COMPARISON OF NEAR-INFRARED SPECTROSCOPY MODELS UTILIZED TO PREDICT LIGNOCELLULOSIC CONSTITUENTS IN WOOD SAMPLES

Paula Seppälä

Bachelor’s thesis
November 2015
Degree Programme in
Environmental Engineering

TAMPEREEN AMMATTIKORKEAKOULU
Tampere University of Applied Sciences

ABSTRACT

Tampereen ammattikorkeakoulu
Tampere University of Applied Sciences
Degree Programme in Environmental Engineering

SEPPÄLÄ, PAULA:
Comparison of near-infrared spectroscopy models utilized to predict lignocellulosic constituents in wood samples

Bachelor's thesis 57 pages, appendices 8 pages
November 2015

To have alternatives for the petroleum-based fuels and chemicals, biomass as a resource shows promising results. There are various ways of converting this abundant feedstock to value-added products by biorefinering, but in order to enable the evaluation of process economics and conversion yields, accurate compositional analysis of the feedstock is required. The standard compositional laboratory analysis are often time consuming, laborious and feedstock constructive. The need for faster biomass analyzing techniques could be met by using near-infrared spectroscopy and mathematical techniques to create predictive models.

This study was made for an Irish company, Celignis Ltd, with the help and knowledge of the CEO, Daniel Hayes. The aim of this work was to evaluate the accuracy predicted lignocellulosic constituents in wood samples by near-infrared spectroscopy models, the main objective being the comparison of models based solely on wood samples and models based on wider variety of also other biomass feedstocks. These models were referred to as local and global, respectively. Concentrations of four constituents; glucose, Klason lignin, xylose and mannose were analyzed and the capability of the models to predict these constituents in external wood sample set analyzed.

The results showed that both of the models were able to predict these components with reasonable accuracy. However, the wood-specific local models did not have increased predictive accuracies for glucose, xylose and mannose compared to the already existing global Celignis models. Even though good improvement was noted for the Klason lignin in the local model, based on these full results it is not justified to move from global to a local-model system.

Key words: biomass analysis, near-infrared spectroscopy, chemometrics, wood

TABLE OF CONTENTS

ABBREVIATIONS AND TERMS

Acid hydrolysis	acid catalysed cleavage of chemical bonds with addition of water
AIA	acid insoluble ash
AIR	acid insoluble residue
ASL	acid soluble lignin
calib	statistics for the calibration set
calib:valid	number of samples in the calibration (calib) and validation (valid) sets
CV	statistics for cross-validation
DJ	samples that had been ground to a particle size < 850 µm
DP	degree of polymerization
Global models	In this work refers to the created models based on average 750 different biomass samples
ISTD	internal standard
KL	Klason lignin (AIR minus AIA)
LAP	Laboratory Analytical Procedure
Local models	In this work refers to the created models based on 110 different wood samples
MC	moisture content (wet basis)
N	number of samples
NIR(S)	near-infrared (spectroscopy)
pred	statistics for the independent validation set
PRESS	prediction error sum of squares
R^2	coefficient of determination
RER	range error ratio
RMSEC	root mean square error of calibration
RMSECV	root mean square error of cross validation
RMSEP	root mean square error of prediction
RPD	ratio of standard error of performance to standard deviation
SD	standard deviation, $\sigma = \sqrt{\frac{\Sigma(y-\bar{y})^2}{N-1}}$, where $\bar{y}$= mean
SDD	standard deviation of the duplicates

SELR	standard error of laboratory as a percentage of mean analyte concentration
SEP	standard error of prediction
SG	Savitzky–Golay derivative
SS	sitka spruce
TAMK	Tampere University of Applied Sciences
Uronic acid	compound derived from sugars by oxidizing hydroxyl group to a carboxylic acid
Wavel.	wavelength region used for model development (nm)

1 INTRODUCTION

The global depletion of natural resources has led to wide research in the field of sustainable production of energy and materials. To overcome the problems associated with global warming, depletion of petroleum reservoirs, substitutes for the current resources are researched. Biorefineries, integrated facilities utilizing different types of biomasses in the production of various bioproducts, fuels, materials, chemicals and power, are seen as a promising alternative. Lignocellulosic feedstocks, such as wood, are abundant, do not directly compete with food resources and have various, though challenging, possibilities in the chemical and energy conversions. (Christopher 2013, 1-5)

To maintain sustainable processing of biomass, to reduce the costs and maximize the production yields, more rapid and accurate methods of compositional analysis are needed. Near-infrared spectroscopy (NIRS) can offer a fast and non-destructive analytical tool in predicting the constituents of interest in the biomass. In order to perform the NIR-analysis and to relate the spectral data to properties of the sample, robust standard reference methods and chemometrics (mathematical and statistical methods to interpret chemical information) are required. Accurate model can be used for rapid prediction of future samples of unknown compositions.

The purpose of this thesis was to compare the precision of the models based solely on wood samples (local, wood-specific) to the models based on larger variety as well as quantity of also other biomass types (global). The accuracy of the calibration was tested on external set of wood samples. The constituents of interest were restricted to glucose, xylose, mannose and Klason lignin for their sufficient concentrations and rather even distribution throughout different wood samples. The analytical procedures and model formation were performed in an Irish company, Celignis Ltd, with the help and knowledge of the CEO, Daniel Hayes.

2 THEORY

To understand the relevance of biomass compositional analysis, this chapter will present a short overview on the background behind the utilization of biomass. The chemical characterization of the studied feedstock, wood, is explained on the second part of this chapter. This theory is the essential base for creating the method used in this work. The main focus will be on the last part, where the use of near-infrared spectroscopy models as analytical tool to predict the constituents in the biomass, the theory behind this instrument and the development of chemometric models based on the NIR spectra are further specified.

2.1 Short overview for the utilization of biomass as a resource for value-added products

Biomass means any organic matter derived from living organism. Technology based on this feedstock is called biotechnology and can be used in many technological applications varying for example from healthcare to agriculture or industries. The conversion of biomass to liquid and gaseous fuels, heat, mechanical or electrical power, or chemicals is called biorefinering. (Yang 2007, 1-14) The main pillars of industrial biobased economy can be seen in Figure 1.

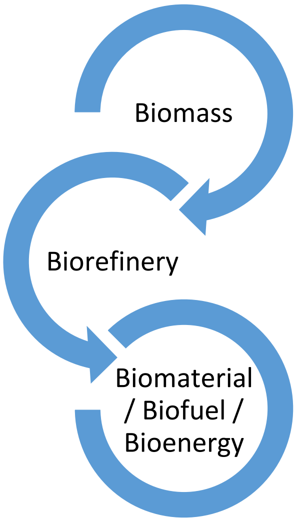


FIGURE 1. The basic pillars of industrial bioeconomy.

Examples of biomass resources include sugar and oily crops, cultivated energy crops, residues from agricultural, forestry or animal operations and municipal and industrial wastes. There are various ways of converting the raw biomass for useful commodities. The specific properties of the feedstock, such as sugar, moisture or energy content as well as other non-chemical factors, for example socio-economics in the country availability of the technology and environmental conditions define which conversion process is the most efficient to the certain biomass in a given locality. (Capareda 2014, 1-32)

These techniques include thermal, chemical or biological processes. An example of a thermal process is pyrolysis, where the organic matter is exposed to high temperature and anaerobic conditions, producing valuable solids, gases and oils. Whereas in a chemical process, the oil in the biomass together with alcohol and a catalyst are used to convert the feedstock to value-added products, such as biodiesel. In a biological conversion process the microbes do the work of converting the sugar into alcohol, or by breaking down the organic matter by anaerobic digestion to biogas. (Capareda 2014, 43-63) The Figure 2 below, adapted from the information by Capareda (2014), shows the schematic figure of these biomass conversion methods and their main products. Pre-treatment refers to various physical, chemical and enzymatic methods used to improve the conversion process, eg. size reduction or acid/alkaline treatments (Kelley 2015).

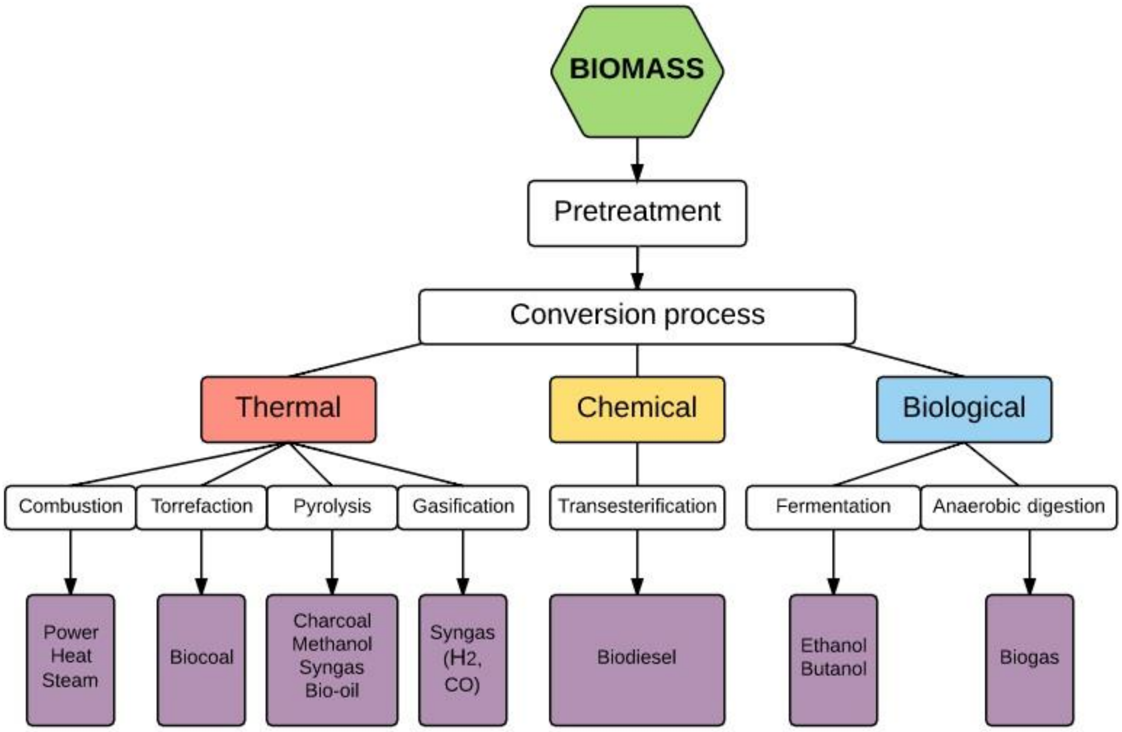


FIGURE 2. Flow chart of biomass conversion processes and their main products.

