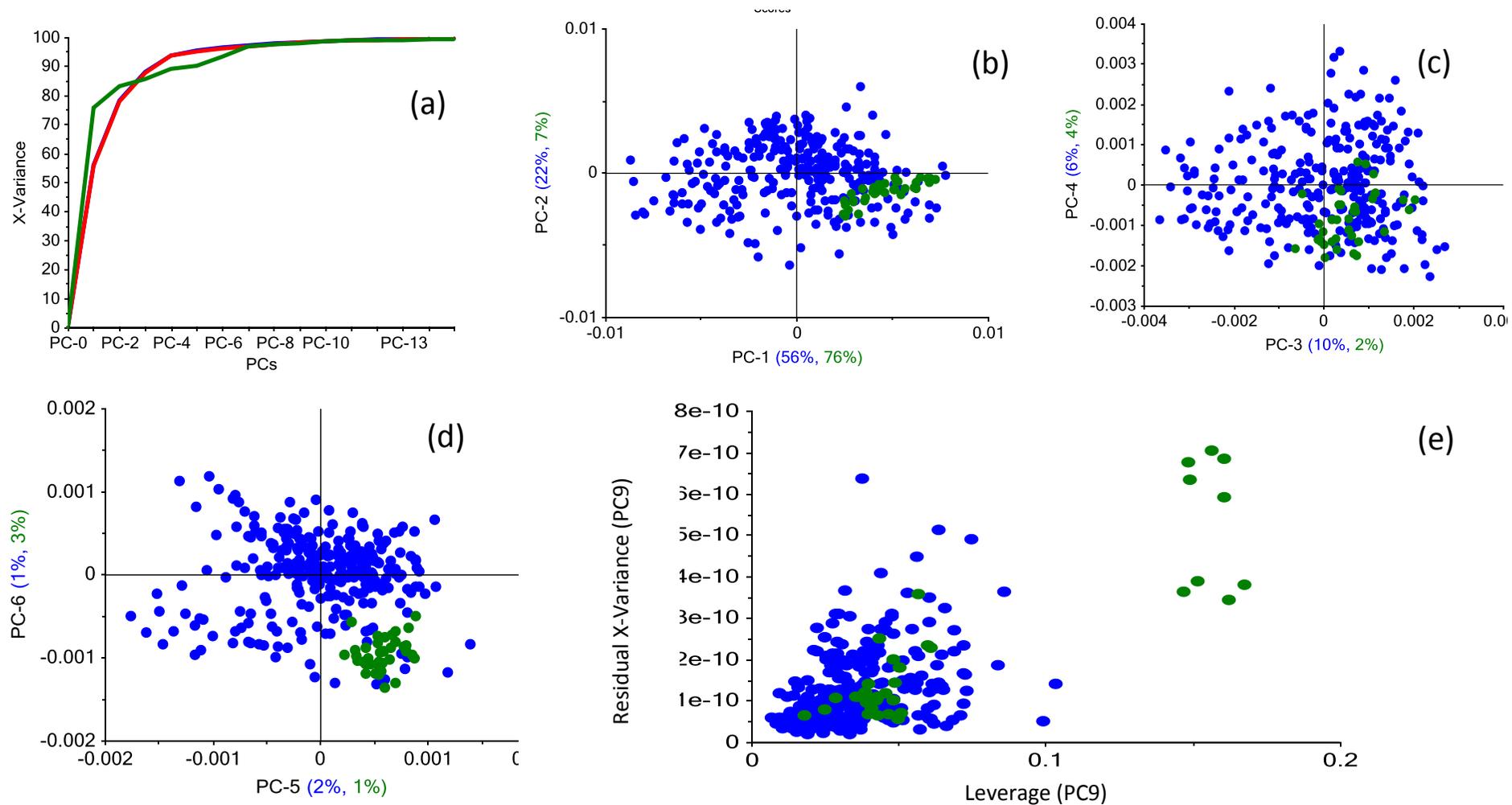


Figure I-4: Plots for the projection of the DS scans of the CTC sugarcane “trash” samples onto the UL *Miscanthus* DS model. (a) Explained X-variance plot (blue line = UL samples in calibration, red line = UL samples in cross-validation, green line = CTC samples); (b) PC1 vs. PC2 scores plot (blue dots = UL samples, green dots = CTC samples); (c) PC3 vs. PC 4 scores plot; (d) PC5 vs. PC6 scores plot; (e) influence plot for a 9 PC model.

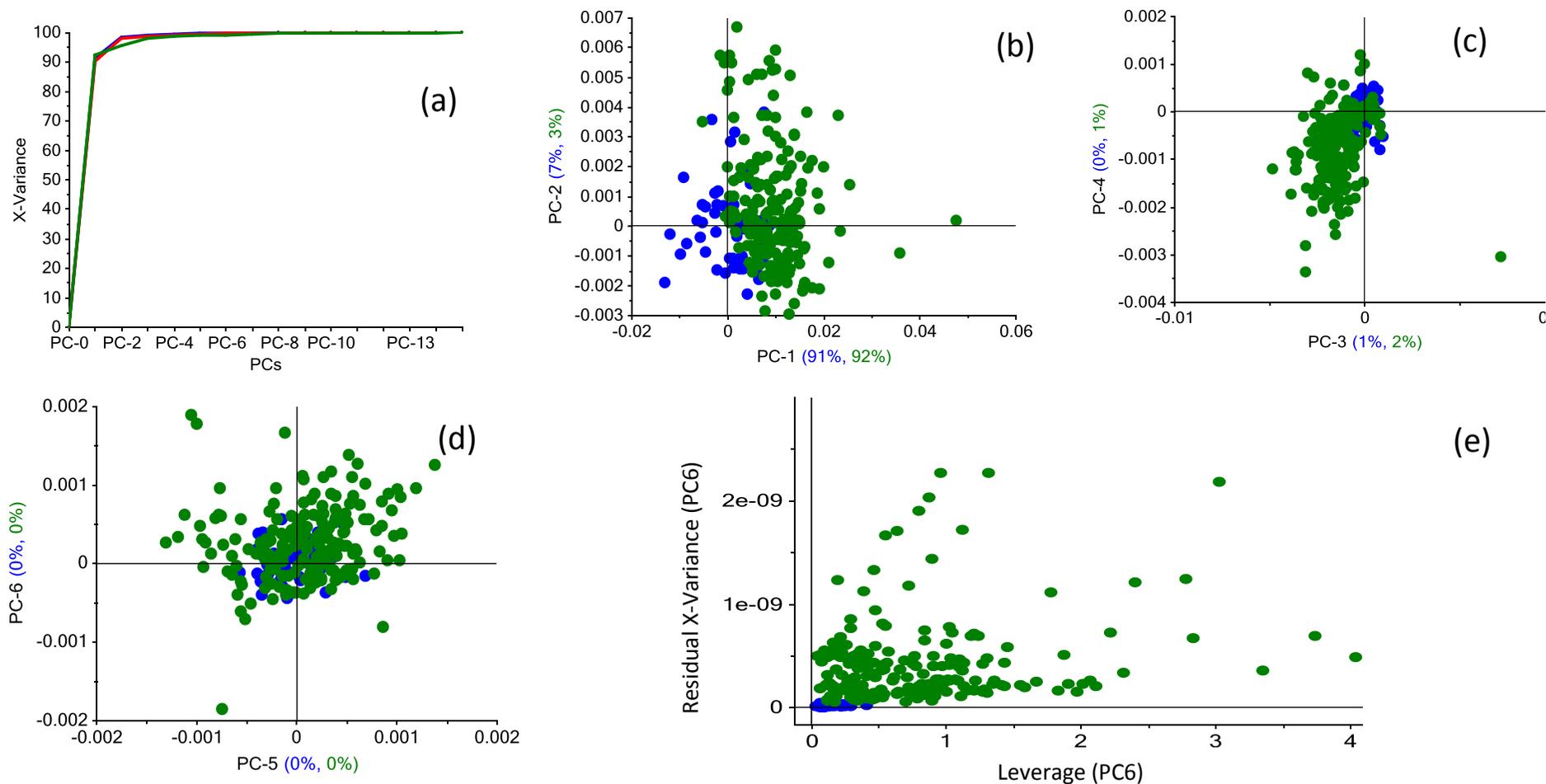


Figure I-5: Plots for the projection of the WU scans of the CTC sugarcane bagasse samples onto the BSES bagasse WU model. (a) Explained X-variance plot (blue line = BSES samples in calibration, red line = BSES samples in cross-validation, green line = CTC samples); (b) PC1 vs. PC2 scores plot (blue dots = BSES samples, green dots = CTC samples); (c) PC3 vs. PC 4 scores plot; (d) PC5 vs. PC6 scores plot; (e) influence plot for a 6 PC model.

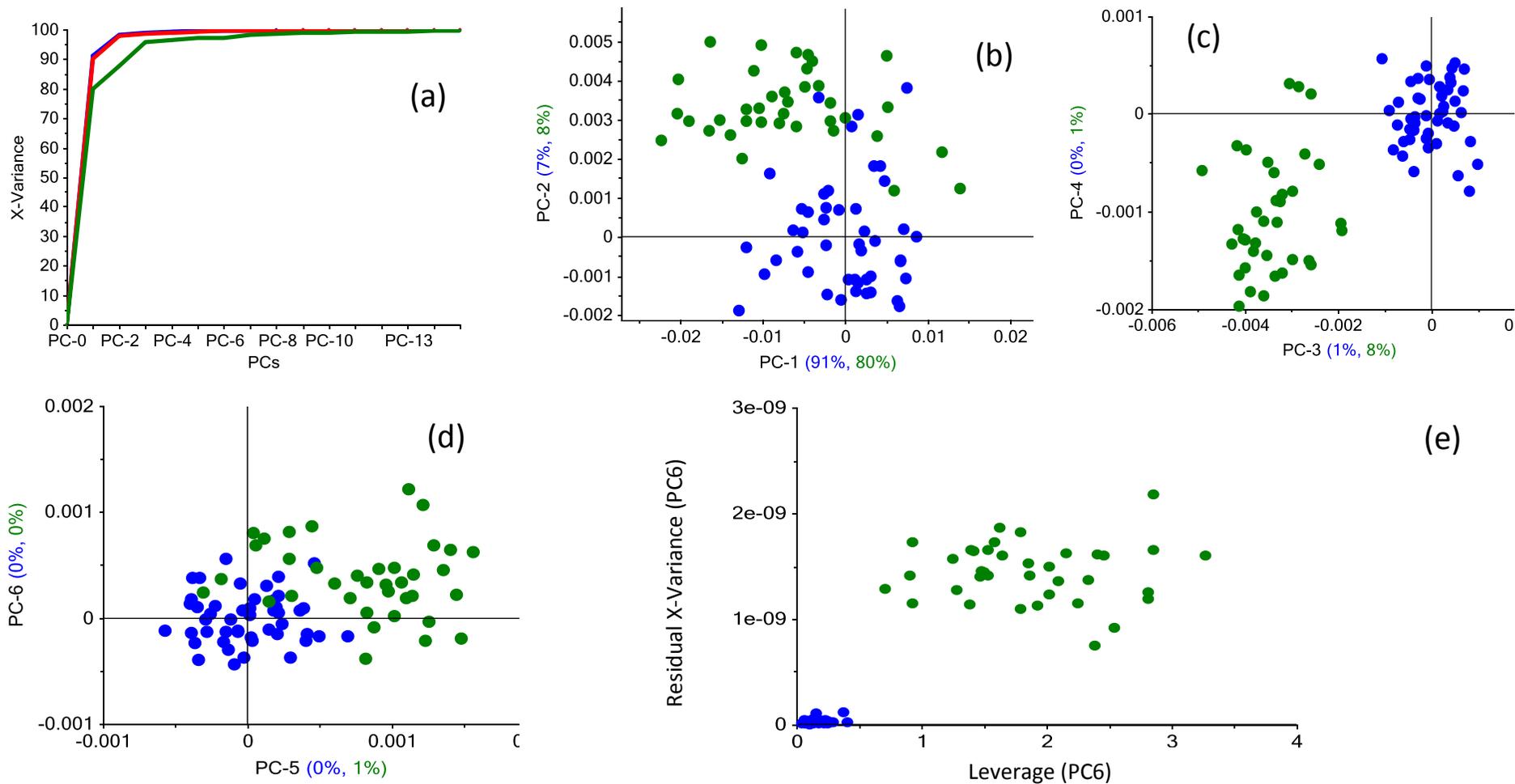


Figure I-6: Plots for the projection of the WU scans of the CTC sugarcane "trash" samples onto the BSES bagasse WU model. (a) Explained X-variance plot (blue line = BSES samples in calibration, red line = BSES samples in cross-validation, green line = CTC samples); (b) PC1 vs. PC2 scores plot (blue dots = BSES samples, green dots = CTC samples); (c) PC3 vs. PC 4 scores plot; (d) PC5 vs. PC6 scores plot; (e) influence plot for a 6 PC model.

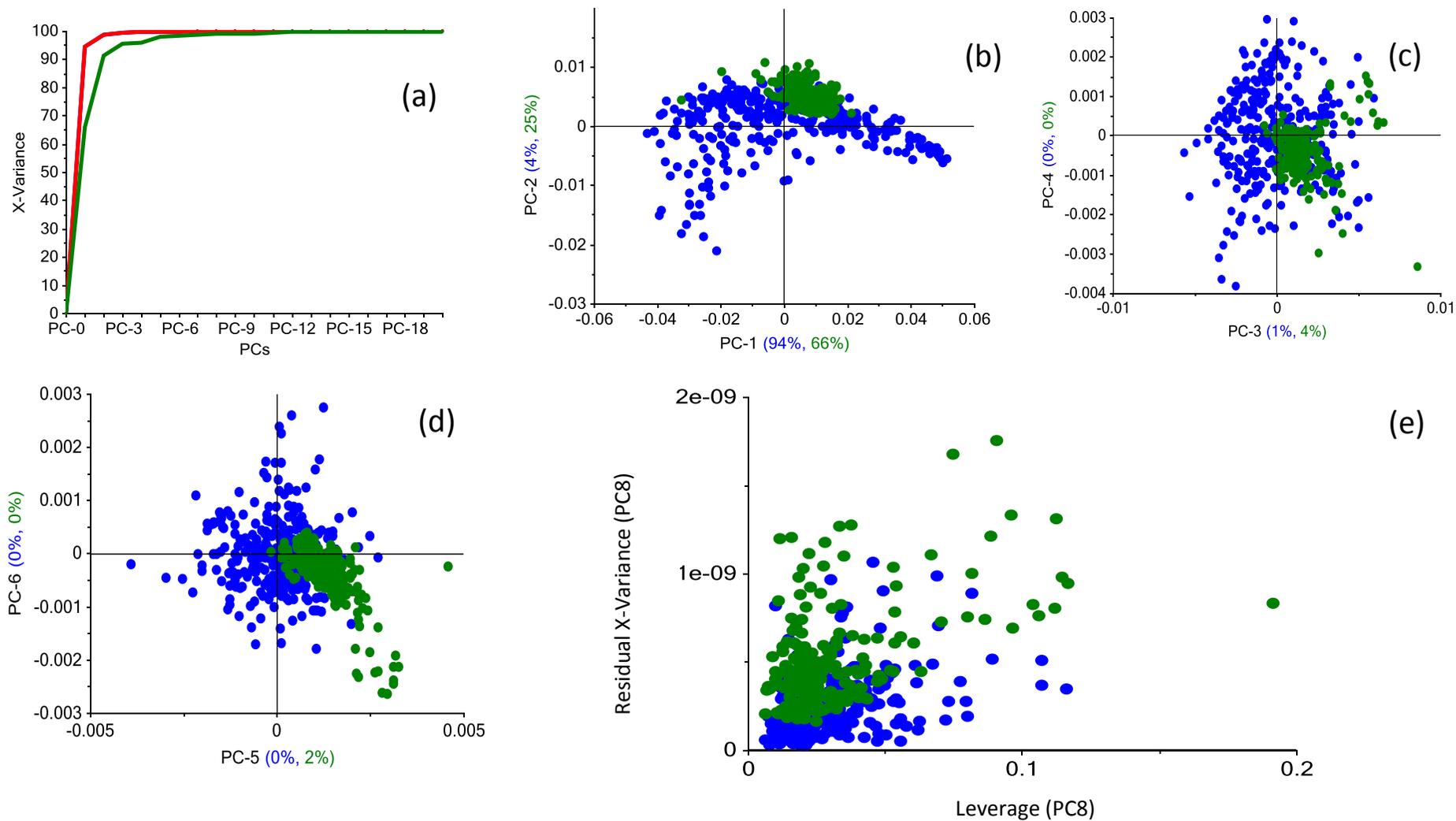


Figure I-7: Plots for the projection of the WU scans of the CTC sugarcane bagasse samples onto the UL *Miscanthus* WU model. (a) Explained X-variance plot (blue line = UL samples in calibration, red line = UL samples in cross-validation, green line = CTC samples); (b) PC1 vs. PC2 scores plot (blue dots = UL samples, green dots = CTC samples); (c) PC3 vs. PC 4 scores plot; (d) PC5 vs. PC6 scores plot; (e) influence plot for an 8 PC model.

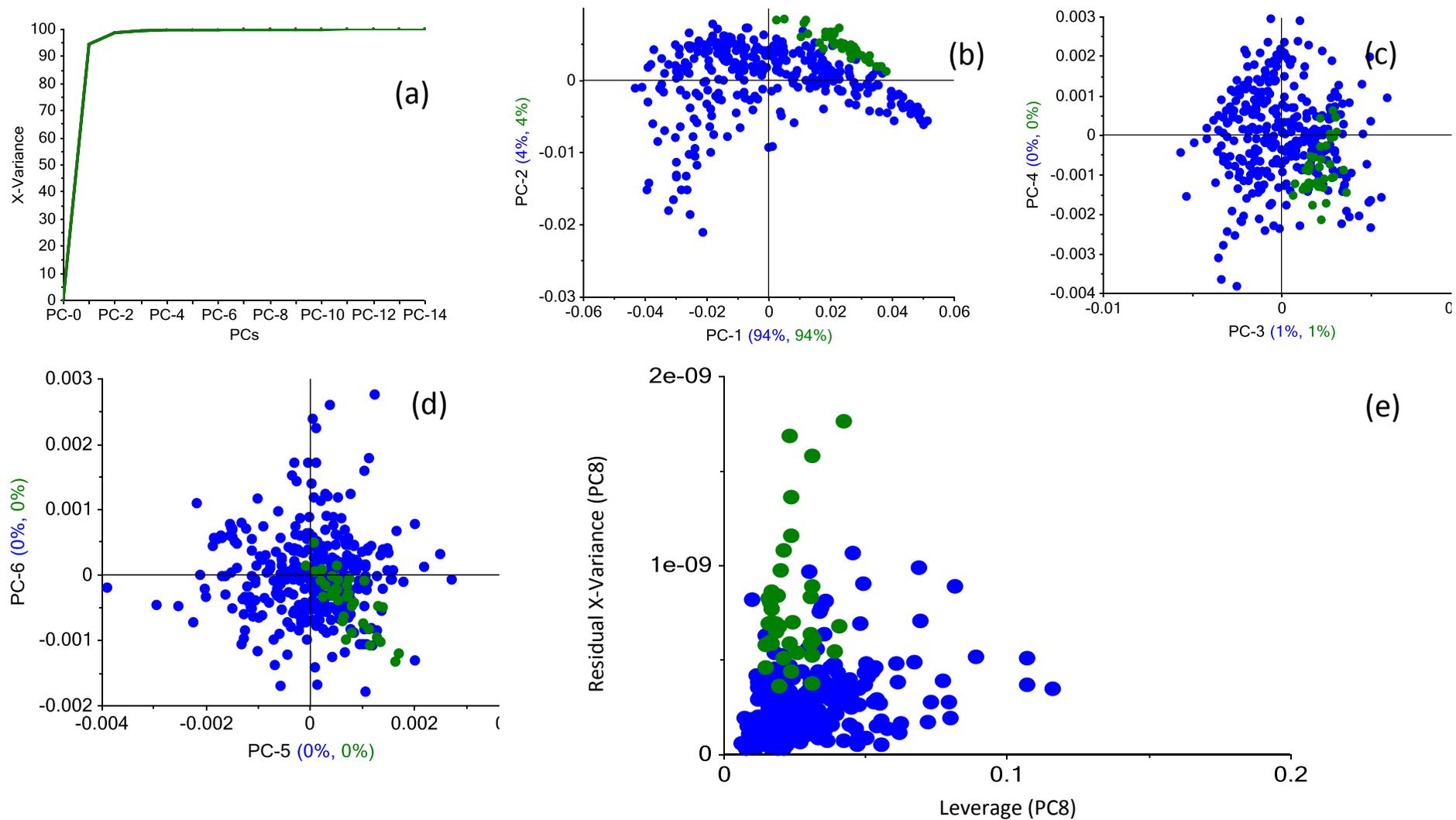


Figure I-8: Plots for the projection of the WU scans of the CTC sugarcane “trash” samples onto the UL *Miscanthus* WU model. (a) Explained X-variance plot (blue line = UL samples in calibration, red line = UL samples in cross-validation, green line = CTC samples); (b) PC1 vs. PC2 scores plot (blue dots = UL samples, green dots = CTC samples); (c) PC3 vs. PC 4 scores plot; (d) PC5 vs. PC6 scores plot; (e) influence plot for an 8 PC model.