When biomass is used in biorefinering, it is classified either being first or second generation, depending on the origin. The first generation biofuels derive mainly from food crops such as sugar, starch or oil crops and are already commercially utilized in biorefineries. This is due to their rather easy degradation into sugar units (more about biomass properties in Chapter 2.2.) as well as other factors, such as already existing technologies and encouraging policies. However, the sustainable production of these crops is under review and the consumption of the already diminishing arable land areas and irrigation water is a concerning factor, since they compete with the food crops. In addition to this, due to the need for fertilizers and indirect land use effects, the 1st generation biofuels might not be produced sustainably. (Sims, Taylor, Saddler & Mabee 2008, 5-8)

These concerning factors have led to wide research in finding more sustainable options, and a rising interest in the field of second generation biofuels derived from lignocellulosic or so-called second generation feedstocks. These sources are mainly composed of crop and forest residues, short-rotation woody crops and wood wastes, and from the organic fraction of municipal waste. With these materials and careful planning it should be possible to avoid compromising the use of scarce arable land areas, remain the sustainability of the whole life cycle of the feedstock and applicable energy/product conversion process. Even though cultivation of energy crops requires arable land, the quality of the soil for the short rotation crops, grasses and woody crops does not have to be that high as for food and fibre crops, which means they would not directly compete with each other. (Sims et al. 2008, 7)

In any of the conversion processes, the overall knowledge of composition of the feedstock is essential. Biomass compositional analysis is of importance for example in the calculation of product yields, in configuration of efficient facilities or in finding of suitable enzymes used in catalysis of the process. In case of wood, as studied in this work, the pulp and paper industries can also benefit from a fast and accurate analysis of their material.

2.2 Chemistry of wood

In chemical terms, a major component of a living tree is water, but on dry basis wood composes mainly of three-dimensional network of polymers of sugar units; cellulose and hemicellulose and the biopolymer “cellular glue” lignin. Together these form the characteristic lignocellulosic structural macrostructure, where cellulose is embedded in a hemicellulose lignin matrix. Other components in lesser amounts include minor polysaccharides starch and pectin, non-structural components (extractives), proteins and inorganic trace constituents. (Rowell 2005, 36-45) The average compositional analysis of different softwood (mostly conifers, generally needle-leaved) and hardwood (mostly flowering plants, generally broadleaved) samples are presented in the Table 1 below. These values are gathered from information by Sjöström. (Sjöström 1981, 208) Even though being an important factor, especially in thermal conversion processes, heating value is not discussed here due to its irrelevance to the topic of sugar and lignin analysis.

TABLE 1. Indicative ranges of the main components of different soft- and hardwood species, % of dry wood.

	Softwoods[a]	Hardwoods[b]
Cellulose	33-42	38-51
Hemicelluloses (total)	19-31	15-34
Glucomannans	14-20	1-4
Xylans	5-11	14-30
Other polysaccharides	3-9	2-4
Lignin	27-31	21-31
Extractives	1.7-5.3	1.2-4.6

[a] Balsam fir, Douglas fir, Eastern hemlock, Common juniper, Monterey pine, Scots pine, Norway spruce, White spruce, Siberian larch

[b] Red maple, Sugar maple, Common beech, Silver birch, Paper birch, Grey alder, River red gum, Blue gum, Yemane, Black wattle, Balsa

2.2.1 Carbohydrates

Major chemical constituents in dry wood are carbohydrates composed of polymers cellulose and hemicellulose with minor portions of other sugar polymers such as starch and pectin.

Cellulose

Cellulose is the most abundant polymer found in nature and a major mass constituent of plant cell wall, approximately 30-50% in dry wood. Glucose, the six carbon hexose sugar with molecular formula $C_6H_{12}O_6$ is the building unit of the polymer cellulose. Several hundreds to thousands of glucose units in cellulose are linked together with beta (β) acetal (carbon with oxygen atoms on each side) glycosidic bonds. This covalent bonding of glucose molecules links the first and fourth carbon of the following molecule together, thus it is referred to as a β-1-4 glycosidic linkage. (Amarasekara 2014, 137-143) This peculiar direction of the β-acetal linkage results in a linear chain that can be seen in Figure 3.

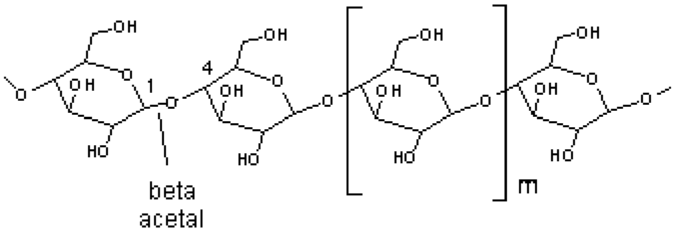


FIGURE 3. Cellulose consists of several glucose units (m) connected by a β-1-4 glycosidic linkage.

Notice in Figure 3 above, the most preferred pyranose (cyclic ring form) structure of glucose in solution. (Berg, Tymoczko & Stryer 2002) Through hydrogen bonding the multiple polar hydroxyl (–OH) groups on the glucose units bind to adjacent chains, which results in microfibrils, strong rod-like formations of cellulose, as can be seen in the Figure 4 below. This highly structural order, or the crystallinity, and length of cellulose microfibrils makes it rather resistant to deconstruction by physical or chemical treatments and provides the strength in the plant cell wall. (Amarasekara 2014, 137-143)

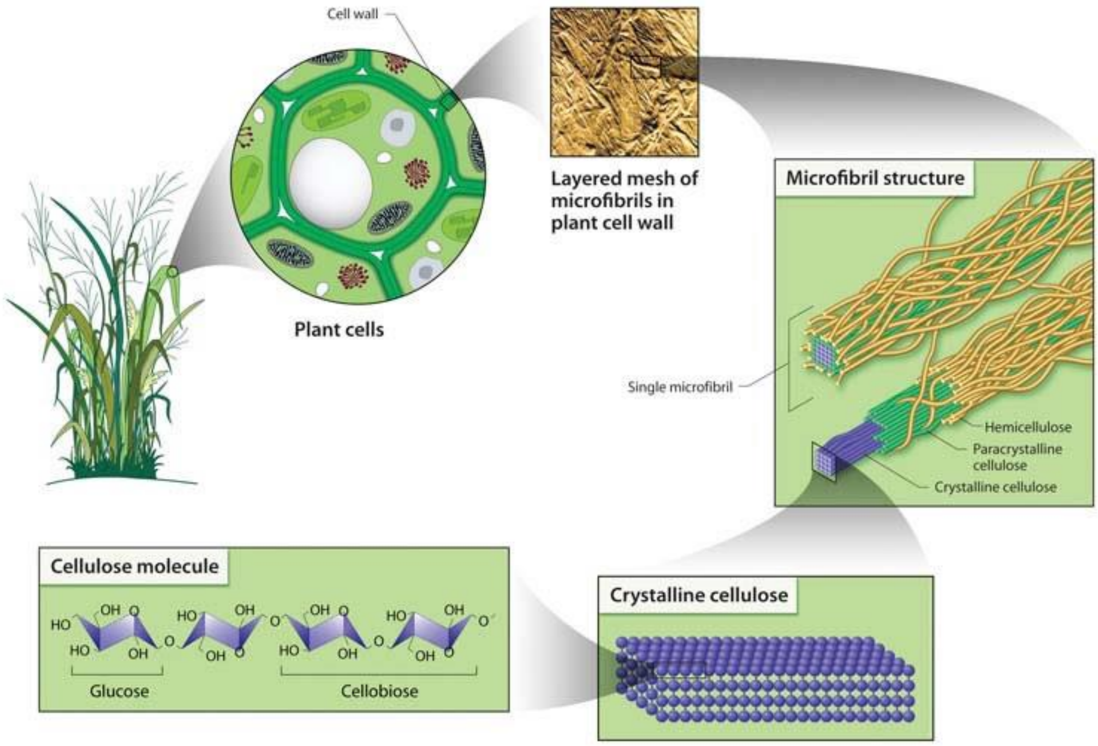


FIGURE 4. Simplified structure of the microfibril structure in plant cell wall. (US DOE 2005)

Due to the increasing packing density, these highly crystalline regions are formed and add up to 65% of the cellulose, the rest being amorphous and lacking this structural order (Rowell 2005, 36-40). In addition to glucose monomers during cellulose breakdown, glucose polymers of two (cellobiose) or more in length (cellodextrins) are also formed. Depending on the type of bacteria or yeast, glucose units or its polymers can be used directly in fermentation. (Shi & Weimer 1996)

Hemicellulose

The second major mass constituent (20-35%) of dry wood is the polymer hemicellulose. Instead of composing only of glucose units, hemicelluloses are co-polymers of two or more sugars and uronic acids. (Amarasekara 2014, 137-143) The monomer sugars include hexoses glucose, mannose, galactose and rhamnose as well as pentoses xylose and arabinose. Hemicelluloses are referred to by the combination of sugars they contain, for example like the primary softwood hemicellulose galactoglucomannan (galactose:glucose:mannose) or major hardwood hemicellulose glucuronoxylan (glucuronic acid:xylose), have varying ratios. (Rowell 2005, 39-43) Thus, it is noted here that since glucose can also be present in predominant amounts in hemicelluloses, approximating cellulose content directly from glucose concentration is not advised.

Figure 5 shows the main chain of both major hemicelluloses as the linear backbone and side groups, in black and in red, for galactoglucomannan and glucuronoxylan respectively (Kelley 2015). In contrast to cellulose's crystalline structure and high degree of polymerization (DP) (around 10 000), hemicellulose is amorphous, thus lacking the long-range order with average DP of 100-200. This makes it more available to acid hydrolysis. (Rowell 2005, 39-43)

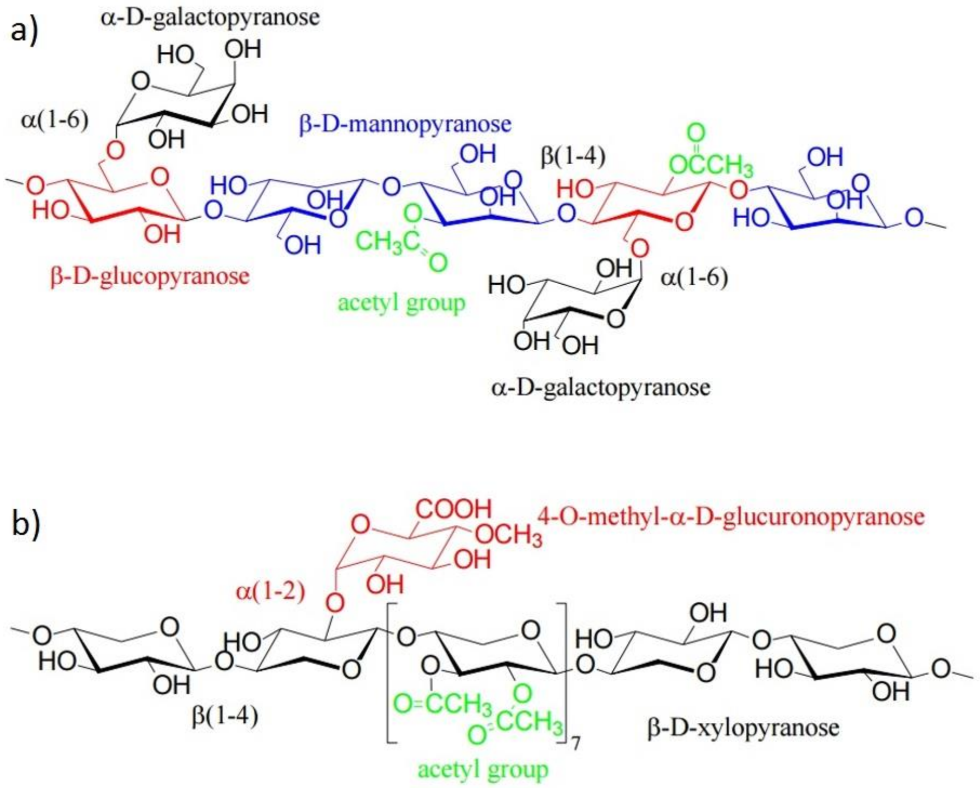


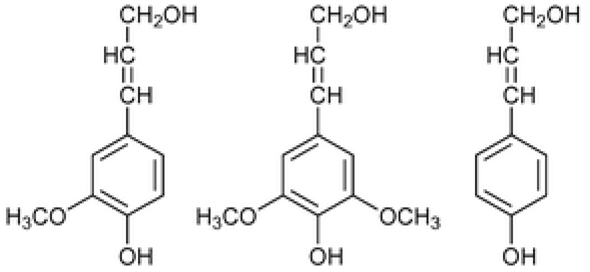
FIGURE 5. Major hemicellulose a) galactoglucomannan in softwood and b) glucuronoxylan in hardwood.

Other polysaccharides

Other minor polysaccharides of wood are starch and pectins. Glucose units bound together with alpha acetal bonds making highly branched or twisted helical chains, compose starch which due to this structure is more prone to cleavage than cellulose. Starch is an important carbohydrate as an energy store, but it is a non-chemically bound component of the wood, thus giving no structural support. Hence, it is removed during extraction, see Chapter 2.2.3 for further explanation. Instead, pectins are structural polysaccharides with repeating galacturonic acid units as a backbone and these compounds are found higher concentrations in inner bark than in the stem. (Rowell 2005, 39-43)

2.2.2 Lignin

Lignin is a highly complex natural polymer which predominant building blocks are phenylpropane units. Structurally, in the cell wall lignin works as a strengthening component, crosslinking the different polysaccharides together. Lignin consists of several substructures and their proportions and linkage modes are hard to estimate with 100% accuracy. (Amarasekara 2014, 137-143) Lignin lacks the single repeating unit like in cellulose and instead consists of complex arrangement of mainly aromatic units. The ratio of these units differ for example between wood types; hardwood contains mainly sinapyl and coniferyl alcohol and softwood almost exclusively of coniferyl alcohol units. Paracoumaryl alcohol is mainly present in grasses. The structure of these units can be seen in Figure 6. (Ek, Gellerstedt & Henriksson 2009, 121-124) The isolation of lignin from the network of cellulose and hemicellulose can be challenging due to the lignin-carbohydrate complexes which are resistant to hydrolysis. (Rowell 2005, 43-45)



Coniferyl alcohol Sinapyl alcohol Paracoumaryl alcohol

FIGURE 6. The most common units of lignin.

2.2.3 Extractive, nitrogenous, inorganic and other elements

The non-structural components of biomass which are not bound to the structure of the material, so called extractives, are removed prior to the analysis of lignocellulosic components. The types and amounts of extractives in woody biomass vary depending on the type of species, season, region of plant and geographical location. Major categories are fats, waxes, terpenoids, polyphenols, monosaccharides and other inorganics. (Cheng, 201, 35-37) Together they present relatively small portion of the wood, normally below 5 % but might be of higher concentration in certain parts of the tree, especially in bark. Even though these constituents have value as a further processed product, for example as adhesives (from tannin, a common phenolic compound) or as cancer medication (eg. Taxol from the bark of Pacific yew) in this work the primary aim was the removal, not the quantification of them. (Christopher 2013, 40-44)

The importance of this removal step is due to the errors these components pose on the final results. If the extractives are not fully removed, part of them can condense to acid insoluble components, resulting in falsely high lignin values. Furthermore, some carbohydrates outside the insoluble cell wall when not extracted might be incorrectly classified as cellulose and hemicellulose, leading to wrong results. Also incomplete hydrolysis may occur when the penetration of the sulphuric acid to the sample is inhibited by hydrophobic extractives. Different extractives are soluble in different solvents, thus the use of water and other organic solvents such as ethanol are utilized in the extraction. The decision on which extraction should be used is dependent on the biomass type. In case woody feedstock the water extractable material is little, thus in this work the samples were subjected to ethanol extraction only. (Sluiter, Ruiz, Scarlata, Sluiter & Templeton 2005)

According to the Laboratory Analytical Procedure (LAP), determined by the National Renewable Energy Laboratory (NREL), the compositional protein analysis is of interest in case of herbaceous feedstocks in order to minimize the interference on subsequent analysis steps. Because of lower protein and nitrogenous compounds concentration in woody biomass and due to the fact that the extraction process already removes portion of it, quantification of protein in this work was not applied. (Sluiter & Sluiter 2011)

The ash content of wood is usually referred to as its inorganic content, since it is determined by incineration at 575 °C, the residue forming of substances with generally no energy value. This material contains different elements, 80% of the ash in wood consisting of calcium, magnesium and potassium. (Rowell 2005, 50) High ash-content of the feedstock can create problems in the energy conversion facilities, since the fly ash and particulates stick on the surface of the boilers, decreasing the heat transfer efficiency and what is more, especially the high chlorine composition of ash can corrode the boilers. (Biedermann & Obernberger 2005) Wood feedstocks without bark have typically ash contents below 1%, which is low compared to herbaceous biomass types that have reported ash values of 2-10%. (Clarke & Preto 2015)

The elemental analysis for the major elements carbon, oxygen, hydrogen, nitrogen and sulphur are calculated through elemental analysis. Their importance is for different aspects; higher carbon and hydrogen values indicate higher heating value of the sample, whereas higher nitrogen, sulphur or chlorine values could lead to higher particulate matter (PM) emission during combustion or result in corrosion in the boilers. (Clarke & Preto 2015)

2.2.4 Moisture

Moisture content of the biomass is a crucial parameter for combustion purposes; the energy needed to drive off the water content will reduce the overall system efficiency and possibly lower the combustion temperature below optimum leading to incomplete combustion (Loo & Koppejan 2008, 9-11). However, in biochemical conversion, such as anaerobic digestion, high moisture content of the biomass is a positive parameter in the process, since it relies on micro-organisms requiring a moist environment (Cheng, 2010). Moisture content is also an important factor in the analytical procedures for carbohydrate quantification where the calculations are based on the dry mass of the sample. (Sluiter & Sluiter 2011) The moisture content of wood is around 40-50 %. (Stenius 2000, 28)

2.3 Near-infrared spectroscopy (NIRS) principles

Spectroscopy investigates the interaction between matter and the photons of electromagnetic radiation (EMR) in forms of both electric and magnetic waves. The range of EMR is called the electromagnetic spectrum and it is divided into different regions based on their frequencies and wavelengths. Wavelength (λ , in nanometer (nm)) is the distance between the peaks of sequential waves, whereas frequency (v in s^{-1}) is the number of waves per second (noted as hertz, Hz). Frequency is inversely proportional to wavelength as can be seen in the following Formula 1:

$$v = \frac{c}{\lambda} \tag{1}$$

Where c is the speed of light in a vacuum (3 x 10^8 m/s). The relationship between the energy carried by the photons of radiation (E) and radiation frequency (v) is following:

$$E = hv = \frac{hc}{\lambda} \tag{2}$$

Where h is the Planck's constant (6.626 x 10^{-34} Joule seconds). In EMR spectrum towards longer wavelengths than visible light is the near-infrared (NIR) region, which covers the wavelengths from about 800 to 2500 nm. Thus, NIR has longer wavelengths and lower frequencies than that of visible light. When different regions of EMR interact with matter, they create various effects in its molecules, depending on the quantity of energy. NIR radiation results in molecular vibrations and rotations. The absorption of the NIR radiation only occurs when the photons emitted resonate with the characteristic vibrations of the chemical bonds of the sample, thus having similar vibrational frequencies. (Kaur 2009, 1-6) The important characteristic requirement here is the molecular dipole (unequal distribution of electric charge) moment change during vibration. Each type of possible vibration within a molecule is called vibrational mode and as the number of atoms within a molecule increase, so does the total number of possible vibrational modes. For this reason the complex molecules of wood (see Chapter 2.2) result in large number of vibrational modes and many peaks in their NIR spectra. (Anderson, Bendell & Groundwater 2004, 26-27) Example of several Eucalyptus NIR spectra are presented in the Figure 7 below adapted from the Forest Quality Pty Ltd, which utilizes this information in their commercial service for predicting Kraft pulp yield for Australian eucalypt growers.