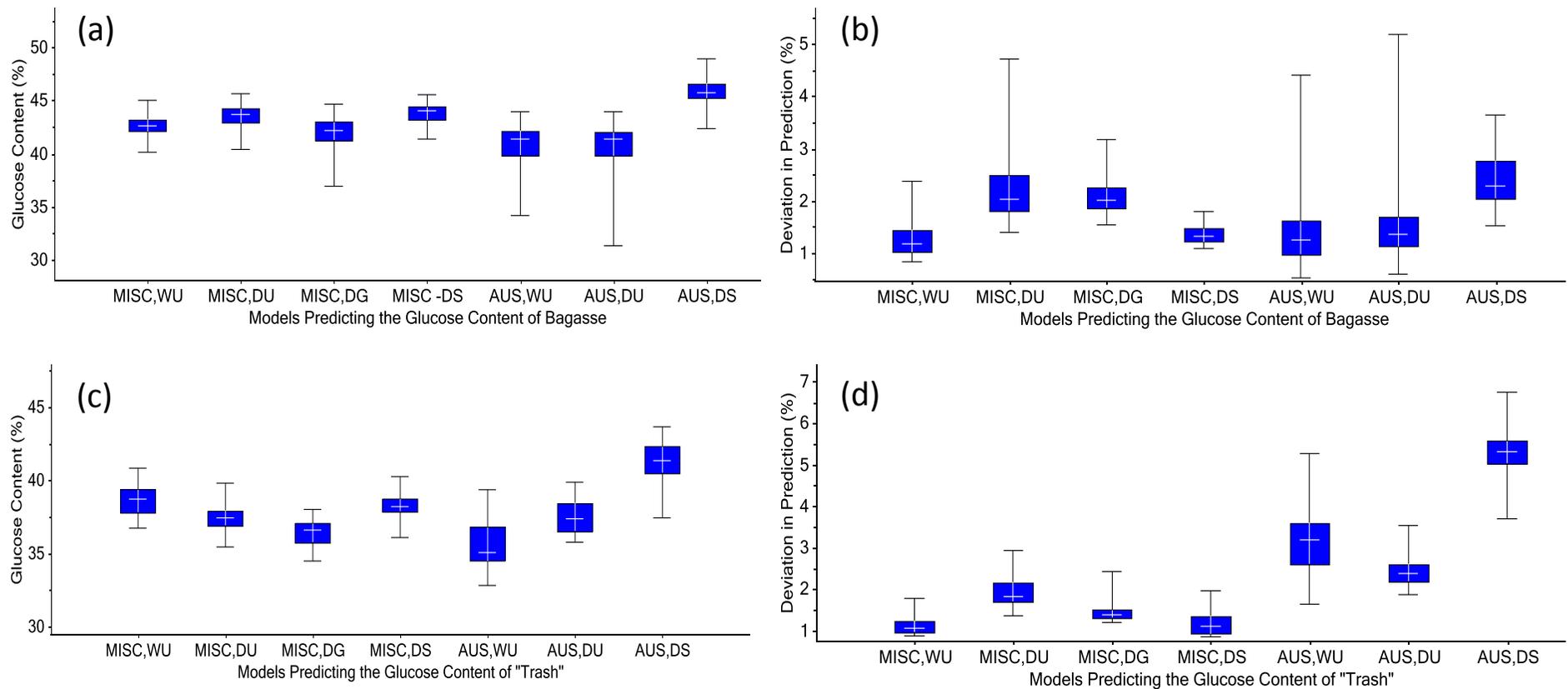
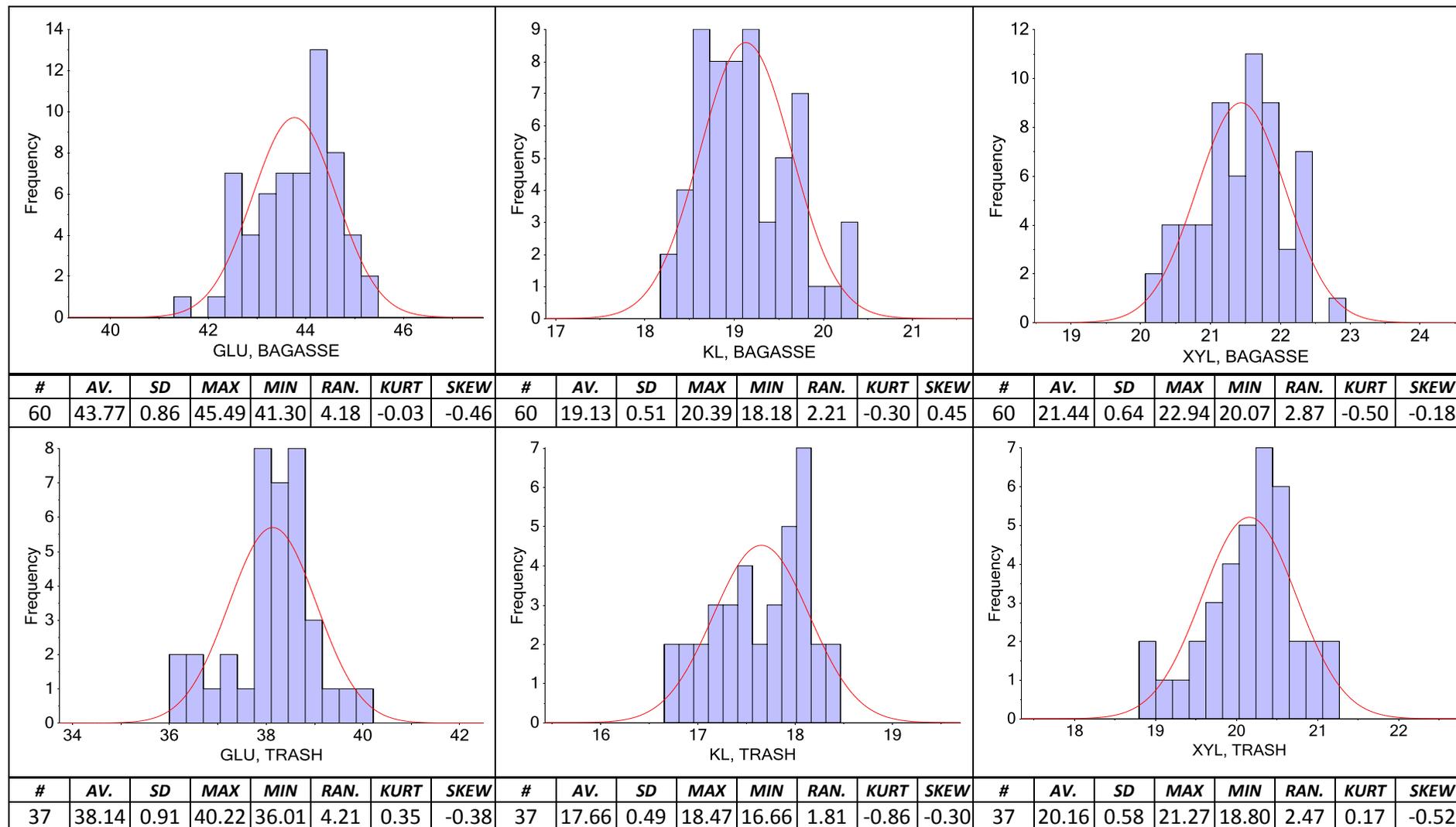


Figure I-9: : Quantile plots for predictions of the glucose contents of the CTC samples using either: the Miscanthus WU glucose model (MISC,WU); the Miscanthus DU glucose model (MISC,DU); the Miscanthus DG glucose model (MISC,DG); the Miscanthus DS glucose model (MISC, DS); the BSES sugarcane bagasse WU glucose model (AUS,WU); the BSES sugarcane bagasse DU glucose model (AUS,DU); or the BSES sugarcane bagasse DS glucose model (AUS,DS). (a) predictions of the glucose content of the CTC bagasse samples; (b) deviation in prediction of the glucose content of the CTC bagasse samples; (c) predictions of the glucose content of the CTC "trash" samples; (b) deviation in prediction of the glucose content of the CTC "trash" samples.

Table I-1: Histograms, with statistics, showing predicted glucose (GLU), Klason lignin (KL), and xylose (XYL) contents of CTC bagasse and trash DS samples. These predictions used the Miscanthus DS model. All figures are presented on a % whole mass dry-matter basis. RAN = range, KURT = kurtosis.



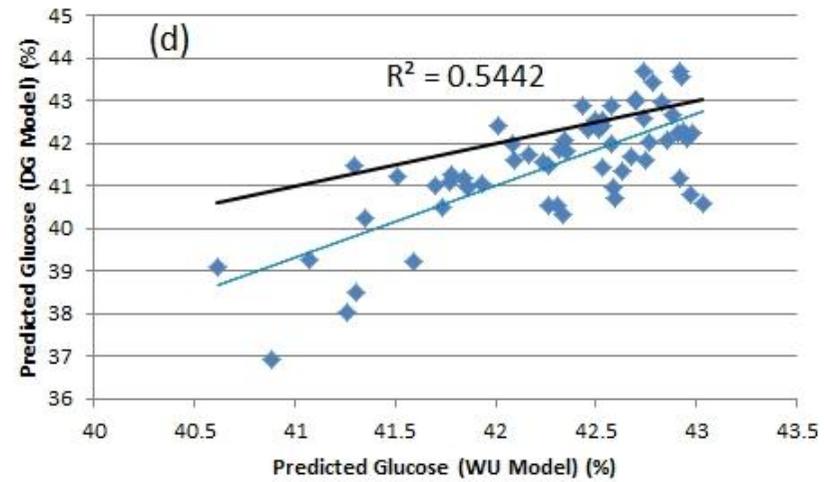
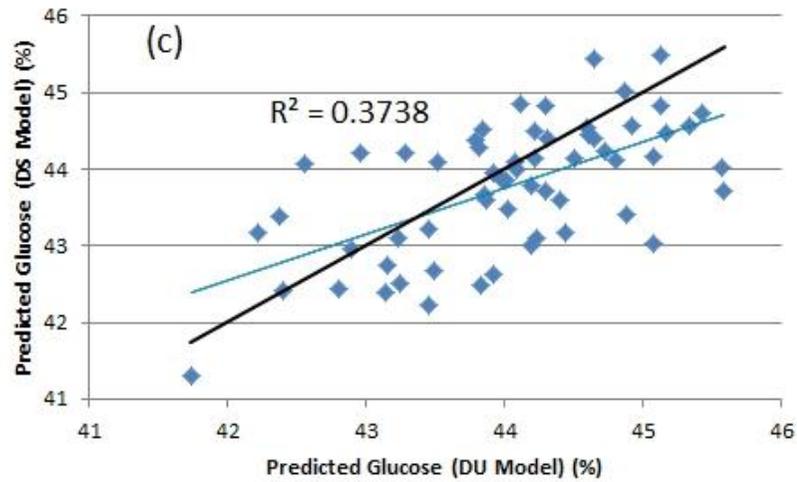
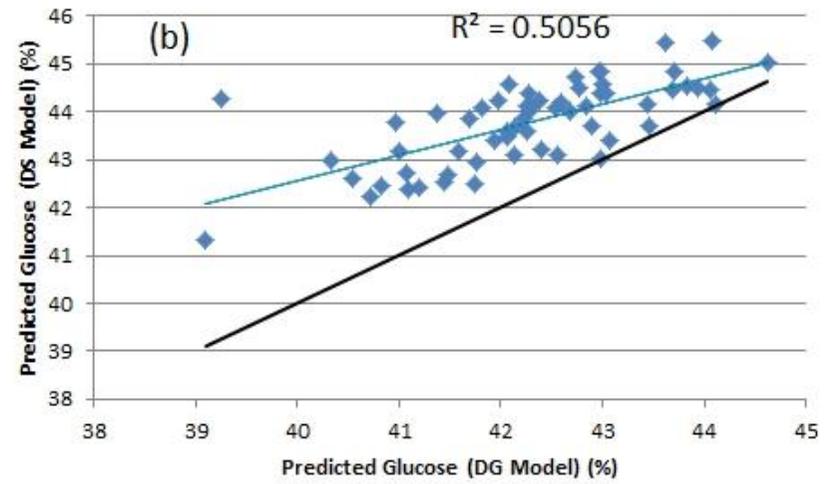
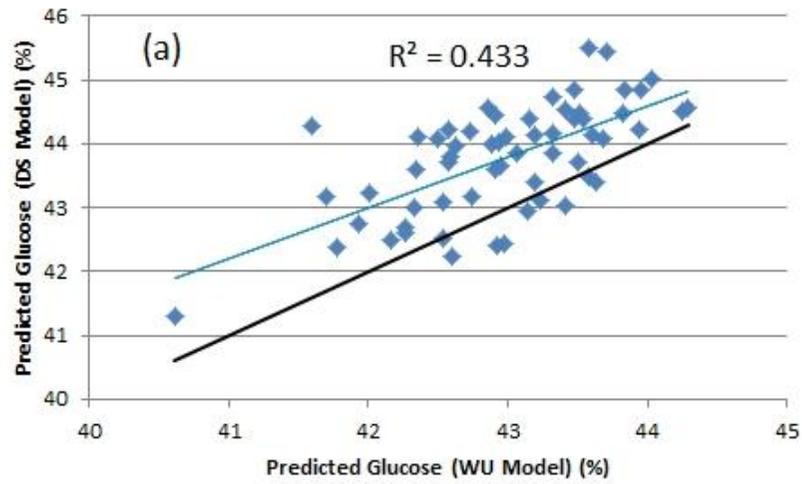


Figure I-10: Predicted glucose contents of CTC bagasse samples using *Miscanthus* models: (a) predictions from WU vs. those from DS model; (b) predictions from DG vs. those from DS model; (c) predictions from DU vs. those from DS model; (d) predictions from WU vs. those from DG model.

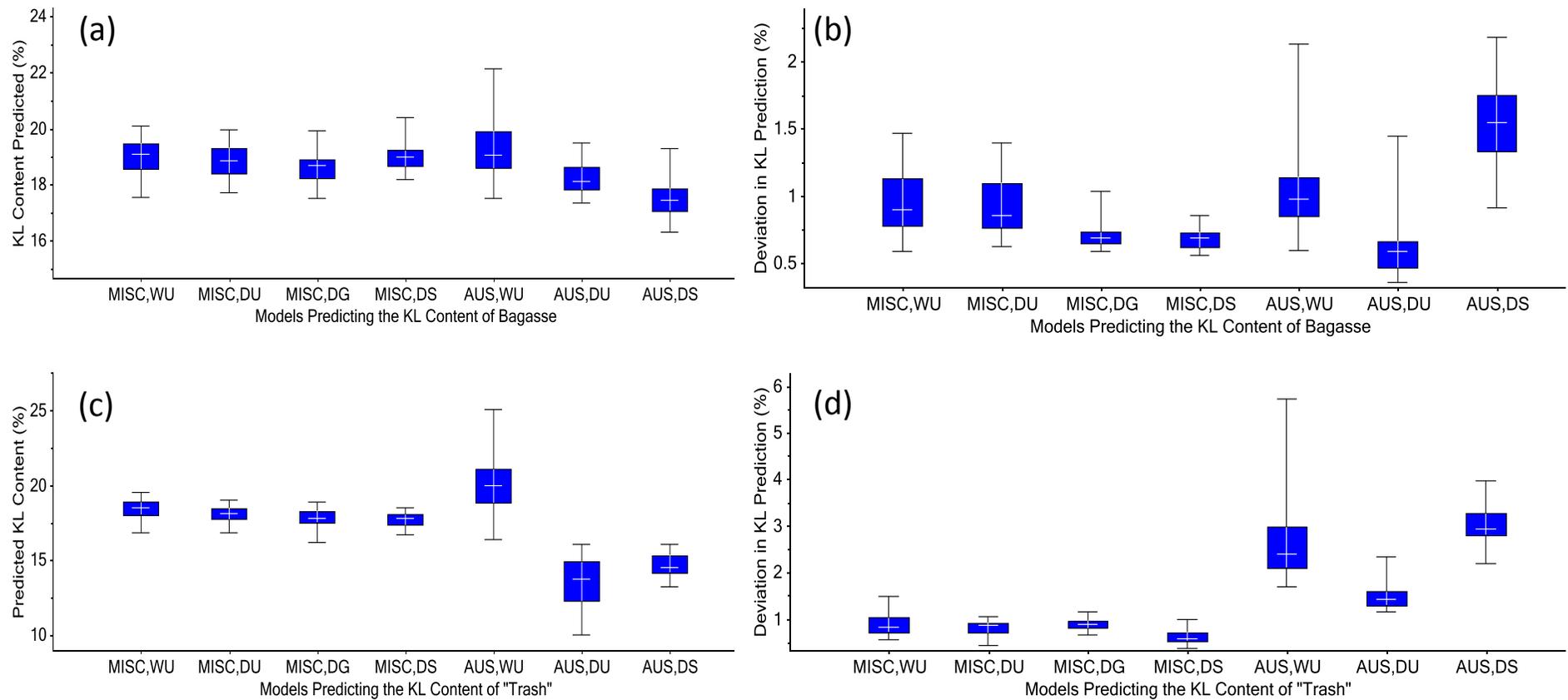


Figure I-11: Quantile plots for predictions of the KL (Klason lignin) contents of the CTC samples using either: the Miscanthus WU KL model (MISC,WU); the Miscanthus DU KL model (MISC,DU); the Miscanthus DG KL model (MISC,DG); the Miscanthus DS KL model (MISC,DS); the BSES sugarcane bagasse WU KL model (AUS,WU); the BSES sugarcane bagasse DU KL model (AUS,DU); or the BSES sugarcane bagasse DS KL model (AUS,DS). (a) predictions of the KL content of the CTC bagasse samples; (b) deviation in prediction of the KL content of the CTC bagasse samples; (c) predictions of the KL content of the CTC "trash" samples; (d) deviation in prediction of the KL content of the CTC "trash" samples.