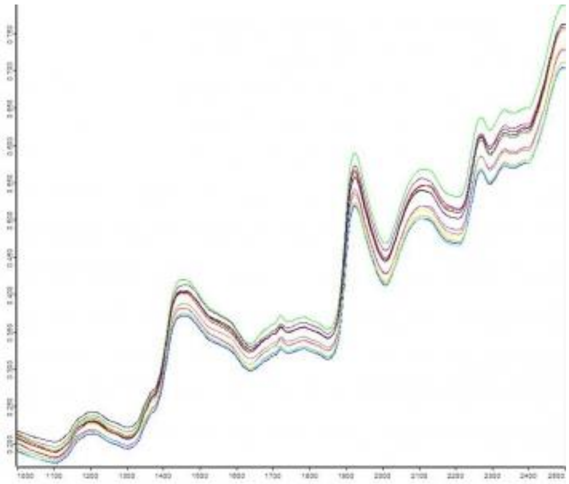


FIGURE 7. Complex NIR (1100-2500 nm) spectra of ground (fine powder), air-dried eucalyptus samples. (Downes 2005)

Especially vibrations from bonds like O–H, C–H and N–H (which are common bonds in organic compounds making especially NIR suitable for detecting these) result in the multiple overtone peaks and combination bands in the NIR spectra (Niederberger *et al.* 2015). This indicates the anharmonic characteristic of these vibrations. Anharmonicity means that in addition to fundamental transitions, which are transitions only occurring between adjacent energy levels (n=0 → n=1), also overtones can occur. First overtone is the transition from the ground state (at which most molecules are at room temperature) to the second energy level (n=0 → n=2), the second from ground state to third level (n=0 → n=3) and so on . (Kaur 2009, 1-6) This is illustrated in the following Figure 8.

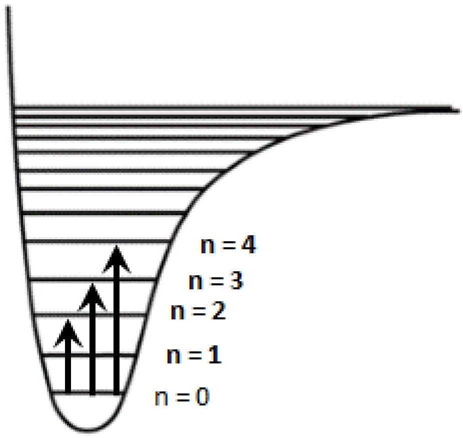


FIGURE 8. Overtone transitions.

The overtone and combination oscillations result in the fact that single peaks cannot be used for identification and a certain molecule has to be recognized from several peaks from its spectrum. (Niederberger *et al.* 2015)

2.4 Background on the use of NIRS for biomass analysis

The earliest studies in the application of NIRS on the analysis of lignocellulosic components of dry and ground biomass for biorefining was conducted by Sanderson et al. (1996) This study had contributions from National Renewable Energy Laboratory (NREL) and since the publication of this paper, NREL have been actively contributing in NIRS biomass analysis (Hayes 2011, 237), as can be noted from the various references to their Laboratory Analytical Procedures (LAPs) in this work. The idea behind the division of data to global and local datasets for predicting lignocellulosic constituents by

NIRS is also not a new concept. Mentions of species-specific vs multi-species calibrations in biomass NIRS analysis are mentioned in the often referred book ‘Handbook of Near-Infrared Analysis’ by Burns and Ciurczak (2008) and by Hodge and Woodbridge (2010) in the journal of near-infrared spectroscopy. However in the latter, these models are referred to as global and species-specific, instead of referring to local, which seems to be the current approach.

Hence, the use of word ‘local’ when describing a calibration model in this work is not to be mixed with the patented LOCAL calibration method developed by J.S. Shenk or local weighted regression (LWR) method. These methods differ to what have been used in this work and presented in this work’s theory (see Chapter 2.5.); LOCAL and LWR methods choose from a large database a small data set tailored to predict a certain unknown sample, thus this method could be referred to as “one at a time analysis”. This small data set is selected by comparing the NIR spectrum of the unknown sample to similar spectrums of samples from the large database. Thus, many PLS models (see Chapter 2.5.5) with multiple factors are created based on these unknown samples on small data sets. (Burns & Ciurczak 2008, 372; Shenk & Westerhaus 1997; Cabassi *et al.* 2005) Reader is advised to notice this difference in meaning of the local based on this method and the meaning of local used in this work: to address the wood-specific models.

2.5 NIRS calibration methodologies

Any problem solving starts with defining the problem. The issue in NIRS calibration is to relate the broad and complex spectral data of the sample to the concentrations of analytes of interest. In combination with representable sample sets, robust analytical chemistry methods, spectral pretreatment, multivariate data analysis (simultaneous analysis of more than one variable) and defining the relevant information to the problem from the results it is possible to receive chemical and physical information of the NIR data. Accurate experiment design gives the analyser the ability to correctly interpret the outcomes and discuss the prediction performance capability of the model. Step by step methods for model development is outlined in the next subchapter. (Burns & Ciurczak 2008, 123-126)

2.5.1 Selection of samples and data set subdivision

When selecting a set of samples used in NIRS calibration, it is highly important that the range of components to be predicted are well presented in the chosen samples. So that both high and low concentrations of components will be accurately predicted, evenly distributed range of data is required. To achieve this, sufficient number of samples with similar but not identical characteristics are chosen. In order to build model that has a high likelihood to perform well, the reference analysis (analysis method used to get the true values) results should be of acceptable reproducibility in order to correctly relate this data to the NIR spectra. (Burns & Ciurzack 2008, 132-140)

In the model development the samples are initially divided into two sets, calibration/teaching set and validation/training set. The calibration set composes of samples with robust reference results and corresponding NIR spectra. This relation between them is used to produce a regression equation. In order to evaluate the model performance on samples outside this calibration set, the model is 'validated' or 'tested' on a validation set of samples. This set composes of samples with corresponding NIR spectra but no information of the 'true' values by the reference method. When the concentration values for the validation samples are then predicted by the formerly developed equation and finally also analysed by the same reference method as for the calibration set, the variance between the values of these two methods can be then used to evaluate the predictive accuracy of the model. In this way testing the precision of the model on external sample set gives a realistic estimate of the model performance. From the sources found by the Author, the basic division of calib/valid sets seemed to be approximately 3:1. (Burns & Ciurzack 2008, 132-140; Esbensen, Dominique, Westad, & Houmoller 2002, 118) Principles for dividing samples to different sets and their subsequent analytical procedures can be seen as a flowchart in Figure 9.

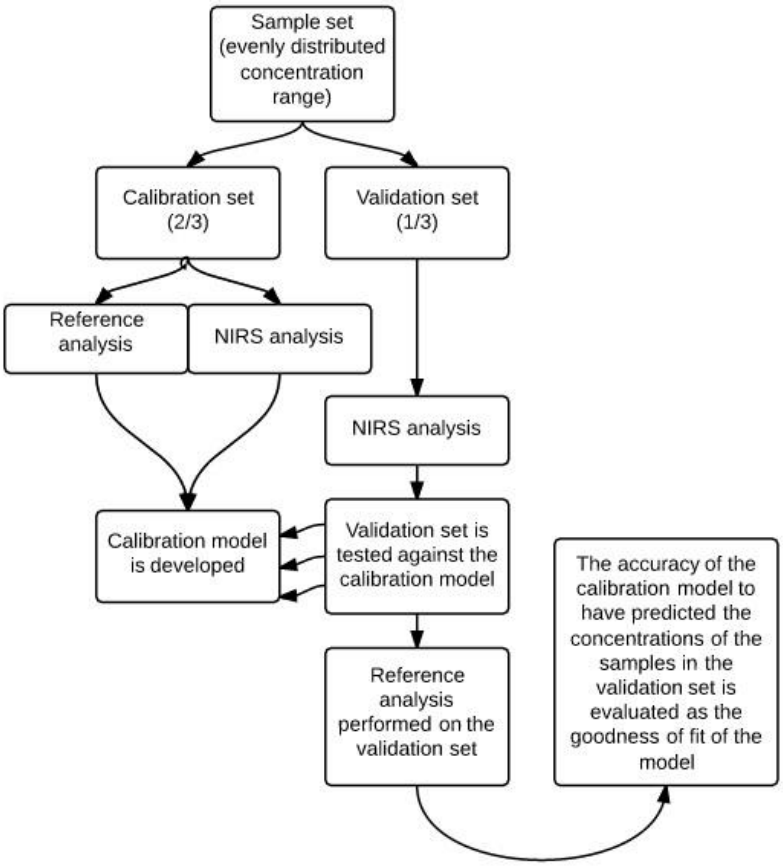

Sample set (evenly distributed concentration range)
Calibration set (2/3)
Validation set (1/3)
Reference analysis
NIRS analysis
NIRS analysis
Calibration model is developed
Validation set is tested against the calibration model
Reference analysis performed on the validation set
The accuracy of the calibration model to have predicted the concentrations of the samples in the validation set is evaluated as the goodness of fit of the model

FIGURE 9. Flowchart of NIRS calibration steps.