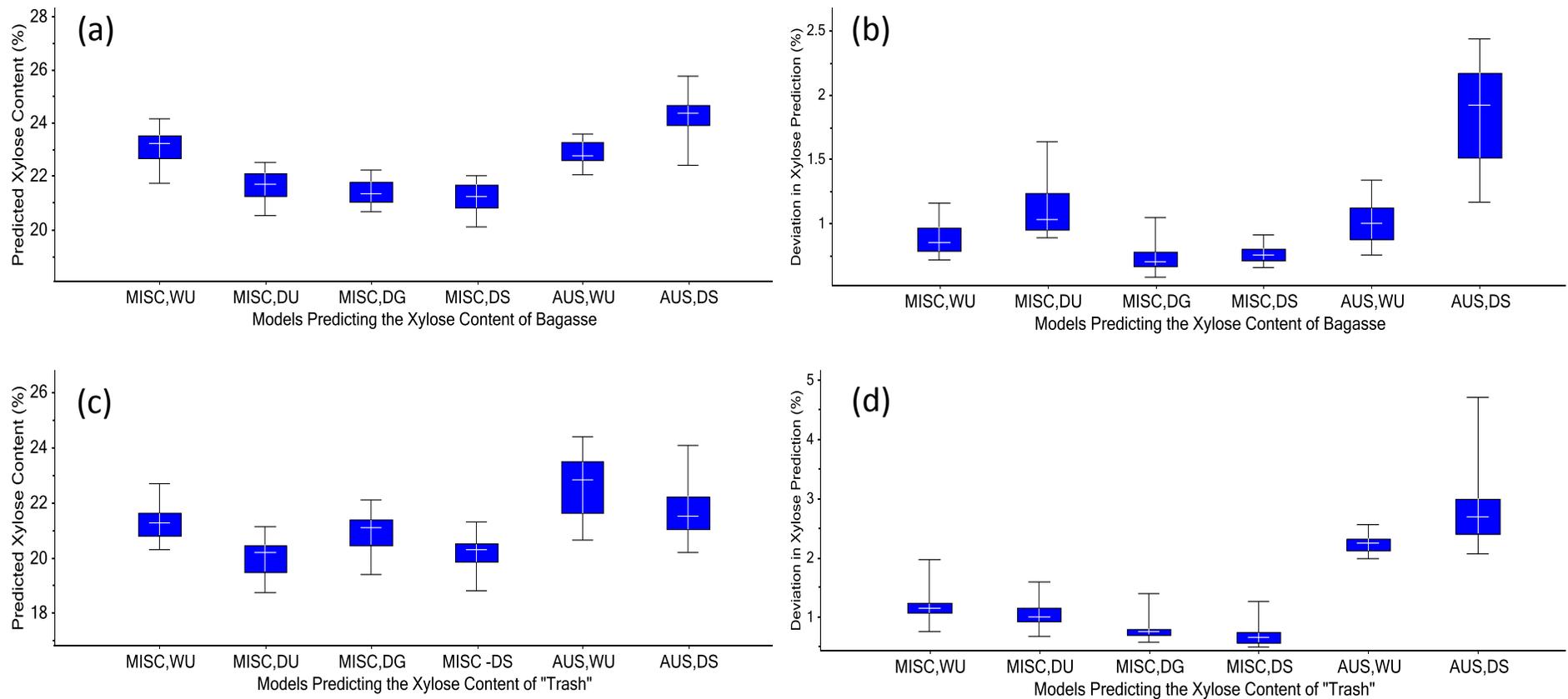


Figure I-12: Quantile plots for predictions of the xylose contents of the CTC samples using either: the Miscanthus WU xylose model (MISC,WU); the Miscanthus DU xylose model (MISC,DU); the Miscanthus DG xylose model (MISC,DG); the Miscanthus DS xylose model (MISC, DS); the BSES sugarcane bagasse WU xylose model (AUS,WU); or the BSES sugarcane bagasse DS xylose model (AUS,DS). (a) predictions of the xylose content of the CTC bagasse samples; (b) deviation in prediction of the xylose content of the CTC bagasse samples; (c) predictions of the xylose content of the CTC "trash" samples; (d) deviation in prediction of the xylose content of the CTC "trash" samples.

Table I-2: PLSR statistics for models developed on a dataset comprising the scans of 23 samples of pretreated *Miscanthus* and the scan of a sample of untreated *Miscanthus*. See Appendix A for descriptions of the terms used. All constituents expressed on % whole dry mass basis.

Constituent	GLU	XYL	ARA	GAL	RHA	MAN	TOT	KL	ASL	AIR	EXTR	EIA	ASH
Pretreatment	SG	SG	SG	SG	SG	SG	SG	SG	SG	SG	SG	SG	SG
Specific	2,2,25,25	2,2,25,25	2,2,25,25	2,2,25,25	2,2,25,25	2,2,25,25	2,2,25,25	2,2,25,25	2,2,25,25	2,2,25,25	2,2,25,25	2,2,25,25	2,2,25,25
PLS- λ 10 ³ nm	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5	1.1-2.5
Calib:Valid	24:0	24:0	24:0	24:0	24:0	24:0	24:0	24:0	24:0	24:0	24:0	24:0	24:0
F-Wold's	6	5	2	4	1	1	7	4	7	4	2	1	1
F-Wold 0.95	6	5	2	4	1	1	7	4	7	4	2	1	1
F-Wold 0.9	6	1	2	3	1	1	6	4	1	4	2	1	1
F-F Test	6	5	2	4	1	1	7	4	7	4	2	1	1
F-Haaland's	6	4	7	7	3	8	6	7	7	8	8	6	7
F.-Min Press	6	5	9	8	5	8	7	7	7	8	9	10	7
F.-UNSCR.	5	5	7	3	5	8	6	4	7	8	9	10	7
R^2_{calib}	0.9941	0.9727	0.9885	0.9823	0.8310	0.9575	0.9923	0.9943	0.9716	0.9916	0.9707	0.9245	0.9663
$Offset_{calib}$	0.3169	0.5324	0.0194	0.0081	0.0173	0.0072	0.5815	0.0401	0.0963	0.0688	0.1373	0.1285	0.0616
RMSEC (%)	0.7785	0.6707	0.0769	0.0232	0.0115	0.0074	0.5674	0.4308	0.1887	0.5312	0.3085	0.1169	0.1000
R^2_{CV}	0.9834	0.9393	0.9509	0.9229	0.3843	0.6974	0.9733	0.9776	0.9204	0.9700	0.5480	0.4348	0.7464
$Slope_{CV}$	0.9931	0.8936	0.9721	0.9854	0.4450	0.7307	0.9645	1.0059	0.9632	0.9827	0.4544	0.4478	0.6242
$Offset_{CV}$	0.2423	2.1994	0.0701	0.0137	0.0553	0.0495	2.8216	0.1108	0.1047	0.2814	2.2775	0.9168	0.6433
RMSECV (%)	1.3184	1.0256	0.1615	0.0503	0.0222	0.0201	1.0677	0.8831	0.3203	1.0169	1.2636	0.3209	0.2883
$BIAS_{CV}$	-0.1319	0.1258	0.0229	0.0070	-0.0016	0.0038	0.1228	0.1522	-0.0199	0.1396	-0.2775	-0.0235	-0.0441
SECV (%)	1.3400	1.0398	0.1633	0.0509	0.0226	0.0202	1.0834	0.8886	0.3265	1.0289	1.2592	0.3270	0.2911
RPD_{CV}	7.7507	3.9861	4.4905	3.5062	1.2646	1.8132	6.1138	6.5600	3.5012	5.7615	1.4617	1.3297	1.9118
RER_{CV}	29.5290	17.4939	15.3097	13.1538	6.6258	8.9050	22.9453	23.0821	13.0461	20.3127	8.1002	6.3923	8.7608
$RMSECV_{MP}$	1.3184	0.9306	0.1411	0.0447	0.0196	0.0201	1.0334	0.8831	0.3203	1.0169	1.2389	0.3026	0.2883

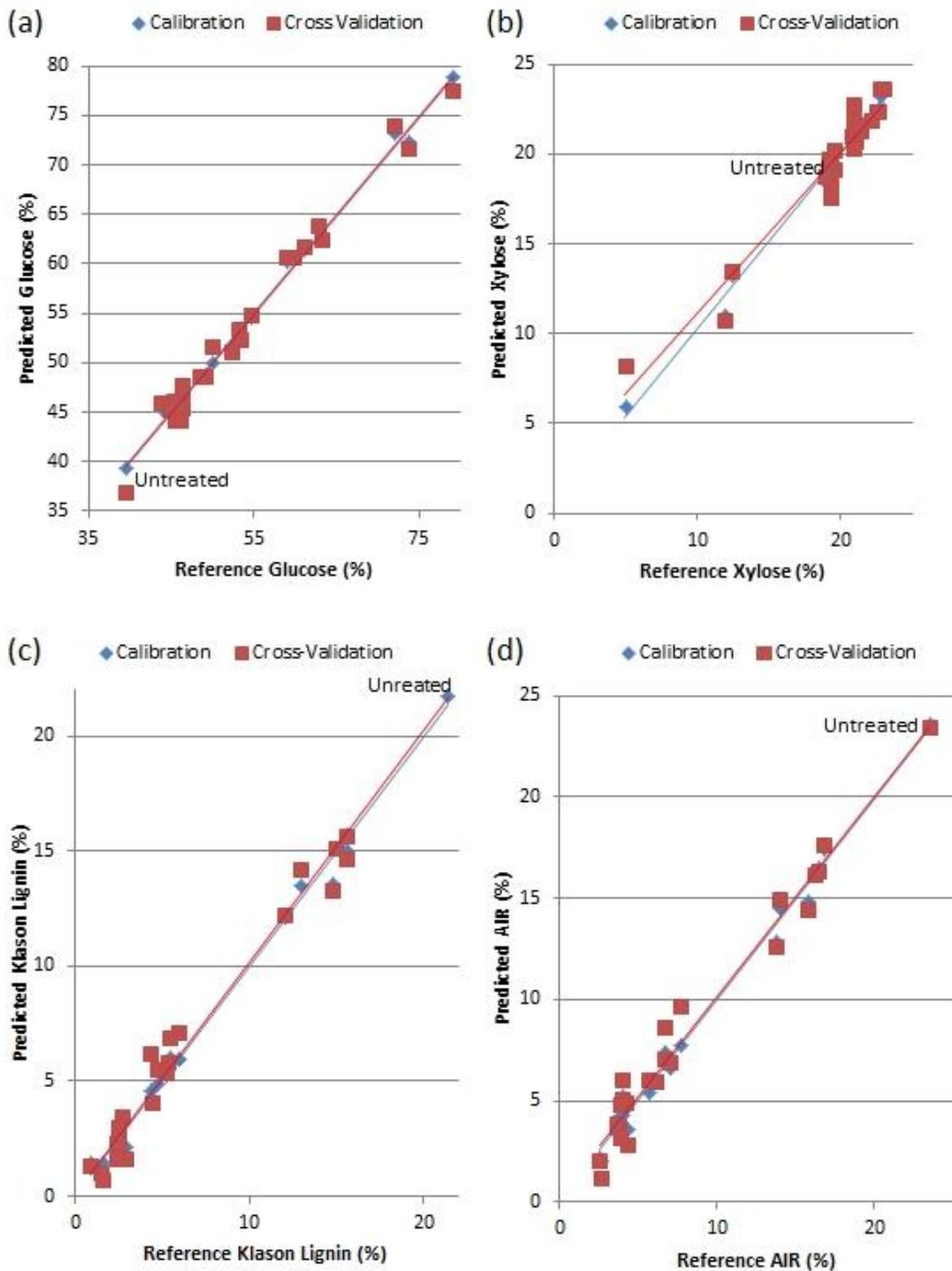


Figure I-13: Predicted y vs. reference y plots for the pretreated *Miscanthus* models. Predictions in calibration labelled by blue diamonds and predictions in full cross-validation labelled by red squares: (a) glucose content; (b) xylose content; (c) Klason lignin content; (d) acid insoluble residue (AIR) content.

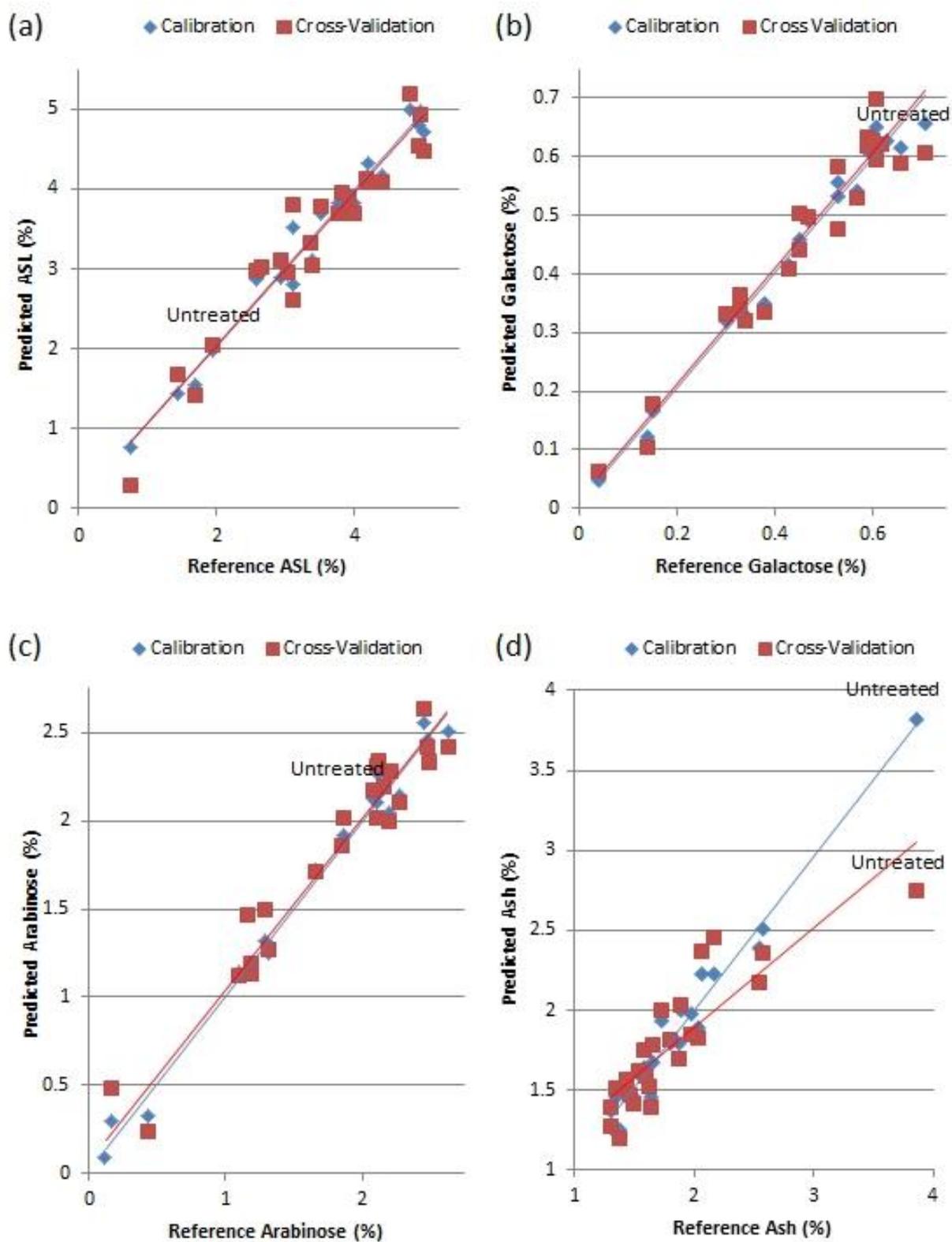


Figure I-14: Predicted y vs. reference y plots for the pretreated *Miscanthus* models. Predictions in calibration labelled by blue diamonds and predictions in full cross-validation labelled by red squares: (a) acid soluble lignin (ASL) content; (b) galactose content; (c) arabinose content; (d) ash content.

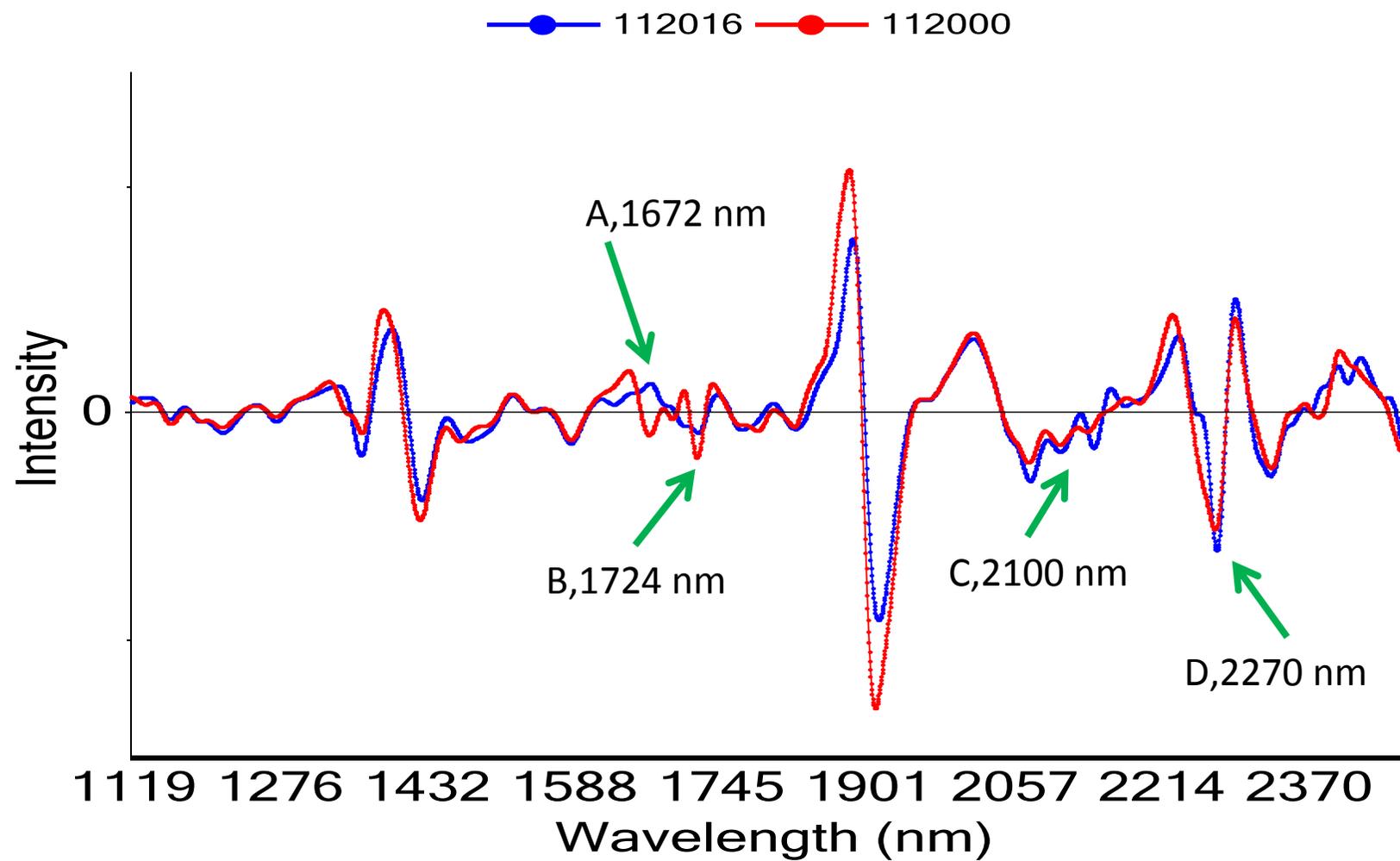


Figure I-15: The spectra, transformed by SG-2,2,25,25, of sample 112000 (no pretreatment) and sample 112016 (strong pretreatment conditions).

Appendix J References

- AACC 1999. Near-Infrared Methods: Guidelines for Model Development and Maintenance - AACC Method 39-00. *Approved Methods of the American Association of Cereal Chemists*. St. Paul, MN: AACC Press.
- ABNEY, W. & FESTING, E. R. 1881. *Phil. Trans. R. Soc.*, 172 887-918.
- ADAMOVIĆ, A., GRUBIĆ, G., MILENKOVIĆ, I., JOVANOVIĆ, R., PROTIC, R., SRETENOVIĆ, L. & STOICEVIĆ, L. 1998. The biodegradation of wheat straw by *Pleurotus ostreatus* mushrooms and its use in cattle feeding. *Animal Feed Science Technology*, 71, 357-362.
- ADEDIPE, O. E., DAWSON-ANDOH, B., SLAHOR, J. & OSBORN, L. 2008. Classification of red oak (*Quercus rubra*) and white oak (*Quercus alba*) wood using a near infrared spectrometer and soft independent modelling of class analogies. *Journal of Near Infrared Spectroscopy*, 16, 49-57.
- ADEN, A., BOZELL, J. J., HOLLADAY, J., WHITE, J. & MANHEIM, A. 2004. *Top Value Added Chemicals From Biomass Joint NREL/PNNL report, August 2004*, National Renewable Energy Laboratory/Pacific Northwest National Laboratory.
- ADLER, P. R., GROSSO, S. J. D. & PARTON, W. J. 2007. Life cycle assessment of net greenhouse-gas flux. *Ecological Applications*, 17, 675-691.
- AGBLEVOR, F. A., CHUM, H. L. & JOHNSON, D. K. 1993. Compositional analysis of NIST biomass standards from the IEA whole feedstock round robin. In: KLASS, D. L. (ed.) *Energy from Biomass and Wastes XVI*. Chicago: Institute of Gas Technology (IGT).
- AGBLEVOR, F. A., EVANS, R. J. & JOHNSON, K. D. 1994. Molecular-beam mass spectrometric analysis of lignocellulosic materials. I. Herbaceous biomass. *J. Anal. Appl. Pyrolysis*, 20, 124-144.
- AHMED, A. E. R. & LABAVITCH, J. M. 1977. A simplified method for accurate determination of cell wall uronide content. *J. Food Biochem.*, 1, 361-365.
- ALDER, E. 1977. Lignin chemistry - past, present and future. *Wood science and technology*, 11, 169-218.
- ALVES, A., SCHWANNINGER, M., PEREIRA, H. & RODRIGUES, J. 2006. Calibration of NIR to assess lignin composition (H/G ratio) in maritime pine wood using analytical pyrolysis as the reference method. *Holzforschung*, 60, 29-31.
- ALVES, E. F., BOSE, S. K., FRANCIS, R. C., COLODETTE, J. L., IAKOVLEV, M. & VAN HEININGEN, A. 2010. Carbohydrate composition of eucalyptus, bagasse and bamboo by a combination of methods. *Carbohydrate Polymers*, 82, 1097-1101.
- AMAN, P. 1993. Composition and structure of cell wall polysaccharides in forages. In: JUNG, H. G., BUXTON, D. R., HATFIELD, R. D. & RALPH, J. (eds.) *Forage cell wall structure and digestibility*. Madison, WI: ASA-CSSA-SSSA.
- AMAN, P. & LINDGREN, E. 1983. Chemical composition and in vitro degradability of individual chemical constituents of six Swedish grasses harvested at different stages of maturity. *Swed. J. Agric. Res.*, 13, 61-67.
- AMAN, P. & NORDKVIST, E. 1983. Chemical composition and in vitro degradability of botanical fractions of cereal straw. *Swed. J. Agric. Res.*, 13, 61-67.
- AN BORD GLAS 1996. *Mushroom Market Report*, An Bord Glas.
- ANGLÈS, M. N., REGUANT, J., MARTÍNEZ, J. M., FARRIOL, X., MONTANÉ, D. & SALVADÓ, J. 1997. Influence of the ash fraction on the mass balance during the summative analysis of high-ash content lignocellulosics. *Bioresource Technology*, 59, 185-193.
- ANTHONY, W. B. Cattle manure as a feed for cattle. Proc. Inter. Symp. Lvstlc. Wastes, 1972. ASAE, 314-318.

- ARNALDS, T., MCELHINNEY, J., FEARN, T. & DOWNEY, G. 2004. A hierarchical discriminant analysis for species identification in raw meat by visible and near infrared spectroscopy. *Journal of Near Infrared Spectroscopy*, 12, 183-188.
- ASPINALL, G. O. 1981. Constituents of plant cell walls. *Ency. Plant Physiol.*, 13, 3-8.
- ASTM 1993. D-1106-84. - Standard Test Method for Acid-Insoluble Lignin. *Annual Book of ASTM Standards*. Philadelphia, PA: American Society for Testing and Materials.
- ASTM STANDARD E131-10 2010. *ASTM E131 - 10 Standard Terminology Relating to Molecular Spectroscopy*, ASTM.
- AXRUP, L., MARKIDES, K. & NILSSON, T. 2000. Using miniature diode array NIR spectrometers for analysing wood chips and bark samples in motion. *Journal of Chemometrics*, 14, 561-572.
- BACIC, A., HARRIS, P. J. & STONE, B. A. 1988. Structure and function of plant cell walls. In: PRIESS, J. (ed.) *The biochemistry of plants*. New York, NY: Academic Press.
- BALABIN, R. M. & SAFIEVA, R. Z. 2011. Biodiesel classification by base stock type (vegetable oil) using near infrared spectroscopy data. *Analytica Chimica Acta*, 689, 190-197.
- BARRON, D. J., EVANS, E. W. & OETERSON, J. M. 1987. *International Journal of Coal Geology*, 8, 1.
- BASCH, A., WASSERMAN, T. & LEWIN, M. 1974. Near-Infrared spectrum of cellulose: A new method for obtaining crystallinity ratios. *Journal of Polymer Science*, 12, 1143-1150.
- BBC. 2007. EU biofuel policy is a 'mistake'.
- BBC 2008. "EU Promises Sustainable Plant Fuel".
- BEHAN, J. L. & SMITH, K. D. 2011. The analysis of glycosylation: a continued need for high pH anion exchange chromatography. *Biomedical Chromatography*, 25, 39-46.
- BEINING, B. A., HOLDEN, N. M., WARD, S. M. & FARRELL, E. P. The prediction of some peat properties for Irish industrial Bogs Using Near Infrared Spectroscopy. Proceedings of IPC Wetland 2000 Conference (August 2000), 2000 Quebec, Canada.
- BEN-GERA, I. & NORRIS, K. H. 1968. *J. Feed. Sci.*, 64
- BEVIN, C., STAUNTON, S. P., STOBE, R., KINGSTON, J. & LONERGAN, G. 2002. On-line use of near infrared spectroscopy in a sugar analysis system (SAS). *Proc. Aust. Soc. Sugar Cane Technol.*, 24, 1-10.
- BJØRSVIK, H.-R. & MARTENS, H. 2001. Data Analysis: Calibration of NIR Instruments by PLS Regression. In: BURNS, D. A. & CIURCZAK, E. W. (eds.) *Handbook of Near Infrared Analysis, Second Edition*. New York: Marcel Dekker.
- BLANCO, M., CASTILLO, M., PEINADO, A. & BENEYTO, R. 2007. Determination of low analyte concentrations by near-infrared spectroscopy: Effect of spectral pretreatments and estimation of multivariate detection limits. *Analytica Chimica Acta*, 581, 318-323.
- BLANCO, M., COELLO, J., ELAAMRANI, M., ITURRIAGA, H. & MASPOCH, S. 1996. Partial least-squares regression for the quantitation of pharmaceutical dosages in control analyses. *Journal of Pharmaceutical and Biomedical Analysis*, 15, 329-338.
- BLAND, E. & MENSUN, M. 1971. Determination of total lignin and polyphenol in eucalypt woods. *Appita* 25, 110-115.
- BNDES 2008. *Sugarcane-based Bioethanol: Energy for Sustainable Development*, Rio de Janeiro, Brazil, BNDES.
- BOUSSARSAR, H., ROGÉ, B. & MATHLOUTHI, M. 2009. Optimization of sugarcane bagasse conversion by hydrothermal treatment for the recovery of xylose. *Bioresource Technology*, 100, 6537-6542.
- BOYSWORTH, M. K. & BOOKSH, K. S. 2001. Aspects of multivariate calibration applied to near-infrared spectroscopy. In: BURNS, D. A. & CIURCZAK, E. W. (eds.) *Handbook of Near Infrared Analysis, Second Edition*. 2 ed. New York: Marcel Dekker.
- BOZELL, J. J. 2001. *Chemicals and Materials from Renewable Resources*, Washington DC, American Chemical Society.
- BRIODY, P. 1997. *Australian Sugar Yearbook*, Rural Press.

- BRODER, J. D. & BARRIER, J. W. 1988. *Paper 88-6007: Producing ethanol and coproducts from multiple feedstocks*, Am. Soc. Agric. Eng.
- BROWNING, B. L. 1967. *Methods of Wood Chemistry*, New York, Wiley Interscience.
- BULLARD, M. 2001. Economics of Miscanthus production. In: JONES, M. B. & WALSH, M. (eds.) *Miscanthus for energy and fibre*. London: James and James, Ltd.
- BULLARD, M. J. & NIXON, P. M. I. 1999. *Miscanthus Agronomy for Fuel and Industrial Uses*, MAFF Scientific Report no. NF0403, London, MAFF.
- BUNGAY, H. R. 1981. *Energy, the Biomass Options*, New York, NY, John Wiley and Sons.
- CACACE, J. E. & MAZZA, G. 2003. Mass transfer process during extraction of phenolic compounds from milled berries. *Journal of Food Engineering*, 59, 379-389.
- CAMO 2010. *The Unscrambler Appendices: Method References*, Oslo, Norway, CAMO Software AS.
- CAMO 2011. *Welcome to The Unscrambler X*, Oslo, Norway, CAMO Software AS.
- CANDOLFI, A., DE MAESSCHALCK, R., JOUAN-RIMBAUD, D., HAILEY, P. A. & MASSART, D. L. 1999. The influence of data pre-processing in the pattern recognition of excipients near-infrared spectra. *Journal of Pharmaceutical and Biomedical Analysis*, 21, 115-132.
- CARDONA, C. A., QUINTERO, J. A. & PAZ, I. C. 2010. Production of bioethanol from sugarcane bagasse: Status and perspectives. *Bioresource Technology*, 101, 4754-4766.
- CARPITA, N. C. & GIBEAUT, D. M. 1993. Structural models of primary cell walls in flowering plants: Consistency of molecular structure with the physical properties of the walls during growing. *Plant Journal*, 3, 1-30.
- CARVALHEIRO, F., DUARTE, L. C. & GIRIO, F. M. 2008. Hemicellulose biorefineries: a review on biomass pretreatments. *Journal of Scientific & Industrial Research*, 67, 16.
- CASEY, J. P. 1980. *Pulp and Paper: Chemistry and Chemical Technology*, New York, NY, John Wiley and Sons.
- CASIERI, C., BUBICI, S., VIOLA, I. & DE LUCA, F. 2004. A low-resolution non-invasive NMR characterization of ancient paper. *Solid State Nuclear Magnetic Resonance*, 26, 65-73.
- CATALDI, T. R. I., CAMPA, C. & BENEDETTO, G. E. 2000. Carbohydrate analysis by high-performance anion-exchange chromatography with pulsed amperometric detection: The potential is still growing. *Fresenius J Anal Chem*, 368, 739-758.
- CEN 1999. *Standardisation of solid biofuels in the Republic of Ireland*, Trinity College, Dublin, Department of Civil, Structural and Environmental Engineering.
- CHEN, D., HU, B., SHAO, X. & SU, Q. 2004. Variable selection by modified IPW (iterative predictor weighting)-PLS (partial least squares) in continuous wavelet regression models. *Analyst*, 129, 664-669.
- CHEN, F. & DIXON, R. A. 2007. Lignin modification improves fermentable sugar yields for biofuel production. *Nat Biotech*, 25, 759-761.
- CHEN, S.-F., MOWERY, R. A., SEVCIK, R. S., SCARLATA, C. J. & CHAMBLISS, C. K. 2010. Compositional Analysis of Water-Soluble Materials in Switchgrass. *Journal of Agricultural and Food Chemistry*, 58, 3251-3258.
- CHERNEY, J. H. & MARTEN, G. C. 1982. Small grain crop forage potential: II. Interrelationships among biological, chemical, morphological and anatomical determinants of quality. *Crop Sci.*, 22, 240-243.
- CHRISTIAN, D. G. & HAASE, E. 2001. Agronomy of Miscanthus. In: JONES, M. B. & WALSH, M. (eds.) *Miscanthus for energy and fibre*. London: James and James, Ltd.
- CHRISTIAN, D. G. & RICHE, A. B. 2000. *Establishing Fuel Specifications of Non-Wood Biomass Crops*, ETSU B/U1/00612/00/00.
- CHUM, H. L., MILNE, T. A., JOHNSON, D. K. & AGBLEVOR, F. A. Feedstock characterisation and recommended procedures. Proc. First Biomass Conf. Of the Americas: Energy Environment, Agriculture and Industry, 1993 Burlington, VT. National Renewable Energy Laboratory, Golden, Colorado, 1685-16703.

- CIURCZAK, E. W. 2001. Principles of Near-Infrared Spectroscopy. In: BURNS, D. A. & CIURCZAK, E. W. (eds.) *Handbook of Near-Infrared Analysis - Second Edition*. 2 ed. New York, NY: Marcel Dekker.
- CLARK, D. H. & LAMB, R. C. 1986. *J. Dairy Sci.*, 69, 136
- CLEAN WASHINGTON CENTER 1997. *Wheat Straw as a Paper Fiber Source*, Seattle, Washington, The Clean Washington Center - A Division of the Pacific Northwest Economic Region (PNWER).
- CLEVELAND, C. J. & KAUFMANN, R. K. 2003. Oil supply and oil politics: déjà vu all over again. *Energy Policy*, 31, 485–489.
- CLEVELAND, W. S. & DEVLIN, S. J. 1988. Locally Weighted Regression: An Approach to Regression Analysis by Local Fitting. *Journal of the American Statistical Association*, 83, 596-610.
- CLIFTON-BROWN, J. C. 1997. *The Importance of Temperature in Controlling Leaf Growth of Miscanthus in Temperate Climates*. PhD, University of Dublin.
- CLIFTON-BROWN, J. C. & LEWANDOWSKI, I. 2002. Screening Miscanthus genotypes in field trials to optimise biomass yield and quality in Southern Germany. *European Journal of Agronomy*, 16, 97-110.
- CLIFTON-BROWN, J. C., LEWANDOWSKI, I., ANDERSSON, B., BASCH, G., CHRISTIAN, D. G., KJELDSSEN, J. B., JORGENSEN, U., MORTENSON, J. V., RICHE, A. B., SCHWARZ, K.-U., TAYEBI, K. & TEIXEIRA, F. 2001a. Performance of 15 *Miscanthus* genotypes at five sites in Europe. *Agron. J.*, 93, 1013-1019.
- CLIFTON-BROWN, J. C., LEWANDOWSKI, I. & JONES, M. B. 2002. MiscanMod: a model for estimating biomass. In: PUDE, R. (ed.) *Anbau und Verwertung von Miscanthus in Europa, Beiträge zu den Agrarwissenschaften*. Wittenschlick/Bonn, Germany: Verlag M. Wehle.
- CLIFTON-BROWN, J. C., LONG, S. P. & JORGENSEN, U. 2001b. Miscanthus productivity. In: JONES, M. B. & WALSH, M. (eds.) *Miscanthus for energy and fibre*. London: James and James.
- COBLENTZ, W. W. 1905. *Investigations of Infrared Spectra Part 1. Publication No. 35*, Carnegie Institute of Washington.
- COLLINS, C., BOLLOCH, O. L. & MEANEY, B. 2005. National Waste Report 2004. Dublin: Environmental Protection Agency.
- COZZOLINO, D., FASSIO, A., FERNANDEZ, E., RESTAINO, E. & LA MANNA, A. 2006a. Measurement of chemical composition in wet whole maize silage by visible and near infrared reflectance spectroscopy. *Animal Feed Science and Technology*, 129, 329-336.
- COZZOLINO, D., VADELL, A., BALLESTEROS, F., GALIETTA, G. & BARLOCCO, N. 2006b. Combining visible and near-infrared spectroscopy with chemometrics to trace muscles from an autochthonous breed of pig produced in Uruguay: a feasibility study. *Analytical and Bioanalytical Chemistry*, 385, 931-936.
- CROSS, J. R. 1983. *Peatlands, Wastelands or Heritage? An Introduction to Bogs*, Dublin, An Foras Taluntais.
- CROWE, M. F. 2000. *National waste database report 1998*, Wexford, Environmental Protection Agency.
- CRUTZEN, P. J., MOSIER, A. N., SMITH, K. A. & WINIWARTER, W. 2007. N₂O release from agro-biofuel production negates global warming reduction by replacing fossil fuels. *Atmos. Chem. Phys. Discuss.*, 7, 11191-11205.
- CSO 2002. *December Livestock Surveys*, Dublin, CSO Eirestat database.
- CSO 2007. *Area, Yield and Production of Crops 2006*.
- CSO. 2009. *Selected Livestock Numbers in December* [Online]. Dublin: Central Statistics Office Ireland. Available: http://www.cso.ie/quicktables/GetQuickTables.aspx?FileName=AAA02.asp&TableName=Selected+Livestock+Numbers+in+December&StatisticalProduct=DB_AA.
- CTC 2005. *Síntese do controle mútuo agroindustrial*, Piracicaba, Sao Paulo, Brazil, Centro de Tecnologia Canavieira.

- CUNHA, J. A., PEREIRA, M. M., VALENTE, L. M. M., DE LA PISCINA, P. R., HOMES, N. & SANTOS, M. R. L. 2011. Waste biomass to liquids: Low temperature conversion of sugarcane bagasse to bio-oil. The effect of combined hydrolysis treatments. *Biomass and Bioenergy*, 35, 2106-2116.
- DARDENNE, P., SINNAEVE, G. & BAETEN, V. 2000. Multivariate calibration and chemometrics for near infrared spectroscopy: which method? *Journal of Near Infrared Spectroscopy*, 8, 229-237.
- DAVIS, B. G. & FAIRBANKS, A. J. 2002. *Carbohydrate Chemistry*, Oxford, UK, Oxford University Press.
- DAVIS, M. K. & CHEN, G. 2007. Graphing Kendall's [tau]. *Computational Statistics & Data Analysis*, 51, 2375-2378.
- DAVIS, M. W. 1998. A rapid modified method for compositional carbohydrate analysis of lignocellulosics by high pH anion-exchange chromatography with pulsed amperometric detection (HPAEC/PAD). *Journal of Wood Chemistry and Technology*, 18, 235-252.
- DE MAESSCHALCK, R., JOUAN-RIMBAUD, D. & MASSART, D. L. 2000. The Mahalanobis distance. *Chemometrics and Intelligent Laboratory Systems*, 50, 1-18.
- DE VRIES, S. & J.F. TER BRAAK, C. 1995. Prediction error in partial least squares regression: a critique on the deviation used in The Unscrambler. *Chemometrics and Intelligent Laboratory Systems*, 30, 239-245.
- DE VRIJE, T., DE HAAS, G. G., TAN, G. B., KEIJSERS, E. R. P. & CLAASSEN, P. A. M. Pretreatment of Miscanthus for hydrogen production by Thermotoga elfii. *International Journal of Hydrogen Energy*, 27, 1381-1390.
- DEHORITY, B. A. 1993. Microbial Ecology of Cell Wall Fermentation. In: JUNG, H. G., BUXTON, D. R., HATFIELD, R. D. & RALPH, J. (eds.) *Forage cell wall structure and digestibility*. Madison, WI: ASA-CSSA-SSSA.
- DELUZIO, K. J., WYSS, U. P., ZEE, B., COSTIGAN, P. A. & SERBIE, C. 1997. Principal component models of knee kinematics and kinetics: Normal vs. pathological gait patterns. *Human Movement Science*, 16, 201-217.
- DENNEY, R. C. & SINCLAIR, R. 1993. *Visible and Ultraviolet Spectroscopy - Analytical Chemistry by Open Learning*, Chichester, England, John Wiley and Sons.
- DHANOVA, M. S., LISTER, S. J., SANDERSON, R. & BARNES, R. J. 1994. The link between multiplicative scatter correction (MSC) and standard normal variate (SNV) transformations of NIR spectra. *Journal of Near Infrared Spectroscopy*, 2, 43-47.
- DINUS, R. J. 2000. *Genetic Modification of Short Rotation Poplar Biomass Feedstock for Efficient Conversion to Ethanol*, Bioenergy Feedstock Development Program, Environmental Sciences Division, Oak Ridge National Laboratory ORNL/Sub/99-4500007253/1.
- DIONEX 1999. *ASE 200 Accelerated Solvent Extractor Operator's Manual*, Sunnyvale, CA, Dionex Corporation.
- DIONEX 2000. *Technical Note 20: Analysis of Carbohydrates by High Performance Anion Exchange Chromatography with Pulsed Amperometric Detection (HPAE-PAD)*, Sunnyvale, CA, Dionex Corporation.
- DIONEX 2003. *Dionex Application Note 92: The Determination of Sugars in Molasses by High-Performance Anion Exchange with Pulsed Amperometric Detection*, Sunnyvale, CA, Dionex Corporation.
- DIONEX 2004a. *AN 325: Extraction of Oils from Oilseeds by Accelerated Solvent Extraction (ASE)*, Sunnyvale, CA, Dionex Corporation.
- DIONEX 2004b. *Product Manual: IonPac NG1 Guard Column*, Sunnyvale, CA, Dionex Corporation.
- DIONEX 2005a. *Product Manual: CarboPac Combined Products*, Sunnyvale, CA, Dionex Corporation.
- DIONEX 2005b. *Product Manual: CarboPac PA20*, Sunnyvale, CA, Dionex Corporation.
- DIONEX 2006. *ICS-3000 Ion Chromatography System Operator's Manual*, Sunnyvale, CA, Dionex Corporation.
- DIONEX 2009. *Application Brief 105: Anions and Organic Acids in Wood Extracts*, Sunnyvale, CA, Dionex Corporation.