Outlier is any sample which is distributed unreasonably from the mean average. This can be due to instrument or laboratory analysis error, in which case the analysis should be repeated if possible. Since the final model statistics will be influenced if the sample with major difference is not rejected from calibration set and because even with robust analytical methods these outliers can still occur, a method for their recognition and reasons for their rejection should be carefully evaluated and presented in model development. Different tests for their recognition are available. International Organization for Standardization (ISO) recommended test is Grubb's test, which compares the deviation of each value from the sample average with the standard deviation of the sample. If this value then exceeds the critical value provided by confidence limit 95% and a specific Grubb's critical value table, the value is considered outlier. This test assumes normally distributed data set. However subjective selection on rejecting the value or not from the model development should be considered based on multivariate expertise. (Miller & Miller 2005, 51)

2.5.2 Spectral pretreatments

Mathematical manipulation of the broad and complex spectral data is done prior to analysis to remove or reduce any unwanted sources of variation, also referred to as ‘noise’ in this case meaning any aspect of the spectra not related to the analytical data. These could include various effects from the instrument (eg. temperature variation) or sample properties (eg. unequal particle size distribution) which can lead to random noise and path length variations and light scattering which can make it difficult to interpret the spectral data. There can also be dominant low frequency sources of variation in the baseline of the spectra, which are not related to the chemical composition of the analyte and may lead to misreading of the data. Prior to calibration the impact of these interferences is either standardized or excluded mathematically by chemometric means. (Beebe, Pell & Seasholtz, 1998) Pretreatments can be done in conjunction with data acquisition where data is reduced by removing or ‘filtering’ some of the noise. However most often the spectral enhancements are done after spectra is collected by different smoothing methods. (Gemperline 2006, 380) Smoothing e.g by decreasing the amplitude of noise-containing frequencies and so called Savitzky–Golay (SG) derivatives (Especially 2nd order, e.g. to resolve peak overlap) are two often used pretreatment methods. Derivatives also remove baseline offsets and can narrow and sharpen the peaks (e.g. making smaller analyte peaks more evident) in the spectra. (Beebe, Pell & Seasholtz, 1998) An example can be seen in the Figure 10 below. Typically a standard set of these pretreatments is included in the software. (CAMO, 2015) Some other pretreatments to remove multiplicative error, nonlinearities and effects from particle size, MSC (multiplicative scatter correction), SNV (standard normal variate), or SNVDT (standard normal variate and detrend) could be used. (Burns & Ciurczak, 2008)

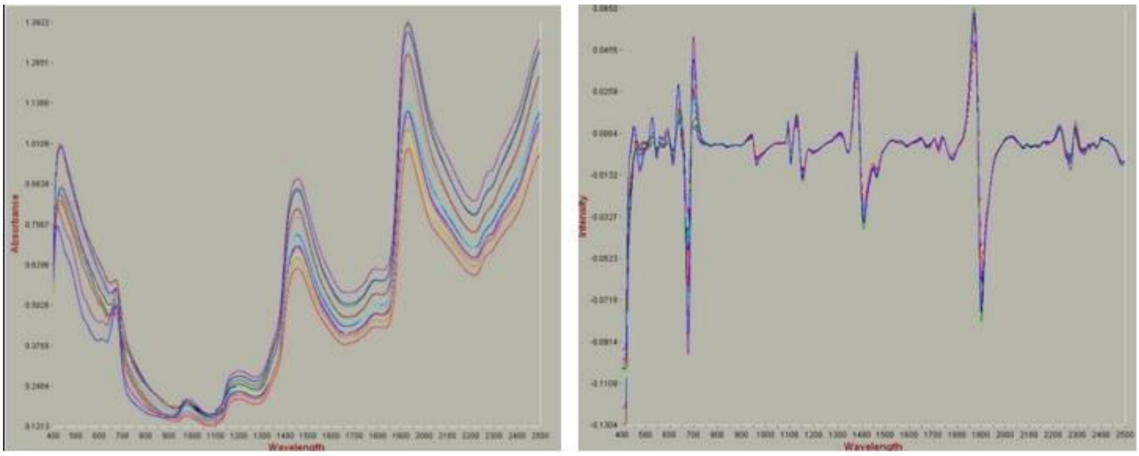


FIGURE 10. Raw spectra and spectra treated with 2nd-order derivative. (Hayes 2010)

2.5.3 Calibration and regression principles

Both molecular absorbance and reflective properties of the sample are variables that result in the received energy that the spectrometer detects. To relate the measured absorbance of NIR by molecules of the sample at a certain wavelength with the concentration of the analyte of interest, Bouguer-Lambert-Beer or commonly Beer's law is used (Burns & Ciurczak 2008, 123-127):

$$A = log_{10}\frac{I_0}{I} = \ Mcd \tag{3}$$

where
A = absorbance
I_0 = intensity of radiation entering the sample
I = intensity of radiation transmitted by the sample
M = molar absorption coefficient at given wavelength (M^{-1} cm^{-1})
c = molar concentration of absorber (mol x dm^{-3})
d = sample path length (cm)

As the transmittance (T) of the sample is the ratio I/I_0 and in NIR reflectance (R) is considered to be relative to transmittance, absorbance (A) can also be noted as:

$$A = log_{10}\frac{1}{R} \tag{4}$$

So ideally, if Beer's law would apply perfectly, absorption at certain wavelength for any substance would be proportional to concentration, thus under standard conditions the only variables would be absorbance and concentration and the relationship could be presented as two dimensional scatter plot. However, to add up the multi-wavelength regression in case of NIR spectra and interferences and to better model the percent concentration (Y) as a function of absorbance (A) response at specific wavelength, the following linear multiple regression which is inverse of the equation (3) could be formed (Burns & Ciurczak 2008, 123-127):

$$Y = B_0 + B_i(log_{10}\frac{1}{R_i})_N + E \tag{5}$$

Where

Y = percent concentration of absorber

B_0 = intercept from regression

B_1= regression coefficient

i = index of the wavelength used and its corresponding reflectance R_i

N = total number of wavelengths used in regression

E = random error

The regression coefficients are found by least squares principle: a regression line is of best fit when the sum of squared distance between predicted and analyzed values (residuals) is minimized. However, this is still a univariate approach where all the variance (how far apart set of values are spread) is consequential of single variable. This excludes the common case in NIR - absorbance at certain wavelength can contribute to different components. The equation also includes baseline errors (e.g. instrument error not directly concentration related) and does not regard overlapping bands. The one-to-one selectivity problem can be solved by performing multivariate analysis, simultaneous analysis of more than one variable. (Griffiths & Haseth 2007, 207-214)

2.5.4 Multivariate calibration

Due to the NIRS instrument noise and drift, laboratory errors, physical variations within sample, scattering at certain wavelengths and other deviations from Beer’s law e.g. due to non-linearity of detection system occur, the ideal univariate equation (3) does not apply. The multivariate approach increases the dimensionality of the problem, due to the amount of multiple variables, for example when two wavelengths and concentration (3 variables) are considered. Thus, instead the equation (3) should be better represented in multiple linear regression (MLR) to meet the requirements for NIR spectra analysis. To simplify the multivariate case, matrix form of this model is presented:

$$\boldsymbol{Y} = \boldsymbol{XB} + \boldsymbol{E} \tag{6}$$

Where ***Y*** is a vector of component concentrations in each experiment, ***X*** is a *sample*-by-*variable* (wavelength) matrix for all observations, ***B*** is a vector of coefficients and ***E*** is a vector of errors for each prediction. (Hayes 2011; Naes, Isaksson, Fearn & Davies 2007) Thus the relations can be written open as:

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_N \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{N1} & x_{N2} & \ldots & x_{Nk} \end{bmatrix} \begin{bmatrix} B_0 \\ B_1 \\ \vdots \\ B_k \end{bmatrix} + \begin{bmatrix} E_0 \\ E_1 \\ \vdots \\ E_k \end{bmatrix} \tag{7}$$

Where N is the number of samples and k number of wavelengths used. Hence, the principle is the same than in multiregression line (Equation 5), but including multiple variables. As for the complexity of the spectral data with large number of wavelengths, corresponding absorbances and relation to values by the reference method, visual interpretation of the regressions in the matrix becomes impossible and mathematical methods for correlating data in multidimensional space has to be applied. This is most commonly done while running data through applicable software. (Burns & Ciurczak 2008, 123-127)

However, this model cannot still overcome the problem of intercorrelation between independent variables (the absorbances at different wavelengths) phenomenon also referred to as multicollinearity. Particularly the wavelengths close to each other in the NIR spectra can have high correlations. This occurs when any column in the observed data can be expressed as a linear combination of other columns or there are more columns than rows, like often is the case with fewer samples than wavelengths in the recorded spectra. (Burns & Ciurczak 2008, 94;214-215) Hence, to avoid this multicollinearity problem there is a need for finding uncorrelated variables, which can be developed by principal components (PCs) as explained in the next subchapter.

2.5.5 Principal component analysis (PCA) and regression analytics

Principal component analysis (PCA) refers to a method used to interpret complex data by building linear multivariate models. (Gemperline 2006, 70) PCA is used to reduce the dimensionality of the data, thus eliminating variations caused by only noise while keeping as much important variation in the data as possible. In order to find the most essential components which can present the characteristics of the data without losing any valuable information, new set of variables, uncorrelated (orthogonal) principal components (PCs) are created. The principle is that the components are ordered in decreasing order according to the amount of variance they contain. (Jolliffe 2002) This can be seen in the graphical illustration in Figure 11.

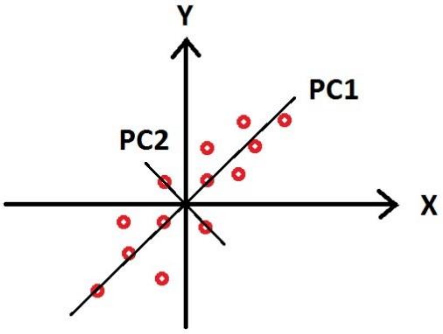


FIGURE 11. Plot on two variables y and x and their average PCs.