- DIONEX 2010. *Chromeleon Client V.6.80 SR10 User Guide* Sunnyvale, California, Dionex Corporation.
- DIOUF, J. 2007. Biofuels should benefit the poor, not the rich. *Financial Times*, 15-08-07.
- DIPARDO, J. 2000. *Outlook for biomass ethanol production and demand* [Online]. Energy Information Administration. Available: <http://www.eia.doe.gov/oiaf/analysispaper/pdf/biomass.pdf> [Accessed 12/7/11 2011].
- DOCO, T., WILLIAMS, P., PAULY, M., O'NEILL, M. A. & PELLERIN, P. 2003. Polysaccharides from grape berry cell walls. Part II. Structural characterization of the xyloglucan polysaccharides. *Carbohydrate Polymers*, 53, 253-261.
- DOWNES, G. M., MEDER, R., EBDON, N., BOND, H., EVANS, R., JOYCE, K. & SOUTHERTON, S. 2010a. Radial variation in cellulose content and Kraft pulp yield in *Eucalyptus nitens* using near-infrared spectral analysis of air-dry wood surfaces. *Journal of Near Infrared Spectroscopy*, 18, 147-155.
- DOWNES, G. M., MEDER, R. & HARWOOD, C. 2010b. A multi-site, multi-species near infrared calibration for the prediction of cellulose content in eucalypt woodmeal. *Journal of Near Infrared Spectroscopy*, 18, 381-387.
- DOYLE, R. 1997. *The Hydrodynamics and Rheology of Peat-Solvent Suspensions*. PhD Thesis, University of Limerick.
- DU TOIT, P., OLIVIER, S. & VAN BILJON, P. 1984. Sugar cane bagasse as a possible source of fermentable carbohydrates. I. Characterisation of bagasse with regard to monosaccharide, hemicellulose and amino acid composition. *Biotechnology and Bioengineering*, 26, 1071-1078.
- DU, Y. P., LIANG, Y. Z., JIANG, J. H., BERRY, R. J. & OZAKI, Y. 2004. Spectral regions selection to improve prediction ability of PLS models by changeable size moving window partial least squares and searching combination moving window partial least squares. *Analytica Chimica Acta*, 501, 183-191.
- DUGUID, K. B., MONTROSS, M. D., RADTKE, C. W., CROFCHECK, C. L., SHEARER, S. A. & HOSKINSON, R. L. 2007. Screening for sugar and ethanol processing characteristics from anatomical fractions of wheat stover. *Biomass and Bioenergy*, 31, 585-592.
- DYNAMOTIVE ENERGY SYSTEMS CORPORATION. Fast pyrolysis of Bagasse to Produce BioOil Fuel for Power Generation. Sugar Conference, 2001, 2001 Trinidad and Tobago, April 2001.
- EC 1999. *Council Directive 1999/31/EC of 26 April 1999 on the landfill of waste*. *Official Journal L 182*, 16/07/1999 European Commission.
- EC 2003. *Directive 2003/30/EC of the European Parliament and of the Council of 8 May 2003 on the promotion of the use of biofuels or other renewable fuels for transport*. *Official Journal of the European Union*, 123, 42-46.
- EC. 2007. *Promoting Biofuels as Credible Alternatives to Oil in Transport*, MEMO/07/5 [Online]. EUROPA. Available: <http://europa.eu/rapid/pressReleasesAction.do?reference=MEMO/07/5&format=HTML&aged=0&language=EN&guiLanguage=en> [Accessed 12/7/11 2011].
- EDWARDS, B. 1991. Energy use in cane and beet factories. *Proceedings of Australian Society of Sugar Cane Technologists*, 13, 227-229.
- EHRMAN, T. 1994a. *LAP-005: Standard method for ash in biomass*, Golden, Colorado, NREL.
- EHRMAN, T. 1994b. *LAP-010: Standard method for the determination of extractives in biomass*, Golden, Colorado, NREL.
- EHRMAN, T. 1994c. *LAP-012: Standard test method for moisture, total solids, and total dissolved solids in biomass slurry and liquid process samples*, Golden, Colorado, NREL.
- EL HAGE, R., CHRUSCIEL, L., DESHARNAIS, L. & BROSE, N. 2010. Effect of autohydrolysis of *Miscanthus x giganteus* on lignin structure and organosolv delignification. *Bioresource Technology*, 101, 9321-9329.
- EPA 1995. *AP 42 - Compilation of Air Pollutant Emission Factors, Volume 1: Stationary Point and Area Sources*, Research Triangle Park, North Carolina, US Environmental Protection Agency.

- EPA 1996 *Waste Characterisation Methodology*, Wexford, Ireland, Environmental Protection Agency.
- EPA 2006. *National Waste Report 2005: Data Update*, Ireland, Environmental Protection Agency.
- ERNST, A. J., FOURAL, Y. & CLARK, T. F. 1960. Rice Straw for Bleached Paper. *Tappi Journal*, 43, 49-53.
- EWANICK, S. & BURA, R. 2011. The effect of biomass moisture content on bioethanol yields from steam pretreated switchgrass and sugarcane bagasse. *Bioresource Technology*, 102, 2651-2658.
- FABER, V. 1994. Clustering and the continuous k-means algorithm. *Los Alamos Science*, 22, 138-144.
- FACKLER, K., SCHWANNINGER, M., GRADINGER, C., HINTERSTOISSER, B. & MESSNER, K. 2007. Qualitative and quantitative changes of beech wood degraded by woodrotting basidiomycetes monitored by Fourier transform infrared spectroscopic methods and multivariate data analysis. *FEMS Microbiol. Lett.*, 271, 162-169.
- FAGAN, C. C., EVERARD, C. D. & MCDONNELL, K. 2011. Prediction of moisture, calorific value, ash and carbon content of two dedicated bioenergy crops using near-infrared spectroscopy. *Bioresource Technology*, 102, 5200-5206.
- FAIR 2000. *Switchgrass (Panicum virgatum L.) as an alternative energy crop in Europe - Initiation of a productivity network - FAIR 5-CT97-3701*, Wageningen, The Netherlands, FAIR.
- FAIX, O., MEIER, D. & BEINHOFF, O. 1989. Analysis of lignocelluloses and lignins from *Arundo donax* L. and *Miscanthus sinensis* Anderss., and hydroliquefaction of *Miscanthus*. *Biomass*, 18, 109-126.
- FAN, L. T., GHARPURAY, M. M. & LEE, Y.-H. 1987. *Cellulose hydrolysis*, Berlin, Springer-Verlag.
- FARONE, W. A. & CUZENS, J. E. 1996. *Method of Producing Sugars Using Strong Acid Hydrolysis of Cellulosic and Hemicellulosic Materials*. US patent application.
- FEARN, T. 2002. Assessing calibrations: SEP, RPD, RER and R². *NIR News*, 13, 12-14.
- FEC CONSULTANTS 1990. *Forestry waste firing of industrial boilers - ETSU-B--1178*, London, ETSU.
- FENGEL, D. & WEGENER, G. 1984. *Wood, Chemistry Ultrastructure and Reactions*, Berlin, De Gruyter.
- FERRETT, G. 2007. Biofuels 'crime against humanity'. 27-10-07. Available: <http://news.bbc.co.uk/1/hi/world/americas/7065061.stm>.
- FITZPATRICK, S. W. 1997. *Production of levulinic acid from carbohydrate-containing materials: U.S. Patent 5,608,105*.
- FORINA, M., CASOLINO, C. & PIZARRO MILLAN, C. 1999. Iterative predictor weighting (IPW) PLS: a technique for the elimination of useless predictors in regression problems. *Journal of Chemometrics*, 13, 165-184.
- FOSS 2004. *Vision Diagnostics*, Hilleroed, Denmark, FOSS Analytical.
- FOSS 2006a. *Vision Software Reference*, Hillerød, Denmark, FOSS.
- FOSS 2006b. *XDS Rapid Content Analyzer Service Manual*, Hilleroed, Denmark, Foss Analytical.
- FOSS. 2011. *Infratec 1241 Grain Analyser* [Online]. Hilleroed, Denmark: FOSS Analytical. Available: <http://www.foss.dk/~media/Files/Documents/Industry%20solution%20documents/Brochures%20and%20data%20sheet/Infratec1241/Infratec1241datasheetGBnew.ashx> [Accessed 13/4/11 2011].
- FRITZ, J. S. & GJERDE, D. T. 1995. *Ion Chromatography*, New York, NY, Wiley.
- FUCHSMAN, C. H. 1980. *Peat : Industrial Chemistry and Technology*, New York, Academic Press.
- GAILLOT, O., GIBSON, C. & HOGAN, J. 2005. Programme for Municipal Waste Characterisation Surveys. Dublin: RPS MCOS.
- GALVÃO, R. K. H., ARAUJO, M. C. U., JOSÉ, G. E., PONTES, M. J. C., SILVA, E. C. & SALDANHA, T. C. B. 2005. A method for calibration and validation subset partitioning. *Talanta*, 67, 736-740.
- GÁMEZ, S., GONZÁLEZ-CABRIALES, J. J., RAMÍREZ, J. A., GARROTE, G. & VÁZQUEZ, M. 2006. Study of the hydrolysis of sugar cane bagasse using phosphoric acid. *Journal of Food Engineering*, 74, 78-88.
- GAUTHIER, T. D. 2001. Detecting Trends Using Spearman's Rank Correlation Coefficient. *Environmental Forensics*, 2, 359-362.

- GELADI, P. & KOWALSKI, B. R. 1986. Partial least squares regression: A tutorial. *Analytica Chimica Acta*, 185, 1-17.
- GÍRIO, F. M., FONSECA, C., CARVALHEIRO, F., DUARTE, L. C., MARQUES, S. & BOGEL-LUKASIK, R. 2010. Hemicelluloses for fuel ethanol: A review. *Bioresource Technology*, 101, 4775-4800.
- GIRISUTA, B. 2007. *Levulinic acid from lignocellulosic biomass*. University of Groningen.
- GOERING, H. K. & VAN SOEST, P. J. 1970. *Forage fibre analyses*. USDA Agric. Handb. 379, Washington, DC, US Govt. Print. Office.
- GOLDSTEIN, I. S. 1991. Overview of the chemical composition of wood. In: LEWIN, M. & GOLDSTEIN, I. S. (eds.) *Wood Structure and Composition*. International Fiber Science and Technology.
- GRAF, A. & KOEHLER, T. 2000. *Oregon Cellulose-Ethanol Study*, Salem, Oregon, USA, Oregon Office of Energy.
- GROGAN, P. & MATTHEW, R. 2001. Review of the potential for soil carbon sequestration under bioenergy crops in the U.K. *MAFF report on contract NF0418*. Institute of Water and Environment, Cranfield University, Silsoe.
- GUO, X.-J., WANG, S.-R., WANG, K.-G., LIU, Q. & LUO, Z.-Y. 2010. Influence of extractives on mechanism of biomass pyrolysis. *Journal of Fuel Chemistry and Technology*, 38, 42-46.
- HA, Y. W. & THOMAS, R. L. 1988. Simultaneous Determination of Neutral Sugars and Uronic Acids in Hydrocolloids. *Journal of Food Science*, 53, 574-577.
- HAALAND, D. M. & THOMAS, E. V. 1988. Partial least-squares methods for spectral analyses. 1. Relation to other quantitative calibration methods and the extraction of qualitative information. *Analytical Chemistry*, 60, 1193-1202.
- HADDAD, P. R. & JACKSON, P. E. 2003. *Ion Chromatography: Principles and Applications*, Amsterdam, The Netherlands, Elsevier.
- HAKKILA, P. 1989. *Utilisation of Residual Forest Biomass*, Berlin, Springer-Verlag.
- HAMELINCK, C., VAN DEN BROEK, R., RICE, B., GILBERT, A., RAGWITZ, M. & TORO, F. 2004. *Liquid Biofuels Strategy Study for Ireland*, Sustainable Energy Ireland.
- HAMELINCK, C. N., VAN HOOIJDONK, G. & FAAIJ, A. 2005. Ethanol from lignocellulosic biomass: techno-economic performance in short-, middle- and long-term. *Biomass and Bioenergy*, 28, 384-410.
- HAMES, B., THOMAS, S. R., SLUITER, A., ROTH, C. J. & TEMPLETON, D. 2003. Rapid biomass analysis - New tools for compositional analysis of corn stover feedstocks and process intermediates from ethanol production. *Applied Biochemistry and Biotechnology*, 105-108, 5-16.
- HAMMOND, R. F. 1979. *The Peatlands of Ireland, Soil Survey Bulletin No. 35*, Dublin, An Foras Taluntais.
- HANSEN, E. M., CHRISTENSEN, B. T., JENSEN, L. S. & KRISTENSEN, K. 2004. Carbon sequestration in soil beneath long-term Miscanthus plantations as determined by ¹³C abundance. *Biomass and Bioenergy*, 26, 97-105.
- HARRIS, P. J. 1990. Plant cell structure and development. In: AKIN, D. E. (ed.) *Microbial and plant opportunities to improve lignocellulose utilisation by ruminants*. New York, N. Y.: Elsevier Sci. Publ. Co.
- HATFIELD, R. D. 1993. Cell Wall Polysaccharide Interactions and Degradability. In: JUNG, H. G., BUXTON, D. R., HATFIELD, R. D. & RALPH, J. (eds.) *Forage cell wall structure and digestibility*. Madison, WI: ASA-CSSA-SSSA.
- HAYASHI, T. 1989. Xyloglucans in the primary cell wall. *Annu. Rev. Plant Physiol.*, 40, 139-168.
- HAYES, D. J. 2008. An Examination of Biorefining Processes, Catalysts and Challenges. *Catalysis Today*, 145, 138-151.
- HAYES, D. J., FITZPATRICK, S. W., HAYES, M. H. B. & ROSS, J. R. H. 2005. The Biofine Process: Production of levulinic acid, furfural and formic acid from lignocellulosic feedstocks. In: KAMM, B., GRUBER, P. R. & KAMM, M. (eds.) *Biorefineries: Industrial Processes and Products*. Weinheim, Germany: Wiley.

- HAYES, D. J. & HAYES, M. H. B. 2009. The role that lignocellulosic feedstocks and various biorefining technologies can play in meeting Ireland's biofuel targets. *Biofpr*, 3, 500-520.
- HAYES, M. H. B. 2006. Biochar and biofuels for a brighter future (letter). *Nature (London)*, 442, 144.
- HEDING, J. B., KOFMAN, P. D. & MORSING, M. 1993. *Lagring af braendiels flis, chunk og braende (Storage of wood fuel chips, chunk and fire wood)*, Lyngby, Denmark, Skovbrugsserien nr 7. Forskningscentret for Skov & Landskab.
- HENNENBERG, W. & STOHMANN, F. 1859. Uber das Erhaltungsfutter volljahrigen Rindviehs. *J. Landwirtsch*, 3, 485-551.
- HENNING, A. & POPPE, S. 1977. *Animal wastes as a feed*, Prague, St. zem. nakl.
- HERSCHEL, W. 1800. *Philos. Trans. R. Soc.*, 90, 255-283.
- HILL, J., NELSON, E., TILMAN, D., POLASKY, S. & TIFFANY, D. 2006a. Environmental, economic and energetic costs and benefits of biodiesel and ethanol biofuels. *Proceedings of the National Academy of Sciences*, 103, 11206-11210.
- HILL, J., NELSON, E., TILMAN, D., POLASKY, S. & TIFFANY, D. 2006b. Environmental, economic, and energetic costs and benefits of biodiesel and ethanol biofuels. *PNAS*, 103, 11206-11210.
- HILLARD, E. P. 1977. *Chemical and nutritive characterization of pig diets and pig faeces - Progress Report, Jul. 9-12. 1977*, Univ. of Melbourne, Australia.
- HINSHAW, J. V. 2006. The thermal conductivity detector. *LCGC North America*, 24.
- HODGE, G. R. & WOODBRIDGE, W. C. 2010. Global Near Infrared models to predict the lignin and cellulose content of pine wood. *Journal of Near Infrared Spectroscopy*, 18, 367-380.
- HODGSON, E. M., J. LISTER, S. J., BRIDGWATER, A. V., CLIFTON-BROWN, J. C. & DONNISON, I. S. 2010. Genotypic and environmentally derived variation in the cell wall composition of *Miscanthus* in relation to its use as a biomass feedstock. *Biomass and Bioenergy*, 34, 652-660.
- HODGSON, E. M., NOWAKOWSKI, D. J., SHIELD, I., RICHE, A., BRIDGWATER, A. V., CLIFTON-BROWN, J. C. & DONNISON, I. S. 2011. Variation in *Miscanthus* chemical composition and implications for conversion by pyrolysis and thermo-chemical bio-refining for fuels and chemicals. *Bioresource Technology*, 102, 3411-3418.
- HOEBLER, C., BARRY, J. L., DAVID, A. & DELORT-LAVAL, J. 1989. Rapid acid hydrolysis of plant cell wall polysaccharides and simplified quantitative determination of their neutral monosaccharides by gas-liquid chromatography. *J. Agr. Food Chem*, 37, 360-367.
- HOPKINS, D. W. 2001. What is a Norris derivative? *NIR News*, 12, 3-5.
- HOUSE OF COMMONS 2008. Are Biofuels Sustainable? First Report of Session 2007-2008.
- HUANG, C., HAN, L., YANG, Z. & LIU, X. 2009. Ultimate analysis and heating value prediction of straw by near infrared spectroscopy. *Waste Management*, 29, 1793-1797.
- HUANG, G., HAN, L. & LIU, X. 2007. Rapid estimation of the composition of animal manure compost by near infrared reflectance spectroscopy. *Journal of Near Infrared Spectroscopy*, 15, 387-394.
- HUDSON, J. B., MITCHELL, C. P., DARDENER, D. & STORRY, P. A comparative study of storage and drying of chips and chunks in the UK. Proc. IEA/BA Task III conference production, storage and utilisation of wood fuels, Vol. II, Uppsala, Dec 6-7, 1988. Swedish University of Agricultural Sciences, Department of Operational Efficiency, Garpenberg, Sweden, 72-89.
- HUSSAIN, I., CHEEKE, P. R. & JOHNSON, D. E. 1996. Evaluation of grass straw:corn juice silage as a ruminant feedstuff: digestibility, straw ammoniation and supplementation with by-pass protein. *Anim. Feed Sci. Technol.*, 57, 1-13.
- IÑÓN, F. A., GARRIGUES, J. M., GARRIGUES, S., MOLINA, A. & DE LA GUARDIA, M. 2003. Selection of calibration set samples in determination of olive oil acidity by partial least squares-attenuated total reflectance-Fourier transform infrared spectroscopy. *Analytica Chimica Acta*, 489, 59-75.
- IPCC 2001. *Climate Change 2001: The Scientific Basis*, Cambridge, UK, Cambridge University Press.

- JACKSON DE MORAES ROCHA, G., MARTIN, C., SOARES, I. B., SOUTO MAIOR, A. M., BAUDEL, H. M. & MORAES DE ABREU, C. A. 2011. Dilute mixed-acid pretreatment of sugarcane bagasse for ethanol production. *Biomass and Bioenergy*, 35, 663-670.
- JANSEN, B., NIEROP, K. G. J., KOTTE, M. C., DE VOOGT, P. & VERSTRATEN, J. M. 2006. The applicability of accelerated solvent extraction (ASE) to extract lipid biomarkers from soils *Applied Geochemistry*, 21, 1006-1015.
- JIANG, Z.-H., YANG, Z., SO, C.-H. & HSE, C.-H. 2007. Rapid prediction of wood crystallinity in Pinus elliotii plantation wood by near-infrared spectroscopy. *J. Wood Sci.*, 53, 449-453.
- JIN, S. & CHEN, H. 2007. Near-infrared analysis of the chemical composition of rice straw. *Industrial Crops & Products*, 26, 207-211.
- JIRJIS, R. 1995. Storage and drying of wood fuel. *Biomass and Bioenergy*, 9, 181-190.
- JOHANSON, J., FOLESTAD, S., JOSEFSON, M., SPAREN, A., ABRAHAMSSON, C., ANSERSSON-ENGELS, S. & SVANBERG, S. 2002. *Appl. Spectrosc.*, 56, 725.
- JONES, J. N., SCHOFIELD, W., MCKENZIE, N. J. & OLSON, B. C. 2002. Factory rate control based on more constant fibre rate. *Proc. Aust. Soc. Sugar Cane Technol.*, 24, 1-7.
- JONES, P. D., SCHIMLECK, L. R., PETER, G. F., DANIELS, R. F. & CLARK III, A. 2006. Nondestructive estimation of wood chemical composition of sections of radial wood strips by diffuse reflectance near infrared spectroscopy. *Wood Science and Technology*, 40, 709-720.
- JORDAN, S. N., MULLEN, G. J. & MURPHY, C. M. Composition variability of spent mushroom compost in Ireland. 14th Irish Environmental Researchers' Colloquium, 2004 University of Limerick.
- JORDAN, S. N., MULLEN, G. J. & MURPHY, M. C. 2008. Composition variability of spent mushroom compost in Ireland. *Biores. Technol.*, 99, 411-418.
- JOUAN-RIMBAUD, D., BOUVERESSE, E., MASSART, D. L. & DE NOORD, O. E. 1999. Detection of prediction outliers and inliers in multivariate calibration. *Analytica Chimica Acta*, 388, 283-301.
- KAACK, K. & SCHWARZ, K.-U. 2001. Morphological and mechanical properties of Miscanthus in relation to harvesting, lodging, and growth conditions. *Industrial Crops and Products*, 14, 145-154.
- KAAR, W. E., COOL, L. G., MERRIMAN, M. M. & BRINK, D. L. 1991. The Complete Analysis of Wood Polysaccharides Using HPLC. *Journal of Wood Chemistry and Technology*, 11, 447 - 463.
- KABEL, M. A., BOS, G., ZEEVALKING, J., VORAGEN, A. G. J. & SCHOLS, H. A. 2007. Effect of pretreatment severity on xylan solubility and enzymatic breakdown of the remaining cellulose from wheat straw. *Bioresource Technology*, 98, 2034-2042.
- KALTSCHMITT, M. & HARTMANN, H. 2000. *Energie aus Biomasse: Grundlagen, Techniken und Verfahren*, Berlin, Germany, Springer-Verlag.
- KATOFISKY, R. 1993. *The production of liquid fuels from biomass*, Princeton, NJ, Princeton University - Center for Energy and Environmental Studies.
- KAVALOV, B. & PETEVES, S. D. 2005. Status and Perspectives of Biomass-to-Liquid Fuels in the European Union. Petten, The Netherlands: European Commission Director General Joint Research Centre.
- KAYE, W. 1975. Resolution and stray light in near infrared spectroscopy. *Appl. Opt.*, 14, 1977-1986.
- KELLEY, S. S., RIALS, T. G., SNELL, R., GROOM, L. H. & SLUITER, A. 2004a. Use of near infrared spectroscopy to measure the chemical and mechanical properties of solid wood. *Wood Sci. Technol.*, 38, 257-276.
- KELLEY, S. S., ROWELL, R. M., DAVIS, M., JURICH, C. K. & IBACH, R. 2004b. Rapid analysis of the chemical composition of agricultural fibers using near infrared spectroscopy and pyrolysis molecular beam mass spectrometry. *Biomass and Bioenergy*, 27, 77-88.
- KERLEY, M. S., FAHEY, G. C., GOULD, J. M. & IANNOTTI, E. L. 1988. Effect of lignification, cellulose crystallinity and enzyme accesible space on the digestibility of plant cell wall carbohydrates by the ruminant. *Food Microstruct.*, 7, 29-65.

- KLASON, P. 1922. Contributions to a more exact knowledge of the chemical composition of spruce wood, part I *Pap. Trade J.*, 74, 45- 51.
- KLASS, D. L. 1981. *Biomass as a nonfossil fuel source*, American Chemical Society.
- KOHEL-KNABNER, I. 2002. The macromolecular organic composition of plant and microbial residues as inputs to soil organic matter. *Soil Biology and Biochemistry*, 34, 139-162.
- KOIZUMI, K., KUBOTA, Y., OZAKI, H., SHIGENOBU, K., FUKUDA, M. & TANIMOTO, T. 1992. Analyses of isomeric mono-O-methyl-D-glucoses, D-glucobioses and D-glucose monophosphates by high-performance anion-exchange chromatography with pulsed amperometric detection. *J. Chromatogr.*, 595, 340-345.
- KONG, X., XIE, J., WU, X., HUANG, Y. & BAO, J. 2005. Rapid Prediction of Acid Detergent Fiber, Neutral Detergent Fiber, and Acid Detergent Lignin of Rice Materials by Near-Infrared Spectroscopy. *Journal of Agricultural and Food Chemistry*, 53, 2843-2848.
- KOURTI, T. & MACGREGOR, J. F. 1995. Process analysis, monitoring and diagnosis, using multivariate projection methods. *Chemometrics and Intelligent Laboratory Systems*, 28, 3-21.
- KRISTENSEN, E. F. 1999. *Winter harvesting of Miscanthus*, Denmark, Bio Base.
- KRONGTAEW, C., MESSNER, K., TERS, T. & FACKLER, K. 2010a. Characterisation of Key Parameters for Biotechnological lignocellulose conversion assessed by FT-NIR spectroscopy part 1: Qualitative analysis of pretreated straw. *BioResources*, 5, 2063-2080.
- KRONGTAEW, C., MESSNER, K., TERS, T. & FACKLER, K. 2010b. Characterisation of Key Parameters for Biotechnological lignocellulose conversion assessed by FT-NIR spectroscopy part II: Quantitative analysis by partial least squares regression. *BioResources*, 5, 2081-2096.
- KRZANOWSKI, W. J. 1987. Cross-Validation in Principal Component Analysis. *Biometrics*, 43, 575-584.
- LABBÉ, N., DE JÉSO, B., LARTIGUE, J.-C., DAUDÉ, G., PÉTRAUD, M. & RATIER, M. 2002. Moisture content and extractive materials in maritime pine wood by Low Field 1H NMR. *Holzforschung*, 56, 25-31.
- LABBÉ, N., YE, X. P., FRANKLIN, J. A., WOMAC, A. R., TYLER, D. D. & RIALS, T. G. 2008. Analysis of switchgrass characteristics using near infrared spectroscopy. *Bioresources*, 3, 1329-1348.
- LAI, Y. Z. & SARKANEN, K. V. 1971. Isolation and structure studies. In: SARKANEN, K. V. & LUDWIG, C. H. (eds.) *Lignins. Occurrence, formation, structure and reactions*. New York: Wiley-Interscience.
- LAIDLAW, R. A. & REID, S. G. 1952. *J. Sci. Food and Agric.*, 3, 19-25.
- LASER, M., SCHULMAN, D., ALLEN, S. G., LICHWA, J., ANTAL, M. J. & LYND, L. R. 2002. A comparison of liquid hot water and steam pretreatments of sugar cane bagasse for bioconversion to ethanol. *Bioresource Technology*, 81, 33-44.
- LAU, C.-C., CHAN, C.-O., CHAU, F.-T. & MOK, D. K.-W. 2009. Rapid analysis of Radix puerariae by near-infrared spectroscopy. *Journal of Chromatography A*, 1216, 2130-2135.
- LAVARACK, B. P., GRIFFIN, G. J. & RODMAN, D. 2002. The acid hydrolysis of sugarcane bagasse hemicellulose to produce xylose, arabinose, glucose and other products. *Biomass and Bioenergy*, 23, 367-380.
- LAWFORD, H. & ROUSSEAU, J. 1997. Fermentation of biomass-derived glucuronic acid by pet expressing recombinants of E. coli B. *Applied Biochemistry and Biotechnology*, 63-65, 221-241.
- LE NGOC HUYEN, T., RÉMOND, C., DHEILLY, R. M. & CHABBERT, B. 2010. Effect of harvesting date on the composition and saccharification of *Miscanthus x giganteus*. *Bioresource Technology*, 101, 8224-8231.
- LEE, Y. C. 1996. Carbohydrate analyses with high-performance anion-exchange chromatography. *Journal of Chromatography A*, 720, 137-149.
- LEHRFELD, J. 1981. Differential gas-liquid chromatography method for determination of uronic acids in carbohydrate mixtures. *Analytical Biochemistry*, 115, 410-418.
- LEWANDOWSKI, I., CLIFTON-BROWN, J. C., ANDERSSON, B., BASCH, G., CHRISTIAN, D. G., JORGENSEN, U., JONES, M. B., RICHE, A. B., SCHWARZ, K.-U., TAYEBI, K. & TEIXEIRA, F. 2003a.

- Environment and harvest time affects the combustion qualities of *Miscanthus* genotypes. *Agron. J.*, 95, 1274-1280.
- LEWANDOWSKI, I. & HEINZ, A. 2003. Delayed harvest of miscanthus-influences on biomass quantity and quality and environmental impacts of energy production. *European Journal of Agronomy*, 19, 45-63.
- LEWANDOWSKI, I. & KICHERER, A. 1997. Combustion quality of biomass: practical relevance and experiments to modify the biomass quality of *Miscanthus x giganteus*. *European Journal of Agronomy*, 6, 163-177.
- LEWANDOWSKI, I., SCURLOCK, J. M. O., LINDVALL, E. & CHRISTOU, M. 2003b. The development and current status of perennial rhizomatous grasses as energy crops in the US and Europe. *Biomass and Bioenergy*, 25, 335-361.
- LI, B., MORRIS, J. & MARTIN, E. B. 2002. Model selection for partial least squares regression. *Chemometrics and Intelligent Laboratory Systems*, 64, 79-89.
- LI, X., XIMENES, E., KIM, Y., SLININGER, M., MEILAN, R., LADISCH, M. & CHAPPLE, C. 2010. Lignin monomer composition affects Arabidopsis cell-wall degradability after liquid hot water pretreatment. *Biotechnology for Biofuels*, 3, 27.
- LINDGREN, E., THEANDER, O. & AMAN, P. 1980. Chemical composition of timothy at different stages of maturation and of residues from feeding value determinations. *Swed. J. Agric. Res.*, 10, 3-10.
- LINDGREN, F., GELADI, P. & WOLD, S. 1993. The kernel algorithm for PLS. *Journal of Chemometrics*, 7, 45-59.
- LIU, L., YE, X. P., SAXTON, A. M. & WOMAC, A. 2010a. Pretreatment of near infrared spectral data in fast biomass analysis. *Journal of Near Infrared Spectroscopy*, 18, 317-331.
- LIU, L., YE, X. P., WOMAC, A. R. & SOKHANSANJ, S. 2010b. Variability of biomass chemical composition and rapid analysis using FT-NIR techniques. *Carbohydrate Polymers*, 81, 820-829.
- LLOYD, S. P. 1982. Least squares quantization in PCM. *IEEE Transactions on Information Theory*, IT-28, 129-137.
- LONG, S. P. 1983. C-4 photosynthesis at low temperatures. *Plant Cell and Environment*, 6, 345-363.
- LONG, S. P. 1999. Environmental responses. In: SAGE, R. F. & MONSON, R. K. (eds.) *C4 Plant Biology*. San Diego: Academic Press.
- LONG, S. P. & BEALE, C. V. 2001. Resource capture by *Miscanthus*. *Miscanthus for Energy and Fibre*. London: James and James.
- LUNDQVIST, J., JACOBS, A., PALM, M., ZACCHI, G., DAHLMAN, O. & STÅLBRAND, H. 2003. Characterization of galactoglucomannan extracted from spruce (*Picea abies*) by heat-fractionation at different conditions. *Carbohydrate Polymers*, 51, 203-211.
- LYND, L. R. 1996. Overview and evaluation of fuel ethanol from cellulosic biomass: technology, economics, the environment, and policy. *Annual Review of Energy and the Environment*, 21, 403-465.
- LYND, L. R., WEIMER, P. J., VAN ZYL, W. H. & PRETORIUS, I. S. 2002. Microbial cellulose utilization: fundamentals and biotechnology. *Microbiology and Molecular Biology Reviews*, 66, 506-77.
- MACKENZIE, D. J. & WYLAM, C. B. 1957. *J. Sci. Food and Agric.*, 8, 38-45.
- MACQUEEN, J. 1967. Some methods for classification and analysis of multivariate observations. In: LE CAM & NEYMAN, J. (eds.) *Proceedings 5th Berkeley Symposium on Mathematical Statistics and Probability*. Berkeley, CA: University of California Press.
- MAEDA, R. N., SERPA, V. I., ROCHA, V. A. L., MESQUITA, R. A. A., ANNA, L. M. M. S., DE CASTRO, A. M., DRIEMEIER, C. E., PEREIRA JR, N. & POLIKARPOV, I. 2011. Enzymatic hydrolysis of pretreated sugar cane bagasse using *Penicillium funiculosum* and *Trichoderma harzianum* cellulases. *Process Biochemistry*, 46, 1196-1201.

- MAEKAWA, E., ICHIZAWA, T. & KOSHIJIMA, T. 1989. An Evaluation of the Acid-Soluble Lignin Determination in Analyses of Lignin by the Sulfuric Acid Method. *Journal of Wood Chemistry and Technology*, 9, 549 - 567.
- MAHER, M. J. 1993. *Spent mushroom compost – Options for use*, Dublin, Teagasc.
- MAL, S. S. 1979. *Solid Fuel Chemistry*, 13, 130.
- MALLEY, D. F., WILLIAMS, P., MCLAUGHLIN, J. & ATKINSON, T. Rapid Analysis of Moisture, Organic Matter and Carbonate in Peat Cores from Northern Ontario by Near-infrared Spectroscopy. Annual Manitoba Soil Science Society Meeting, 2007 Winnipeg, Manitoba.
- MAO, H., GENCO, J. M., YOON, S.-H., VAN HEININGEN, A. & PENDSE, H. 2008. Technical Economic Evaluation of a Hardwood Biorefinery Using the "Near-Neutral" Hemicellulose Pre-Extraction Process. *Journal of Biobased Materials and Bioenergy*, 2, 177-185.
- MARCHAL, L. M., ZONDERVAN, J., BERGSMAN, J., BEEFTINK, H. H. & TRAMPER, J. 2001. Monte Carlo simulation of the α -amylolysis of amylopectin potato starch. Part I: modeling of the structure of amylopectin. *Bioproc. Biosys. Eng.*, 24, 163-170.
- MARK, H. 2001a. Data analysis: Multilinear regression and principal component analysis. In: BURNS, D. A. & CIURCZAK, E. W. (eds.) *Handbook of Near Infrared Analysis, Second Edition*. 2 ed. New York: Marcel Dekker.
- MARK, H. 2001b. Qualitative discriminant analysis. In: BURNS, D. A. & CIURCZAK, E. W. (eds.) *Handbook of Near Infrared Analysis, Second Edition*. New York: Marcel Dekker.
- MARK, T., DARBY, P. & SALASSI, M. Energy Cane Usage for Cellulosic Ethanol: Estimation of Feedstock Costs. Southern Agricultural Economics Association - 2009 Annual Meeting, January 31-February 3, 2009 Atlanta, Georgia
- MARTEN, G. C., HALGERSON, J. L. & SLEPER, D. A. 1988. Near Infrared reflectance spectroscopy evaluation of ruminal fermentation and cellulase digestion of diverse forages. *Crop Sci.*, 28, 163–167.
- MARTENS, H., NIELSEN, J. P. & ENGELSEN, S. B. 2003. Light Scattering and Light Absorbance Separated by Extended Multiplicative Signal Correction. Application to Near-Infrared Transmission Analysis of Powder Mixtures. *Analytical Chemistry*, 75, 394-404.
- MARTIN, M. E. & ABER, J. D. 1994. Analyses of forest foliage III: Determining nitrogen, lignin and cellulose in fresh leaves using near infrared reflectance data. *Journal of Near Infrared Spectroscopy*, 2, 25-32.
- MASSART, D. L., VANDEGINSTE, B. G. M., LEWI, P. J., SMEYERS-VERBEKE, J., BUYDENS, L. M. C. & DE JONG, S. 1998a. *Handbook of Chemometrics and Qualimetrics: Part A*, Elsevier Science Ltd.
- MASSART, D. L., VANDEGINSTE, B. G. M., LEWI, P. J., SMEYERS-VERBEKE, J., BUYDENS, L. M. C. & DE JONG, S. 1998b. *Handbook of Chemometrics and Qualimetrics: Part B*, Elsevier Science Ltd.
- MCCUTCHEON, G. A. 1997. *A study of the dry matter and nutrient value of pig slurry*. M. Sc., National University of Ireland.
- MCDONALD, P. 1981. *The biochemistry of silage*, New York, John Wiley and Sons.
- MCELHINNEY, J., DOWNEY, G. & FEARN, T. 1999. Chemometric processing of visible and near infrared reflectance spectra for species identification in selected raw homogenised meats. *Journal of Near Infrared Spectroscopy*, 7, 145-154.
- MCLELLAN, T. M., ABER, J. D., MARTIN, M. E., MELILLO, J. M. & NADELHOFFER, K. J. 1991. Determination of nitrogen, lignin, and cellulose content of decomposing leaf material by near infrared reflectance spectroscopy. *Canadian Journal of Forest Research*, 21, 1684-1688.
- MCMURROUGH, I. & ROSE, A. H. 1967. Effect of growth rate and substrate limitation on the composition and structure of the cell wall of *Saccharomyces cerevisiae*. *Biochem. J.*, 105, 189-203.
- MCTIERNAN, K. B., COÛTEAUX, M., BERG, B., BERG, M. P., DE ANTA, R. C., GALLARDO, A., KRATZ, W., PIUSSI, P., REMACLE, J. & DE SANTO, A. V. 2003. Changes in chemical composition of *Pinus sylvestris* needle litter during decomposition along a European coniferous forest climatic transect. *Soil Biology and Biochemistry*, 35, 801-812.