Even though it is not possible to visually show the higher-order components in two-dimensional plot, the principle stays the same. Thus, PC3 is simultaneously orthogonal to both PC1 and PC2 and includes the third largest variance. The maximum number of components is either number of objects minus one or number of variables, whichever is smaller. Though, when introducing more PCs to the system and especially in the higher-order principal components the variance gets smaller and smaller, consisting of mostly random error and noice. Thus, too many components or factors can cause overfitting of the model, when adding new terms does not give valuable information of the studied relationship. The basis for the analyzer to decide how many components are needed to sufficient data presentation can be obtained by standard statistical methods. (Esbensen *et al.* 2002, 27-36)

Principal component regression (PCR) and partial least squares regression (PLSR)

As such, PCA and multiple linear regression set the basis of regression analysis technique called principal component regression (PCR) which is the application perspective of PCA used for calibration. Even though PCR is an efficient tool against collinear data, it has few downsides. There is no guarantee that the first PC contains just only the correlated information to the variable of interest, there can be also other kinds of variations in the PC1. Since some of this irrelevant variation can dominate initial PCs, it can happen that some relevant variance is lost in the higher order PCs. Partial least squares (PLS) regression as utilized in this work is a procedure similar to PCR, with the differentiation that the variance in $\boldsymbol{Y}$ (represents the reference values for the samples) and covariance between $\boldsymbol{X}$ (represents the spectra of the samples in the calibration set) and $\boldsymbol{Y}$ is used to build the components. This generally results to fewer components or 'factors' than in PCR. (Esbensen *et al.* 2002, 135-139) PLS1 means that only one constituent (e.g. concentration of glucose) is used, thus PLS2 is used for double constituents. (Hayes 2010)

Determination of the optimum number of factors

An example of standard method for finding the suitable number of factors include predicted residual sum of squares (PRESS) as used in this work. This procedure first leaves one sample outside the calibration set while using one factor and develops a calibration with all the rest of the samples. Then the left out sample is predicted using this formula and the residuals recorded. Then this technique is repeated leaving each sample outside the set and finally the sum of squares of residuals noted. After this one by one new factors are added and process repeated until maximum amount of factors is reached. The best model for calibration is considered as least number of factors and sum of squares for residuals. From the xy-scatter plot with the PRESS value in x-axis and number of factors in y-axis it can be then seen, which minimum PRESS value corresponds to the smallest number of factors in the calibration model. (Burns & Ciurczak 2008, 148) In this work Haaland and Thomas (1988) criterion was used, which compares the amount of factors that gives the minimum PRESS via F-test, in the following manner:

$$F(m) = \frac{PRESS(m)}{PRESS(m^*)} \tag{8}$$

Where

m^*= factors associated with the model that gives the minimum PRESS

m = all models with fewer factors ($m < m^*$)

The purpose of comparing the relation between these models is to find the optimum model with lowest number of factors so that PRESS for this model is not majorly greater than PRESS for the model with m^* factors. As for the optimum number of factors, the smallest m is chosen so that $F(m)< F_{\mathrm{N,N,\alpha}}$ where N is the number of samples and α=0.25. (Haaland & Thomas 1988, 1200-1201)

2.5.6 Calibration and regression statistics

To compare the regressions between reference analysis and model-predicted values, thus to determine the capability of the NIRS models to predict constituents of interest, variety of different statistics are used. Most important of these are the ones for the validation set. (Hayes, 2011) Unless otherwise noted the equations are from Burns & Ciurzcak. (Burns & Ciurczak 2008, 140-148)

Dispersion of data points

The total variation (SS_T) in the calculated (y) values is calculated as sum of deviations of the true value from the corresponding mean ($\bar{y}$). The values are squared to reach only positive values and to prevent negative and positive values from cancelling each other. Thus, when N is the number of samples, the total sum of squares is calculated as follows:

$$SS_T = \sum_{i=1}^{N} (y_i - \bar{y})^2 \tag{9}$$

The total sum of squares (SS_T) equals the sum of squares of regression (SS_R), which is the sum of squares of deviations of the predicted value ($\hat{y}$) from the mean value of responding variable, and sum of squares for residuals (SS_D), the total error between the model and real data, thus the sum of the individual deviations (residuals) between calculated (y) and predicted ($\hat{y}$) values.

$$SS_R = \sum_{i=1}^{N} (\hat{y}_i - \bar{y})^2 \tag{10}$$

$$SS_D = \sum_{i=1}^{N} (y_i - \hat{y}_i)^2 \tag{11}$$

Correlation: Coefficient of determination (R^2) and correlation coefficient (r)

To compare the correlation between two dispersed data points and their linear depencencies, correlation, a unitless measure is used. To describe how well the data fits the line of regression, coefficient of determination (R^2), the ratio between SS_R/SS_T is used. It indicates the percentage of explained variation to total variation. For example if

R^2= 0.90, then 90% of the total variation in the variable y can be explained by the linear relationship between $\hat{y}$ and y, while 10% of the total variation of y cannot be explained.

$$R^2 = \frac{SS_R}{SS_T} \tag{12}$$

However, this is the unadjusted correlation, which is expected to increase whenever new variable is added to the model. For this reason it can give an overly positive estimation of the model performance only based on the fact that it has more terms. When adding variables to adjusted $\bar{R}^2$ data, actually correlative ones will increase the value and variables without strong correlation decrease it. In case of multiple regression statistics, this gives better indication of the goodness of the model. (Frost 2013)

$$\bar{R}^2 = 1 - \frac{SS_D/(N-K-1)}{SS_T\ /\ (\mathrm{N}-1)} \tag{13}$$

Where N is the number of samples used and K the number of wavelengths used in the model. (Burns & Ciurczak 2008, 140-148) For a simple linear regression model the correlation coefficient is used, where the sign of r depends on the negative/positive slope of the linear line, and can be calculated as:

$$r = \pm\sqrt{R^2} \tag{14}$$

Root mean square error (RMSE)

Other important statistic for accuracy comparison is the root mean square error. It estimates how much error there is between the true (y) and predicted $(\hat{y})$ values by giving the average of the sum of the deviations as seen in the equations below.

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{n}(y_i - \hat{y_i})^2} \tag{15}$$

Based on this equation, root mean square error can be applied on three subsets: calibration (RMSEC), cross-validation (RMSECV) or prediction (RMSEP), thus the RMSE equation differing by y and $\hat{y}$ values. The RMSEC defines the deviation between the calibration values and the curve used to fit the calibration data. As only such, this parameter gives overly optimistic values of the model performance. RMSECV is determined by removing each sample (called leave-one-out cross-validation) or specific amount of samples as subsets (called holdout method or K-fold cross-validation) and model built from the remaining samples. (Performed identically to PRESS, see last part of the Chapter 2.5.5.) The properties of the removed sample/s are then estimated by the model and this is done to each sample or subset. RMSECV is the average value of sum of these differences. For future samples outside the model, thus the most important performance parameter is the RMSE of prediction, RMSEP. The estimated $(\hat{y})$ value is determined by using the calibration model on an external data set called validation set. This error value is presented as percentage. (Burns & Ciurczak 2008, 220-221)

Standard error of prediction (SEP)

Standard deviation of predicted residuals or SEP, measures difference between repeated measurements (precision) compared to RMSEP which measures the accuracy (difference between true and predicted value). Bias is the average difference between predicted and true value in validation set. Thus, if bias is small, these two values are similar.

$$SEP^2 \approx RMSEP^2 - BIAS^2 \qquad (16)$$

Ratio of standard error of performance to standard deviation (RPD)

$$RPD = \frac{sd}{SEP} \qquad (17)$$

This is dimensionless statistics and can be compared between different models, higher values indicating increase in accuracy. All these statistics were used to assess the precision of prediction. (Williams & Norris, 1987)

When RPD is of following values the calibration is:
≥ 2.5 – suitable for screening and breeding programs
≥ 5 – acceptable for quality control
≥ 8 – good for process control, development and applied research

Range error ratio (RER)

RER equals to the range in compositional values (max-min value) divided by SEP.

$$RPD = \frac{range}{SEP} \tag{18}$$

The RPD/RER ratio is often 4/5:1 and depends on the distribution of samples in the validation test. When there are no outliers and data is evenly distributed, RER can work as good quality indicator of the model, even better than RPD. (Fearn 2000)

When RER is of following values the calibration is:
≥ 4 – acceptable for sample screening
≥ 10 – acceptable for quality control
≥ 15 – good for quantification

Test for significant differences between two models: SEP values

When comparing calibration models and evaluating which would be better in prediction, there is a need to consider the errors between the models and if they are actually significant. Otherwise it could just happen that the results would be reversed when using another validation set. A test found from Naes *et al.* describes the method for comparing significant differencies between two SEP values of two models. First the coefficient of correlation (r, see equation 14) between the prediction residuals from these two sets is calculated and noted as r in the following equation (Naes *et al.* 2007, 166-170):

$$\kappa = 1 + \frac{2(1 - r^2)t^2{}_{(N_p-2),0.025}}{N_p - 2} \tag{19}$$

Where
N_p = number of samples in prediction set
$t^2{}_{(N_p-2),0.025}$= the upper 2.5% percentile of a t-distribution with $N_p - 2$ degrees of freedom, t=2 is a good approximation for most purposes
Then calculate:

$$\mathrm{L} = \sqrt{(\kappa + \sqrt{(\kappa^2 - 1)}} \tag{20}$$

And:

$$\frac{SEP_1}{SEP_2} x \ \frac{1}{L} \ and \ \frac{SEP_1}{SEP_2} x \ L \tag{21}$$

These two equations then give the lower and upper limits of a 95% confidence interval for the ratio of the true standard deviations. If the range of these lower and upper limits include 1, the SEPs are not significantly different at the 5 % level. Thus when the number of samples in prediction set is small and r is not close to 1, the range of the limits can be quite wide before there is significant difference.

3 METHODS

The methods for selecting the samples, the wet-chemical analytical procedures behind the reference results and steps for building the predictive NIR models are further defined in this chapter. The wet chemistry compositional analysis methods have roots as far as in the turn of 20th century. Modern NIR technologies as presented here rely heavily on computer and have advanced rapidly since 1970s. The current developed and modified method the chemical laboratory analysis is written in separate Laboratory Analytical Procedure (LAPs) by National Renewable Energy Laboratory (NREL), which is also adapted as a standard method by The American Society for Testing and Materials (ASTM). (Sluiter, Ruiz, Scarlata, Sluiter & Templeton 2010) This so-called two-stage acid hydrolysis for the determination of structural carbohydrates and lignin is also used in this work. In a visual interpretation below (Figure 12), the steps of both the reference analysis (wet-chemical) and NIRS analysis are shown.