- MILNE, T. A., CHUM, H. L., AGBLEVOR, F. A. & JOHNSON, D. K. 1992. Standardised analytical methods. *Biomass Bioenergy*, 2, 341-366.
- MITCHELL, C. P. 1995. New cultural treatments and yield optimisation. *Biomass Bioenergy*, 9, 11-34.
- MITSUI, K., INAGAKI, T. & TSUCHIKAWA, S. 2008. Monitoring of hydroxyl groups in wood during heat treatment using NIR spectroscopy. *Biomacromolecules*, 9, 286-288.
- MOLLER, J. 2009. Gravimetric determination of acid detergent fiber and lignin in feed: interlaboratory study. *J AOAC Int.*, 92, 74-90.
- MOLLER, R., TOONEN, M., VAN BEILEN, J., SALENTIJJN, E. & CLAYTON, D. 2007. *Crop Platforms for Cell Wall Biorefining: Lignocellulose Feedstocks - Outputs from the EPOBIO Project*, Newbury, England, CPL Press.
- MONTANE, D., FARRIOL, X., SALVADÓ, J., JOLLEZ, P. & CHORNET, E. 1998. Application of steam explosion to the fractionation and rapid vapor-phase alkaline pulping of wheat straw. *Biomass and Bioenergy*, 14, 261-276.
- MONTEITH, J. L. 1977. Climate and the efficiency of crop production in Britain. *Philosophical Transactions of the Royal Society of London*, 281, 277-294.
- MONTEITH, J. L. 1978. Reassessment of maximum growth rates for C3 and C4 crops. *Experimental Agriculture*, 14, 1-5.
- MONTES, J. M., MIRDITA, V., PRASAD, K. V. S. V., BLUMMEL, M., DHILLON, B. S. & MELCHINGER, A. E. 2009. A new near infrared spectroscopy sample presentation unit for measuring feeding quality of maize stover *Journal of Near Infrared Spectroscopy*, 17, 195-201.
- MOSIER, N. S., HENDRICKSON, R., HO, N., SEDLAK, M. & LADISCH, M. R. 2005. Optimization of pH controlled liquid hot water pretreatment of corn stover. *Bioresource Technology*, 96, 1986-1993.
- MURPHY, J. D. & MCCARTHY, K. 2005. Ethanol production from energy crops and wastes for use as a transport fuel in Ireland. *Applied Energy*, 82, 148-166.
- MURRAY, I. & WILLIAMS, P. C. 1987. Chemical Principles of Near-Infrared Technology. In: WILLIAMS, P. C. & NORRIS, K. H. (eds.) *Near-Infrared Technology in the Agricultural and Food Industries*. St. Paul, MN.: Am. Assoc. Cereal Chem.
- MUSHA, Y. & GORING, D. A. I. 1974. Klason and acid-soluble lignin content of hardwoods. *Wood. Sci.*, 7, 133-134.
- NAES, T. 1987. The design of calibration in near infra-red reflectance analysis by clustering. *Journal of Chemometrics*, 1, 121-134.
- NAES, T. & ISAKSSON, T. 1991. Splitting of calibration data by cluster analysis. *Journal of Chemometrics*, 5, 49-65.
- NAES, T., ISAKSSON, T., FEARN, T. & DAVIES, T. 2007. *A User-Friendly Guide to Multivariate Calibration and Classification*, Chichester, UK, NIR Publications.
- NAIDENOV, V. I., KOPERIN, I. F. & PURIN, V. R. 1982. Physical and chemical properties of the ash of waste wood. *Lesnoi Zhurnal*, 2, 115-117.
- NEUREITER, M., DANNER, H., THOMASSER, C., SAIDI, B. & BRAUN, R. 2002. Dilute-acid hydrolysis of sugarcane bagasse at varying conditions. *Applied Biochemistry and Biotechnology*, 98-100, 49-58.
- NGUYEN, Q. 1998. *Milestone Completion Report: Evaluation of a Two-Stage Dilute Sulfuric Acid Hydrolysis Process*. Internal Report, Golden, Colorado, National Renewable Energy Laboratory.
- NIE, G. Y., LONG, S. P. & BAKER, N. R. 1992. The effects of development at suboptimal growth temperatures on photosynthetic capacity and susceptibility to chilling-dependent photoinhibition in Zea-Mays. *Physiologia Plantarum*, 85, 554-60.
- NIST. 2011. *National Institute of Standards and Technology - Report of Investigation: Reference Material 8492, Eastern Cottonwood Whole Biomass Feedstock* [Online]. Available: <https://www-s.nist.gov/srmors/reports/8492.pdf?CFID=1395767&CFTOKEN=d81fb3dfdc7be2c8->

- EF0A4A24-F021-876E-8D48F7B9C20A8055&jsessionid=f03054dbf854d7d3b4b31d124f21c5c69827 2011].
- NIXON, P. M. I. & BULLARD, M. J. 2001. Is Miscanthus suited to the whole of England and Wales? Preliminary studies. *Aspects of Applied Biology*, 65, 91-99.
- NKANSAH, K., DAWSON-ANDOH, B. & SLAHOR, J. 2010. Rapid characterization of biomass using near infrared spectroscopy coupled with multivariate data analysis: Part 1 yellow-poplar (*Liriodendron tulipifera* L.). *Biores. Technol.*, 101, 4570-4576.
- NORDKVIST, E. & AMAN, P. 1986. Changes during growth in anatomical and chemical composition and in vitro degradability of lucerne. *J. Sci. Food and Agric. Chem.*, 39, 473-477.
- NORRIS, K. H. 1984. *Reflectance Spectroscopy in Modern Methods of Food Analysis*, Westport, CT, AVI.
- NORRIS, K. H., BARNES, R. F., MOORE, J. E. & SHENK, J. S. 1976. Predicting forage quality by infrared reflectance spectroscopy. *Journal of Animal Science*, 43, 889-897.
- O'SHEA, M. G., STAUNTON, S. P. & BURLEIGH, M. 2010. Implementation of on-line near infrared (NIR) technologies for the analysis of cane, bagasse and raw sugar in sugar factories to improve performance. *Proc. Int. Sco. Sugar Cane Technol.*, 27, In Press.
- O'SHEA, M. G., STAUNTON, S. P., DONALD, D. & SIMPSON, J. 2011. Developing laboratory near infrared (NIR) instruments for the analysis of sugar factory products. *Proc. Aust. Soc. Sugar Cane Technol.*, 33, In Press.
- OECD 1997. *OECD Economic Surveys - Ireland 1997*, Paris, OECD.
- ONO, K., HIRAIDE, M. & AMARI, M. 2003. Determination of lignin, holocellulose, and organic solvent extractives in fresh leaf, litterfall, and organic material on forest floor using near-infrared reflectance spectroscopy. *Journal of Forest Research*, 8, 191-198.
- OSMAN, E. A. & GOSS, J. R. 1983. Ash chemical composition, deformation and fusion temperatures for wood and agricultural residues. *Am. Soc. Agric. Engineers Annual Meeting*, 83, 1-16.
- OSTEN, D. W. 1988. Selection of optimal regression models via cross-validation. *Journal of Chemometrics*, 2, 39-48.
- PAPATHEOFANOUS, M. G., KOUKIOS, E. G., MARTON, G. & DENCS, J. Characterisation of Miscanthus sinensis potential as an industrial and energy feedstock. In: CHARTIER, P., FERRERO, G. L., HENIUS, U. M., HULTBERG, S., SACHAU, J. & WIINBLAD, M., eds. Biomass for Energy and the Environment - Proceedings of the 9th European Bioenergy Conference, 24-27 June 1996, 1996 Copenhagen, Denmark. Elsevier Science Ltd., 504-8.
- PASQUINI, C. 2003. Near Infrared Spectroscopy: Fundamentals, Practical Aspects and Analytical Applications. *J. Braz. Chem. Soc.*, 14, 198-219.
- PAULY, M., ALBERSHEIM, P., DARVILL, A. G. & YORK, W. S. 1999. Molecular domains of the cellulose/xyloglucans network in the cell walls of the higher plants. *Plant Journal*, 20, 629-639.
- PEARCE, G. R. Generation and utilization of wastes from intensive piggeries. 48th ANZAAS Congress, 1977 Melbourne.
- PENN STATE UNIVERSITY. 2011. *Review Structures Used to Identify Grasses* [Online]. Available: http://www.forages.psu.edu/topics/species_variety_trials/commonpagrasses/structures.html [Accessed 15/4/11 2011].
- PENNINGTON, N. L. & BAKER, C. W. 1990. *Sugar: A User's Guide to Sucrose*, New York, New York, Van Nostrand Reinhold.
- PHILLIPS, S., ADEN, A., JECHURA, J., DAYTON, D. & EGGEMAN, T. 2007. *Thermochemical Ethanol via Indirect Gasification and Mixed Alcohol Synthesis of Lignocellulosic Biomass. Technical Report NREL/TP-510-41168*, Golden, Colorado, National Renewable Energy Laboratory.
- POHLANDT, K., STRECKER, M. & MARUTSZKY, R. 1993. Ash from the combustion of wood treated with inorganic wood preservatives; element composition and leaching. *Chemosphere*, 26, 2121-2128.

- POKE, F. S. 2006. Predicting Extractives, Lignin, and Cellulose Contents Using Near Infrared Spectroscopy on Solid Wood in *Eucalyptus globulus*. *Journal of Wood Chemistry and Technology*, 26, 187-199.
- POKE, F. S., WRIGHT, J. K. & RAYMOND, C. A. 2004. Predicting extractives and lignin contents in *Eucalyptus globulus* using near infrared reflectance analysis. *Journal of Wood Chemistry and Technology*, 24, 55-67.
- POLLOCK, J. S., O'HARA, I. M. & GRIFFIN, K. G. 2007. Aligning the drivers in the value chain - a new cane payment system for Mackay sugar. *Proc. Aust. Soc. Sugar Cane Technol.*, 29, 1-8.
- POPE, G., MCDOWALL, R., MASSEY, W. & STAUNTON, S. P. 2004. The use of NIR spectroscopy in a cane quality incentive scheme. *Proc. Aust. Soc. Sugar Cane Technol.*, 26, 1-8.
- PORTER, E. Poolbeg 2 original climate chapter calculations.xls. Presentation to An Bórd Pleanála, 26th April 2007, Croke Park, Dublin, 2007.
- POWER, N., MURPHY, J. D. & MCKEOGH, E. 2008. What crop rotation will provide optimal first-generation ethanol production in Ireland, from technical and economic perspectives? *Renewable Energy*, 33, 1444-1454.
- QUARAMBY, C. & ALLEN, S. E. 1989. Organic constituents. In: ALLEN, S. E. (ed.) *Chemical Analysis of Ecological Materials*. Oxford: Blackwell.
- QURESHI, N. & BLASCHEK, H. P. 1999. Production of Acetone Butanol Ethanol (ABE) by a Hyper-Producing Mutant Strain of *Clostridium beijerinckii* BA101 and Recovery by Pervaporation. *Biotechnol. Prog.*, 15, 594-602.
- RAISKILA, S., PULKKINEN, M., LAAKSO, T., FAGERSTEDT, K., LÖIJA, M., MAHLBERG, R., PAAJANEN, L., RITSCHKOFF, A.-C. & SARANPÄ, P. 2007. FTIR spectroscopic prediction of Klason and acid soluble lignin variation in Norway Spruce cutting clones *Silva Fennica*, 41, 351-371.
- RAMASWAMY, V., BOUCHER, O., HAIGH, J., HAUGLUSTAINE, D., HAYWOOD, J., MYHRE, G., NAKAJIMA, T., SHI, G. Y. & SOLOMON, S. 2001. Chapter 6. Radiative forcing of climate change. In: HOUGHTON, J. T., DING, D. J., GRIGGS, D. J., NOGUER, M., VAN DER LINDEN, P. J., DAI, X., MASKELL, K. & JOHNSON, C. A. (eds.) *Climate Change 2001: The scientific basis. Contribution of Working Group I to the third assessment report of the Intergovernmental Panel on Climate Change (IPCC)*. Cambridge University Press.
- RANATUNGA, T. D., JERVIS, J., HELM, R. F., MCMILLAN, J. D. & WOOLEY, R. J. 2000. The effect of overliming on the toxicity of dilute acid pretreated lignocellulosics: the role of inorganics, uronic acids and ether-soluble organics. *Enzyme and Microbial Technology*, 27, 240-247.
- RÄNNAR, S., LINDGREN, F., GELADI, P. & WOLD, S. 1994. A PLS kernel algorithm for data sets with many variables and fewer objects. Part 1: Theory and algorithm. *Journal of Chemometrics*, 8, 111-125.
- RAYMOND, C. R. & SCHIMLECK, L. R. 2002. Development of near infrared reflectance analysis calibrations for estimating genetic parameters for cellulose content in *Eucalyptus globulus*. *Can. J. For. Res.*, 32, 170-176.
- REES, D. A. 1967. *The Shapes of Molecules: Carbohydrate Polymers*, Contemporary Science Paperbacks.
- REEVES III, J. B. & VAN KESSEL, J. A. S. 2002. Spectroscopic analysis of dried manures. Near- versus mid-infrared diffuse reflectance spectroscopy for the analysis of dried dairy manures. *Journal of Near Infrared Spectroscopy*, 10, 93-101.
- REICHER, F., GANTER, J., RECHIA, C., SIERAKOWSKI, M. & GORIN, P. 1994. Uneven O-acetyl distribution in a lightly acetylated d-xylan from sugar cane. *Ciencia y Cultura (Sao Paulo)* 46, 283-285.
- RICE, B. 2001. *IENICA Report from the Republic of Ireland*, Oakpark, Carlow, TEGASC.
- RITTER, G. J., MITCHELL, R. L. & SEBORG, R. M. 1933. Some Factors that Influence the Conversion of Cellulosic Materials to Sugar. *Journal of the American Chemical Society*, 55, 2989-2991.
- RITTER, G. J., SEBORG, R. M. & MITCHELL, R. L. 1932. Factors affecting quantitative determination of lignin by 72% sulfuric acid method. *Ind. Eng. Chem.*, 4, 202-204.

- ROBERTS, C., WORKMAN, J. & REEVES, J. 2004. *Near-Infrared Spectroscopy in Agriculture*, Madison, WI, ASA-CSSA-SSSA.
- ROBERTS, J. C. 1996. *The Chemistry of Paper*, Cambridge, UK, Royal Society of Chemistry.
- ROCHE, D. 2006. National Strategy on Biodegradable Waste. Dublin, Ireland: Minister for Environment, Heritage and Local Government.
- ROCKLIN, R. D., CLARKE, A. P. & WEITZHANDLER, M. 1998. Improved Long-Term Reproducibility for Pulsed Amperometric Detection of Carbohydrates via a New Quadruple-Potential Waveform. *Analytical Chemistry*, 70, 1496-1501.
- RODRÍGUEZ-CHONG, A., RAMÍREZ, J. A., GARROTE, G. & VÁZQUEZ, M. 2004. Hydrolysis of sugar cane bagasse using nitric acid: a kinetic assessment. *Journal of Food Engineering*, 61, 143-152.
- ROUESSAC, F. & ROUESSAC, A. 1998. *Chemical Analysis: Modern Instrumentation Methods and Techniques*, New York, NY, John Wiley and Sons.
- ROYLE, D. J., HUNTER, T. & PEI, M. H. 1992. *Evaluation of the biology and importance of diseases and pests in willow plantations, ETSU Contractors report ETSU B 1258*, Long Ashton Research Station, ETSU.
- RPS MCOS 2004. *An Assessment of the Renewable Energy Resource Potential of Dry Agricultural Residues in Ireland*, Sustainable Energy Ireland.
- RUTHERFORD, I. & BELL, A. 1992. Economic Appraisal. In: RUTHERFORD, I. (ed.) *The Potential of Miscanthus as a Fuel Crop, ETSU B 1354*. Harwell: ETSU.
- SANDERSON, M. A., AGBLEVOR, F., COLLINS, M. & JOHNSON, D. K. 1996. Compositional analysis of biomass feedstocks by near infrared reflectance spectroscopy. *Biomass and Bioenergy*, 11, 365-370.
- SASKA, M. & OZER, E. 1995. Aqueous extraction of sugarcane bagasse hemicellulose and production of xylose syrup. *Biotechnology and Bioengineering*, 45, 517-523.
- SCHAFFLER, K. J., DUNSMORE, A. N. & MEYER, J. H. 1993. Rapid analysis of sugar products by near infra red spectroscopy. *Proceedings of The South African Sugar Technologists' Association*, June - 1993, 222-9.
- SCHIMLECK, L. R., WRIGHT, P. J., MICHELL, A. J. & WALLIS, A. F. A. 1997. Near-infrared spectra and chemical compositions of *E. globulus* and *E. nitens* plantation woods. *Appita*, 50, 40-46.
- SCHMER, M. R., VOGEL, K. P., MITCHELL, R. B. & PERRIN, R. K. 2008. Net energy of cellulosic ethanol from switchgrass. *PNAS*, 105, 464-469.
- SCHNEIDER, M. H. Correcting dry matter loss calculations in biomass fuels containing volatile extractives. Proc. IEA/BA Task VI Activity 5, 1995 Garpenberg, Sweden, 13-16 June, 1994. Swedish University of Agricultural Sciences, Department of Operational Efficiency.
- SHALIZI, C. 2009. *Distances between Clustering, Hierarchical Clustering - Lecture of 14 September 2009* [Online]. Available: <http://www.stat.cmu.edu/~cshalizi/350/lectures/08/lecture-08.pdf> [Accessed 12/7/11 2011].
- SHARMA, H. S. S. 1996. Compositional analysis of neutral detergent, acid detergent, lignin and humus fractions of mushroom compost. *Thermochimica Acta*, 285, 211-220.
- SHENG, C. & AZEVEDO, J. L. T. 2005. Estimating the higher heating value of biomass fuels from basic analysis data *Biomass and Bioenergy*, 28, 499-507.
- SHENK, J. S., WORKMAN, J. J. & WESTERHAUS, M. O. 2008. Application of NIR Spectroscopy to Agricultural Products. In: BURNS, D. A. & CIURCZAK, E. W. (eds.) *Handbook of Near-Infrared Analysis - Third Edition*. 3 ed. Boca Raton, Florida: CRC Press.
- SIESLER, H. W. 2008. Basic Principles of Near-Infrared Spectroscopy. In: BURNS, D. A. & CIURCZAK, E. W. (eds.) *Handbook of Near-Infrared Analysis - Third Edition*. 3 ed. Boca Raton, Florida: CRC Press.
- SIMPSON, J., STAUNTON, S. P. & O'SHEA, M. G. 2011. A Review of NIR Applications for Process Control Purposes. *Proc. Aust. Soc. Sugar Cane Technol.*, 33, 1-8.

- SINDHU, R., BINOD, P., SATYANAGALAKSHMI, K., JANU, K., SAJNA, K., KURIEN, N., SUKUMARAN, R. & PANDEY, A. 2010. Formic Acid as a Potential Pretreatment Agent for the Conversion of Sugarcane Bagasse to Bioethanol. *Applied Biochemistry and Biotechnology*, 162, 2313-2323.
- SINELLI, N., CERRETANI, L., EGIDIO, V. D., BENDINI, A. & CASIRAGHI, E. 2010. Application of near (NIR) infrared and mid (MIR) infrared spectroscopy as a rapid tool to classify extra virgin olive oil on the basis of fruity attribute intensity. *Food Research International*, 43, 369-375.
- SINNAEVE, G., DARDENNE, P. & AGNEESSENS, R. 1994. Global or local? A choice for NIR calibrations in analyses of forage quality. *Journal of Near Infrared Spectroscopy*, 2, 163-175.
- SJOSTROM, E. 1981. *Wood Chemistry: Fundamentals and Applications*, London, UK, Academic Press, Inc.
- SLUITER, A., HAMES, B., RUIZ, R., SCARLATA, C., SLUITER, J., TEMPLETON, D. & CROCKER, D. 2006a. *Determination of Structural Carbohydrates and Lignin in Biomass*, Golden, Colorado, National Renewable Energy Laboratory.
- SLUITER, J., SCARLATA, C., SLUITER, A., HAMES, B. & THANH, N.-D. Automating Biomass Analysis Using Accelerated Solvent Extraction *In: LABORATORY, N. R. E.*, ed. 27th Symposium on Biotechnology for Fuels and Chemicals, 2006b Denver, CO.
- SLUITER, J. B., RUIZ, R. O., SCARLATA, C. J., SLUITER, A. D. & TEMPLETON, D. W. 2010. Compositional Analysis of Lignocellulosic Feedstocks. 1. Review and Description of Methods. *Journal of Agricultural and Food Chemistry*, 58, 9043-9053.
- SMITH, D. 1973a. The nonstructural carbohydrates. *In: BUTLER, G. W. & BAILEY, R. W. (eds.) Chemistry and Biochemistry of Herbage*. New York: Academic Press.
- SMITH, L. W. Nutritive evaluations of animal manures. *In: INGLETT, G. E.*, ed. Processing agricultural and municipal wastes., 1973b Westport, CT. Avi. Publ. Co.
- SNOWMAN, J. W. 1988. *Downstream processes: equipment and techniques*, New York, Alan R. Liss, Inc.
- SOLOMON, B. D. & BANERJEE, A. 2006. A global survey of hydrogen energy research, development and policy. *Energy Policy*, 34, 781-792.
- SOLOMON, B. D., BARNES, J. R. & HALVORSEN, K. E. 2007. Grain and cellulosic ethanol: History, economics, and energy policy. *Biomass and Bioenergy*, 31, 416-425.
- SONDEREGGER, M. & SAUER, U. 2003. Evolutionary engineering of *Saccharomyces cerevisiae* for anaerobic growth on xylose. *Applied and Environmental Microbiology*, 69, 1990-1998.
- SOUTHGATE, D. A. T. 1976. The analysis of dietary fibre. *In: SPILLER, G. A. & AMEN, R. J. (eds.) Fibre in human nutrition*. Boca Raton, FL: CRC Press.
- SPINK, J. & BRITT, C. 1998. *Crops for set-aside land: an economic and environmental appraisal*, England, ADAS.
- STAUNTON, S. P. & WARDROP, K. 2006. Development of an online bagasse analysis system using NIR spectroscopy. *Proc. Aust. Soc. Sugar Cane Technol.*, 28, 1-8.
- STOMBAUGH, S. K., JUNGB, H. G., ORFA, J. H. & SOMERSA, D. A. 2000. Genotypic and Environmental Variation in Soybean Seed Cell Wall Polysaccharides. *Crop Science*, 40, 408-412.
- STYLES, D. & JONES, M. B. 2007. Current and future financial competitiveness of electricity and heat from energy crops: A case study from Ireland. *Energy Policy*, 35, 4355-4367.
- STYLES, D., THORNE, F. & JONES, M. B. 2008. Energy crops in Ireland: An economic comparison of willow and *Miscanthus* production with conventional farming systems. *Biomass and Bioenergy*, 32, 407-421.
- SUGIYAMA, J. 1985. Lattice images from ultrathin sections of cellulose microfibrils in the cell wall of *Valonia macrophysa* Kutz. *Planta*, 166, 161-168.
- SULLIVAN, J. & DOUEK, M. 1994. Determination of carbohydrates in wood, pulp and process liquor samples by high-performance anion-exchange chromatography with pulsed amperometric detection. *Journal of Chromatography A*, 671, 339-350.

- SUN, J.-X., SUN, R., SUN, X.-F. & SU, Y. 2004a. Fractional and physico-chemical characterization of hemicelluloses from ultrasonic irradiated sugarcane bagasse. *Carbohydrate Research*, 339, 291-300.
- SUN, J. X., SUN, X. F., ZHAO, H. & SUN, R. C. 2004b. Isolation and characterization of cellulose from sugarcane bagasse. *Polymer Degradation and Stability*, 84, 331-339.
- SUN, R., LAWThER, J. M. & BANKS, W. B. 1996. Fractional and structural characterization of wheat straw hemicelluloses. *Carbohydrate Polymers*, 29, 325-331.
- SUN, Y. & CHENG, J. 2002. Hydrolysis of lignocellulosic materials for ethanol production: a review *Bioresource Technology*, 83, 1-11.
- SUNDSTOL, F., KOSSILA, V., THEANDER, O. & VESTERGAARD THOMSEN, K. 1978. Evaluation of the feeding value of straw. A comparison of the laboratory methods in the Nordic countries. *Acta Agric. Scand.*, 28, 10-16.
- SUNETHANOL. 2007. Available: <http://biopact.com/2007/08/sunethanol-secures-funding-for.html> [2007].
- SUSOTT, R. A., DEGROOT, W. F. & SHAFIZADEH, F. 1975. Heat content of natural fuels. *J. Fire and Flammability*, 6, 311-325.
- SVERZUT, C. B., VERMA, L. R. & D., F. A. 1987. Sugarcane analysis using near infrared spectroscopy. *Transactions of the ASAE*, 30, 255-8.
- SWAN, B. 1965. Isolation of acid-soluble lignin from the Klason lignin determination. *Svensk Papperstidn*, 68, 791-795.
- SZMANT, H. H. 1989. *Organic Building Blocks of the Chemical Industry*, New York, Wiley & Sons.
- TAPPI 1991. *TAPPI standard Um 250. Acid-soluble lignin in wood and pulp*, Atlanta, GA, TAPPI J.
- TAPPI 1997. *Solvent extractives of wood and pulp.*, Technical Association of the Pulp and Paper Industry.
- TAPPI 1998. *Acid-insoluble lignin in wood and pulp*, Technical Association of the Pulp and Paper Industry.
- TAUB, D. R. & LERDAU, M. T. 2000. Relationship between leaf nitrogen and photosynthetic rate for three NAD-ME and three NADP-ME C4 grasses. *American Journal of Botany*, 87, 412-417.
- TEAGASC 1994. *Monaghan Agricultural Waste Management Study.*, Wexford, Ireland, Teagasc.
- TEAGASC 2002. *Census of Mushroom Production*, Dublin., Teagasc.
- TEMPLETON, D. & EHRMAN, T. 1995. *Laboratory Analytical Procedure LAP-003: Determination of Acid-Insoluble Lignin in Biomass*, Golden, Colorado, NREL.
- TEMPLETON, D. W., SCARLATA, C. J., SLUITER, J. B. & WOLFRUM, E. J. 2010. Compositional Analysis of Lignocellulosic Feedstocks. 2. Method Uncertainties. *Journal of Agricultural and Food Chemistry*, 58, 9054-9062.
- TEWARI, J., MEHROTRA, R. & IRUDAYARAJ, R. 2003. Direct near infrared analysis of sugar cane clear juice using a fibre-optic transmittance probe. *Journal of Near Infrared Spectroscopy*, 11, 351-356.
- THAMMASOUK, K., TANDJO, D. & PENNER, M. H. 1997. Influence of Extractives on the Analysis of Herbaceous Biomass. *Journal of Agricultural and Food Chemistry*, 45, 437-443.
- THE ECONOMIST. 2007. The end of cheap food. *The Economist*, 6-12-07.
- THEANDER, O. 1983. Advances in the chemical characterisation and analytical determination of dietary fibre components. In: BIRCH, G. G. & PARKER, K. J. (eds.) *Dietary fibre*. London: Appl. Sci. Publ.
- THEANDER, O. 1985. Chemical investigations in the Swedish agrobioenergy project. In: PALZ, W. (ed.) *Energy from biomass*. London: Appl. Sci. Publ.
- THEANDER, O. 1991. Chemical characterisation of some potential lignocellulosic materials in the Swedish Agro-fibre project. In: GALETTI, G. C. (ed.) *Production and utilisation of lignocellulosics*. London: Appl. Sci. Publ.
- THEANDER, O. & AMAN, P. 1980. Chemical composition of some forages and various residues from feeding value determinations. *J. Sci. Food and Agric.*, 31, 31-37.

- THEANDER, O. & AMAN, P. 1984. Anatomical and chemical characteristics. *In: SUNDSTOL, F. & OWEN, E. (eds.) Straw and other fibrous by-products as feed.* Amsterdam: Elsevier Sci. Publ.
- THEANDER, O., AMAN, P., WESTERLUND, E., ANDERSSON, R. & PETTERSSON, D. 1995. Total dietary fiber determined as neutral sugar residues, uronic acid residues, and Klason lignin (The Uppsala Method): Collaborative Study. *J. AOAC Int.*, 78, 1030-1044.
- THEANDER, O. & WESTERLUND, E. 1993. Quantitative Analysis of Cell Wall Components. *In: JUNG, H. G., BUXTON, D. R., HATFIELD, R. D. & RALPH, J. (eds.) Forage Cell Wall Structure and Digestibility.* Madison, WI: ASA-CSSA-SSSA.
- THOMAS, B. 1972. Beitrage zur Nomenklatur und Analytik pflanzlicher Zellwandsubstanzen Getreide. *Mehl Brot*, 26, 158-169.
- TIJMENSEN, M. J. A., FAAIJ, A. P. C., VAN HARDEVELD, M. R. M. & HAMELINCK, C. N. 2002. Exploration of the possibilities for production of Fischer Tropsch liquids and power via biomass gasification. *Biomass and Bioenergy*, 23, 129-152.
- TILMAN, D., HILL, J. & LEHMAN, C. 2006. Carbon-negative biofuels from low-input high-diversity grassland biomass. *Science*, 314, 1598-1600.
- TIMELL, T. E. 1965. *Adv. Carbohyd. Chem.*, 20, 410.
- TIMOKHIN, B. V., BARANSKY, V. A. & ELISEEVA, G. D. 1999. Levulinic acid in organic synthesis. *Russian Chemical Reviews*, 68, 73-84.
- TSUCHIKAWA, S. & SIESLER, H. W. 2003a. Near-infrared spectroscopic monitoring of the diffusion process of deuterium-labeled molecules in wood. Part I: Softwood. *Appl. Spectrosc.*, 57, 667-674.
- TSUCHIKAWA, S. & SIESLER, H. W. 2003b. Near-infrared spectroscopic monitoring of the diffusion process of deuterium-labeled molecules in wood. Part II: Hardwood. *Appl. Spectrosc.*, 57, 675-681.
- TSUCHIKAWA, S., YONENOBU, H. & SIESLER, H. W. 2005. Near-infrared spectroscopic observation of the ageing process in archaeological wood using a deuterium exchange method. *Analyst*, 130, 379-384.
- ÜNER, B., KARAMAN, İ., TANRIVERDI, H. & ÖZDEMİR, D. 2011. Determination of lignin and extractive content of Turkish Pine trees using near infrared spectroscopy and multivariate calibration. *Wood science and Technology*, 45, 121-134.
- US CONGRESS 2007. Energy Independence and Security Act of 2007.
- VAN DEN BROEK, R., TEEUWISSE, S., HEALION, K., KENT, T., VAN WIJK, A., FAAIJ, A. & TURKENBURG, W. 2001. Potentials for electricity production from wood in Ireland. *Energy Fuels*, 26, 991-1013.
- VAN LIER, J. C., VAN GINKEL, J. T., STRAATSMA, G., GERRITS, J. P. G. & VAN GRIENSVEN, L. J. L. D. 1994. Composting of mushroom substrate in a fermentation tunnel compost parameters and a mathematical model. *Netherlands Journal of Agricultural Science*, 42, 271-292.
- VAN SOEST, P. J. 1963a. Use of detergents in analysis of fibrous feeds II A rapid method for the determination of fibre and lignin. *J. Assoc. Off. Agric. Chem.*, 48, 829-835.
- VAN SOEST, P. J. 1963b. Use of detergents in the analysis of fibrous feeds. I. Preparation of fibre residues of low nitrogen content. *J. Assoc. Off. Agric. Chem.*, 46, 825-829.
- VAN SOEST, P. J. 1982. *Nutritional ecology of the ruminant*, Corvallis, OR, O and B Books.
- VAN SOEST, P. J. 1993. Cell Wall Matrix Interactions and Degradation - Session Synopsis. *In: JUNG, H. G., BUXTON, D. R., HATFIELD, R. D. & RALPH, J. (eds.) Forage cell wall structure and digestibility.* Madison, WI: ASA-CSSA-SSSA.
- VAN SOEST, P. J. & WINE, R. H. 1967. Use of detergent in the analysis of fibrous feeds IV Determination of plant cell-wall constituents. *J. Assoc. Off. Anal. Chem.*, 50, 50-55.
- VAN SOEST, P. J. & WINE, R. H. 1968. Determination of lignin and cellulose in acid-detergent fibre with permanganate. *J. Assoc. Off. Anal. Chem.*, 51, 780-785.
- VAVROVA, P., STENBERG, B., KARSISTO, M., KITUNEN, V., TUPANILA, T. & LAIHO, R. 2008. Near Infrared Reflectance Spectroscopy for Characterization of Plant Litter Quality: Towards a

- Simpler Way of Predicting Carbon Turnover in Peatlands? In: VYMAZAL, J. (ed.) *Wastewater Treatment, Plant Dynamics and Management in Constructed and Natural Wetlands*. New York: Springer-Science.
- VENTURI, P., GIGLER, J. K. & HUISMAN, W. 1999. Economical and technical comparison between herbaceous (*Miscanthus x giganteus*) and woody energy crops (*Salix viminalis*). *Renewable Energy*, 16, 1023-1026.
- VENTURI, P. & VENTURI, G. 2003. Analysis of energy comparison in European agricultural systems. *Biomass and Bioenergy*, 25, 235-255.
- VERGNOUX, A., GIULIANO, M., LE DRÉAU, Y., KISTER, J., DUPUY, N. & DOUMENQ, P. 2009. Monitoring of the evolution of an industrial compost and prediction of some compost properties by NIR spectroscopy. *Science of The Total Environment*, 407, 2390-2403.
- VISSER, P. & PIGNATELLI, V. 2001. Utilisation of *Miscanthus*. In: B., J. M. & WALSH, M. (eds.) *Miscanthus for energy and fibre*. London: James and James, Ltd.
- VON POST, L. International Soil Science Conference, Helsingfors, 1924. 287.
- WALSH, M. 1999. *Economic Analysis of production of the three most promising energy crops in Ireland*, Denmark, BTGS.
- WANG, M., WU, M. & HUO, H. 2007. Life-cycle energy and greenhouse gas emission impacts of different corn ethanol plant types. *Environmental Research Letters*, Online.
- WARD JR., J. H. 1963. Hierarchical grouping to optimize an objective function. *J. Am. Stat. Assoc.*, 58, 236-244.
- WATANABE, A., MORITA, S. & OZAKI, Y. 2006. Temperature-dependent structural changes in hydrogen bonds in microcrystalline cellulose studied by infrared and nearinfrared spectroscopy with perturbation-correlation moving-window two-dimensional correlation analysis. *Appl. Spectrosc.*, 60, 611-618.
- WESTMORELAND, A. H., PEATEY, G. M. & PAYNE, R. C. 2005. Implementation of the NIR based cane analysis system at the Maryborough sugar factory limited. *Proc. Aust. Soc. Sugar Cane Technol.*, 27, 378-86.
- WICKHOLM, K. 2001. *Structural elements in native celluloses*, Sweden, Royal Institute of Technology.
- WIESENBERG, G. L. B., SCHWARK, L. & SCHMIDT, M. W. I. 2004. Improved automated extraction and separation procedure for soil lipid analyses. *European Journal of Soil Science*, 55, 349-356.
- WILKIE, K. C. B. 1979. The hemicelluloses of grasses and cereals. *Adv. Carbohyd. Chem. Biochem.*, 36, 215-264.
- WILLIAMS, B. C., MCMULLAN, J. T. & MCCAHEY, S. 2000. *Investigation into the recovery of energy from spent mushroom compost. Final Report*, University of Ulster, Northern Ireland Centre for Energy Research and Technology.
- WILLIAMS, B. C., MCMULLAN, J. T. & MCCAHEY, S. 2001. An initial assessment of spent mushroom compost as a potential energy feedstock. *Bioresource Technology*, 79, 227-30.
- WILLIAMS, R. B. 2007. *Biofuels from Municipal Wastes - Background Discussion Paper*, California, University of California.
- WILSON, J. R. 1990. Influence of plant anatomy on digestion and fibre breakdown. In: AKIN, D. E. (ed.) *Microbial and plant opportunities to improve the utilisation of lignocellulose by ruminants*. New York, NY: Elsevier Sci. Publ.
- WILSON, J. R. 1993. Organisation of forage plant tissues. In: JUNG, H. J., BUXTON, D. R., HATFIELD, R. D. & RALPH, J. (eds.) *Forage cell wall structure and digestibility*. Madison, WI: Amer. Soc. Agron.
- WISELOGEL, A. E., AGBLEVOR, F. A., JOHNSON, D. K., DEUTCH, S., FENNELL, J. A. & SANDERSON, M. A. 1996. Compositional changes during storage of large round switchgrass bales. *Bioresource Technology*, 56, 103-109.
- WITWER, S. H. 1974. *Bioscience*. 24, 216.
- WOLD, S. 1976. Pattern recognition by means of disjoint principal components models. *Pattern Recogn.*, 8, 127-139.

- WOLD, S. 1978. Cross-Validatory Estimation of the Number of Components in Factor and Principal Components Models. *Technometrics*, 20, 397-405.
- WOLD, S., SJOSTROM, M. & ERIKSSON, L. 2001. PLS-Regression: a basic tool of chemometrics. *Chemometrics and Intelligent Laboratory Systems*, 58, 109-130.
- WOLF, D. D. & FISKE, D. A. 1995. *Planting and managing switchgrass for forage, wildlife, and conservation.*, Virginia Polytechnic Institute and State University, Extension agronomist, Forages. Virginia Tech.
- WORKMAN, J. J. 1996. Process chemometrics/spectroscopy terminology. *NIR News*, 7, 15-16.
- WORKMAN, J. J. 2001. NIR Spectroscopy Calibration Basics. In: BURNS, D. A. & CIURCZAK, E. W. (eds.) *Handbook of Near Infrared Analysis, Second Edition*. New York: Marcel Dekker.
- WORKMAN, J. J. & BURNS, D. A. 2001. Commercial NIR Instrumentation. *Handbook of Near-Infrared Analysis - Second Edition*. New York, NY: Marcel Dekker
- WORRALL, J. J. & ANDERSON, K. M. 1993. Sample Preparation for Analysis of Wood Sugars by Anion Chromatography. *Journal of Wood Chemistry and Technology*, 13, 429-437.
- WRIGHT, P. J. & WALLIS, A. F. A. 1998. Rapid determination of cellulose in plantation eucalypt woods to predict kraft pulp yields. *Tappi J.*, 81, 126.
- WYMAN, C. E., DALE, B. E., ELANDER, R. T., HOLTZAPPLE, M., LADISCH, M. R. & LEE, Y. Y. 2005. Coordinated development of leading biomass pretreatment technologies. *Bioresource Technology*, 96, 1959-1966.
- YASUDA, S., FUKUSHIMA, K. & KAKEHI, A. 2001. Formation and chemical structures of acid soluble lignin I. Sulfuric acid treatment time and acid soluble lignin content of hardwood. *J Wood Sci*, 47.
- YE, X. P., LIU, L., HAYES, D., WOMAC, A., HONG, K. & SOKHANSANJ, S. 2008. Fast classification and compositional analysis of cornstover fractions using Fourier transform near-infrared techniques. *Biores. Technol.*, 99, 7323-7332.
- YONENOBU, H., TSUCHIKAWA, S. & SATO, K. 2009. Near-infrared spectroscopic analysis of aging degradation in antique washi paper using a deuterium exchange method. *Vibrational Spectroscopy*, 51, 100-104.
- YOSHIDA, M., LIU, Y., UCHIDA, S., KAWARADA, K., UKAGAMI, Y., ICHINOSE, H., KANEKO, S. & FUKUDA, K. 2008. Effects of cellulose crystallinity, hemicellulose, and lignin on the enzymatic hydrolysis of *Miscanthus sinensis* to monosaccharides. *Bioscience, biotechnology, and biochemistry*, 72, 805-10.
- YU, J. & STAHL, H. 2008. Microbial utilization and biopolyester synthesis of bagasse hydrolysates. *Bioresource Technology*, 99, 8042-8048.
- ZYMARK 1999. *Turbovap LV Evaporator Operator's Manual*, Hopkinton, MA, Zymark Corporation.