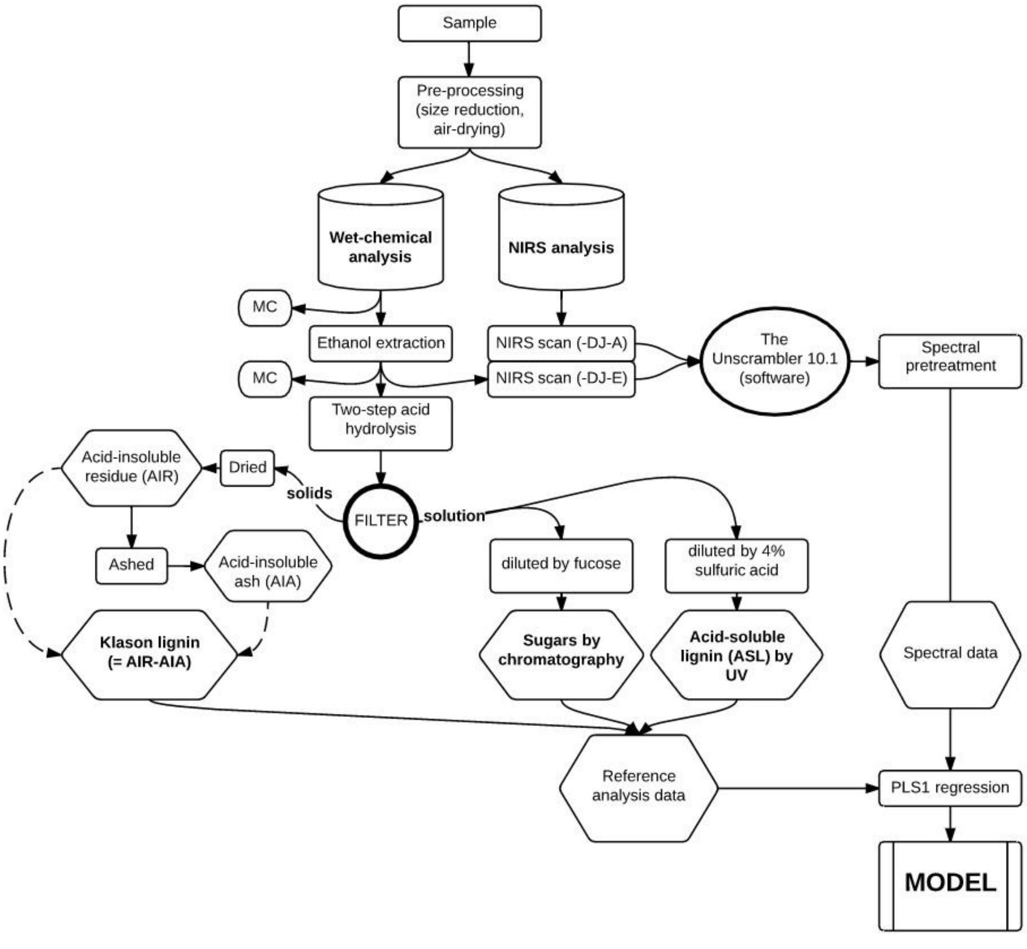


FIGURE 12. Flow chart of analyses.

3.1 Sample selection and pre-processing

For the global data the calibration models were developed using a wide variety of biomass types; e.g. agricultural residues, energy crops, industrial and municipal wastes. (Complete list of types in Appendix II) Also a large number of samples, on average 750, including the same 110 wood samples used in the wood-specific model, were included in the global model development. The 110 wood samples that were used in the development of the wood-specific calibration model were different species presenting softwoods; pine, spruce, and hardwoods; paulownia, ash, alder, birch, poplar and eucalyptus. (See Appendix I) Since the species, region, age and part of the tree influence in its chemical composition, to reach a representative data set, the collection of samples covering the whole concentration range was essential. The external validation set consisted of 20 wood samples. (See Appendix III)

For example when considering the parts of the tree, bark and foliage are in general higher in ash and extractives than the stem thus having lower polysaccharide content. Bark has also higher lignin content and foliage lower than the rest of the plant stem. (Häkkilä 1989, 148-159) Foliage was not included due to greater differences in composition - higher in extractives and lower in lignin than other parts of the tree (Sariyildiz & Anderson, 2005). It was approximated that these differences would alter the even distribution of the calibration data, thus skew the equation curve, considering the low amount (110) of the samples. However, even though higher ash, lignin and extractive content, bark was included due to the many samples analysed, thus being a consistent variable in the analytical range. Also parts such as tops, branches, stems and wood were included.

Samples excluded from the wood-specific model development were the same consistent outliers in the global model (also excluded), probably indicating a laboratory error in labelling, during the analytical procedures or in NIR spectra collection. Excluded samples can be seen in Appendix I for each constituent. To assure that no outliers remained, Grubb’s test as defined in Chapter 2.5.1 was used for the wood-specific calibration set. Reader is advised to get further acquainted with this test by the Grubb’s paper. Notice that in this paper the critical values for alpha=0.025 for one sided test are analogous to alpha=0.05 for two sided test, the latter used in this work. (Grubbs 1969, 4)

Particle size reduction (<0.85mm, samples referred to as ‘DJ’) for wet-chemical analysis was taken into consideration for the accuracy and efficiency of the subsequent analysis procedures. Air-dried, grinded samples (<0.85mm) were used in wet-chemical and NIR analysis.

3.2 Wet-chemical analysis

In order to build a predictive NIR model, primary compositional analysis of the samples and a robust reference method is required, since the regression between these values and the NIR spectra is the basis for the calibration model. This subchapter specifies the different steps of this reference analysis, referred to as wet-chemical analysis due to the nature of the procedures. To complete a batch of samples, approximately 2 weeks’ time altogether is needed for this whole analysis.

The results were considered to meet the precision criteria, if the standard deviation of the constituent value for each of the duplicates (SDD) did not exceed certain limits, which can be seen in the following Table 2.

TABLE 2. Standard deviation of duplicates (SDD) for different constituents. (Hayes 2011)

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.

3.2.1 Moisture and ash

The moisture content by convection oven method described in LAP was used (Sluiter, Hames, Hyman, Payne, Ruiz, Scarlata, Sluiter, Templeton & Wolfe 2008). Moisture content was determined as the mass loss of the sample placed in an oven at 105 °C until constant weight was reached. Samples (0.2-0.5 grams) used in this work were dried overnight in duplicates in Memmert UF 260 oven and weighed after they were cooled down to room temperature in a desiccator. Subsequently the sample with known moisture content was placed in Nabertherm L-240H1SN muffle furnace for ashing, which maintained the temperature at 575 °C for 3 hours. Ash content was expressed as percentage of residue after this dry oxidation, like determined in LAP ‘Determination of Ash in Biomass’ (Sluiter, Hames, Ruiz, Scarlata, Sluiter & Templeton 2005).

3.2.2 Extractives removal

The extractives components (defined in Chapter 2.2.3) were removed according to ‘Determination of Extractives in Biomass’ (LAP) by utilizing Dionex Accelerated Solvent Extractor (ASE) 200 with ethanol as solvent (Sluiter *et al.* 2005). The method performed used 11ml ASE cells with 1-6 g sample, 95% ethanol, 1500 PSI pressure, 100 °C temperature, heating time of 5min and static cycle time of 7 min. Each sample went through three static cycles, total flush volume of 150% and purging of 120 sec. After the extraction was finished, the solid residue in the ASE cell was emptied to a container and left to air dry for 2 days. The standard method (see Chapter 3.2.1) was then used to determine the moisture content. Amount of extractives in the sample was determined as the loss in dry matter during the extraction. All samples were run in duplicates.

3.2.3 Two-step acid hydrolysis

Procedure similar to what is described by NREL (Sluiter, Hames, Ruiz, Scarlata, Sluiter, Templeton & Crocker, 2008) was employed on the extracted sample. First step included adding approximately 300mg of the sample to a pressure tube, followed by 3mL of 72% sulphuric acid (H_2SO_4) by an automatic titrator. This concentrated acid can hydrolyse even the crystalline region of cellulose. The sample and the acid were then thoroughly mixed by using a glass rod and the pressure tube was placed into a water bath, where it stayed for a period of 1h at constant temperature. (30 °C) Care was taken that no sample stayed adherent to the sides by proper mixing every 10 min. After completion of the 60-minute hydrolysis, 84mL of deionized water was added to the pressure tube in order to achieve 4 % acid concentration. The tube was then sealed and inverted several times to achieve thorough mixing. These steps were repeated for three other samples and their duplicates.

Three pressure tubes, referred to as sugar recovery solutions (SRS), each containing 10mL of a known sugar solution (mixture of D-(+)glucose, D-(+)xylose, D-(+)galactose, - L(+)arabinose, and D-(+)mannose) and 350 µl of H_2SO_4 were then prepared. These were used to approximate the sugar losses during the secondary step of the hydrolysis, which was completed by transferring all the pressure tubes (11) to an autoclave, where 121 °C was maintained for 60 min.

PICTURE 1. Pressure tubes with hydrolyzed samples after autoclaving. (Photo: Gladys Batisson)

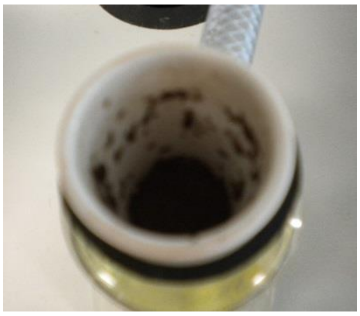
Once the temperature dropped to 80 °C, the pressure tubes were removed and let to cool down to room temperature. Using vacuum suction the contents of the tubes were then filtered through filter crucibles of known weight as can be seen in the following photo.

PICTURE 2. Filtering the two-stage acid hydrolysis liquids by vacuum suction.

During acid hydrolysis lignin fractionates into two parts: solid acid-insoluble and liquid acid-soluble lignin (ASL). The filtered hydrolysates solutions gained from the previous step as can be seen in Picture 3 were stored for determination of acid-soluble lignin (ASL) by ultraviolet-spectroscopy (see Chapter 3.2.4) and carbohydrates by chromatography (see Chapter 3.2.5).

PICTURE 3. Hydrolysates. Lodgepole pine in the front. (Photo: Daniel Hayes)

The acid-insoluble portion of lignin is referred to as Klason lignin, and it is counted as the organic fraction of the acid insoluble residue. This residue was gained by carefully washing the remaining solids from the pressure tubes through the filters with deionized water. The filters with the solid residues as can be seen in Picture 4 were then dried overnight at 105 °C, after which the acid insoluble residue (AIR) content was determined. The acid insoluble ash (AIA) was determined by ashing the dried mass in this filter crucible and Klason lignin (KL) determined as the organic fraction of the residue: AIR minus AIA.



PICTURE 4. The solid residue from acid hydrolysis ready for drying and the determination of acid-insoluble residue (AIR). (Photo: Gladys Batisson)

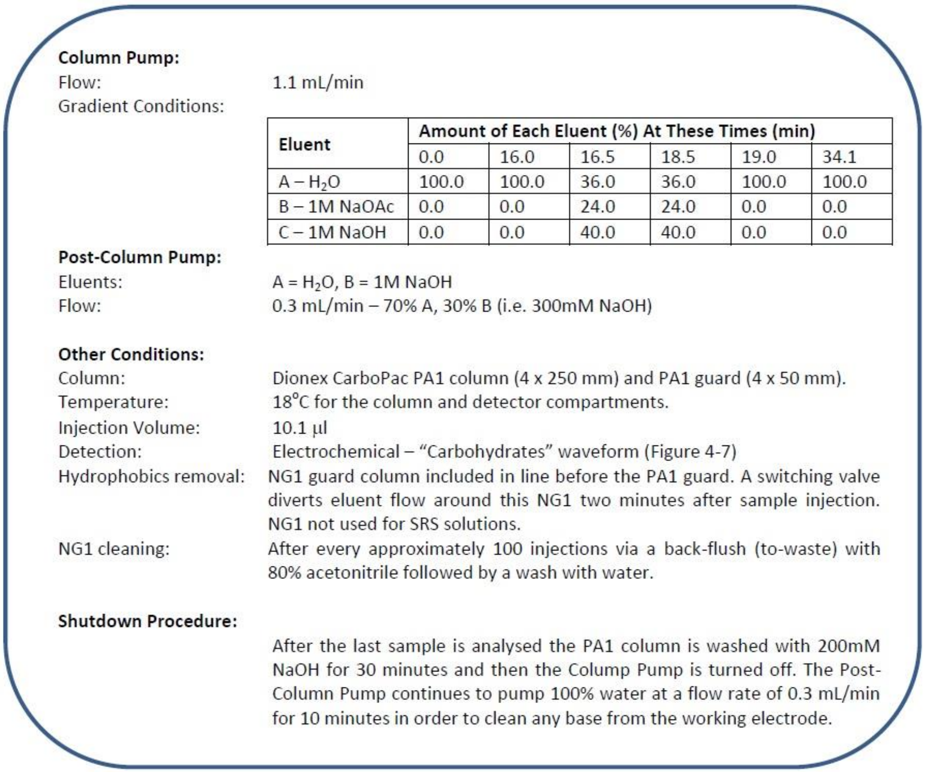
3.2.4 Determination of acid-soluble lignin by ultra-violet spectrometer

The method for determining the ASL is presented here for clarification, however only Klason lignin was predicted in this work. The hydrolysate gathered from the previously defined procedure includes the ASL and it was measured by ultra-violet (UV) spectroscopy. In this work, HP Agilent 8452A diode array spectrophotometer and 4% sulfuric acid (H_2SO_4) as a background blank were used. Appropriate amount of the hydrolysate was placed in a 3 mL quartz cuvette and diluted with 4 % H_2SO_4 solution until the absorbance fell into the range 0.7-1.0. Each sample was analysed in duplicates, reproducibility being ± 0.05 absorbance units. Since wavelength recommendation value for woody feedstocks was found to be 240nm, this and absorptivity constant of 110 L/g/cm were used to calculate the ASL. (Sluiter et al., 2008)

3.2.5 Determination of sugars by chromatography

This work utilized the chromatographic conditions seen in Figure 13, described by Hayes (2011). The determination of these optimum conditions for carbohydrate analysis were adapted from the Davis (1998) protocol, with some modifications made by the researchers at the US Forest Products Laboratory. (Hayes, 2011) In contrast to the NREL method where sample is neutralized before chromatography, this study found out that the peaks of the sugars in the spectrum actually became sharper with an increase in the sulphate (SO_4^{2-}, salt of sulfuric acid) load compared with aqueous samples. (Davis 1998, 250) But in order to prevent variability from an absolute response of an analyte (e.g. analyte losses), internal standard (ISTD) is used as a dilutor and the calibration is based on the ratio between the analyte and the standard. ISTD should be similar but not identical to the analyte of interest. Dilution factor 1/5 and fucose as ISTD was used. Instrument used was DIONEX ICS-3000 ion chromatography system with an electrochemical detector (using Pulsed Amperometric Detection, PAD), a gradient pump, a temperature controlled column and detector enclosure plus an AS50 autosampler.

The diluted hydrolysates were first filtered through 0.2 lm Teflon syringe filters and placed in vials for the analysis. Figure 13 shows the amount of diluted sample injected by the autosampler into the chromatographer, the column used to separate the sugars and how hydrophobic material was removed. Also the column conditions are stated and from the gradient conditions can be seen that after 16 min of eluent flow the separation of sugars occurred and after that the column was regenerated and re-equiblirated prior to the injection of next sample. Shutdown procedure for the instrument is also explained. This method allowed the separation of fucose, arabinose, galactose, rhamnose, glucose, xylose, and mannose. At regular intervals during the procedure sugar standard samples of known concentration (the fucose 1/5 diluted SRS from the hydrolysis) were injected. This was done in order to determine the response factors of all the sugars of interest compared with the internal standard fucose.



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_2O	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 = H_2O, 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 Columb 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.

FIGURE 13. Chromatographic conditions used for sugar analysis. Adapted from (Hayes 2011)

3.3 NIR model development

In this chapter the method for collection of the NIR spectra of a sample and the subsequent spectral modifications are further explained. The time of only scanning via NIR takes approximately 5 min. The FOSS XDS monochromator with Rapid Content Analyser (RCA) module and the Vision 3.5 software were used to scan the samples with near-infrared radiation in the wavelengths of 1100-2500 nm. It was found out by Hayes (2012) that the inclusion of wavelengths <1100nm did not improve model performance of any constituent, thus not utilized in this work. Internal standard was scanned approximately every 30min to standardize the scans for e.g. temperature and air humidity changes. The sample was placed in a coarse rectangular cell, which cell window moved in eight different positions, in each which four spectra was collected to cover the window of all the cell. This resulted in 32 spectra that were then averaged to represent as homogeneous distribution of the sample as possible. Data was recorded at 0.5 nm absorbance (log[1/reflectance]) intervals, which added up to 2800 data points per each spectrum. Each sample was scanned in duplicates.

Samples consisting of more than 10% extractives (27 of the total 110) were NIR scanned both before and after extraction, labelled (NIR code)-DJ-A and (NIR code)-DJ-E, respectively, as can be seen in Appendix 1. However, as can be noted from the Chapter 2.2.3 on extractives, it was essential to use the wet-chemical data after extraction for both of the scans, but the original 'A-spectrum' was compensated for the extractives in the model development.

From the Vision software the spectra were then exported and imported into The Unscrambler 10.1 (Camo Software AS). This software implemented all the spectral treatments as well as the model development. Since it was studied that the raw spectral data tended to have lower R^2 values and require more PCs than the models utilizing e.g. second derivative, second-order Savitzky-Golay (SG) derivative with smoothing using a 2nd-order polynomial with gap-segment of 25 data points (nm) from both the left and right sides were used. It appeared that some important spectral data was lost with MSC, SNV, or SNVDT transforms and derivative seemed to be better choice for pretreatment. (Hayes, 2011) For each of the constituent of interest (glucose, xylose, Klason lignin and mannose) a model was developed. One variable partial least squares (PLS1) regression was used for each constituent's spectrum when building the model.

4 RESULTS

In this chapter the results are presented both visually as graphics as in a table format in numerical values. The chapter is divided to first present the local model following the global. The most important variations in the results are noted, especially while comparing the two different models in the last subchapter.

4.1 Local NIRS models

It can be seen that there is a rather good representation of sample types (Appendix I) and approximately even distribution between compositions of glucose, xylose and mannose in the wood-specific (local) calibration set (Table 3).

TABLE 3. Descriptive statistics for the local calibration set (wet-chemical data).

	Glucose	Klason lignin	Xylose	Mannose
Min	11.38	15.36	2.24	0.34
Max	48.61	55.76	23.10	13.23
Range	37.23	40.40	20.86	12.90
STD	7.69	7.19	5.00	3.98
Average	32.86	28.39	8.17	5.28

Correlations between selected constituents can be seen in Table 4. Relationships of note include the negative correlations between Klason lignin and the major sugars glucose and xylose, and between mannose and xylose. Clear positive correlation was seen between mannose and glucose. These dynamics are equivalent to the characteristics of cellulose, hemicellulose and lignin distributions in wood, see Chapter 2.2.

TABLE 4. Correlation statistics for the local calibration set (wet-chemical data).

	Glucose	Klason lignin	Xylose	Mannose
Glucose	1	-0.51	0.14	0.45
Klason lignin		1	-0.61	0.15
Xylose			1	-0.62
Mannose				1

The regression between model predicted and reference analysis values (in % of dry matter) for both calibration and validation sets for wood models are shown for the four different constituents in the following Figure 14. It can be seen that the slopes of regression do not differ much, referring that the compositions in both sets were analogous. Validation set is approximately evenly distributed across the whole concentration range in all cases, however for Klason lignin against the calibration set the validation set shows lower upper range.

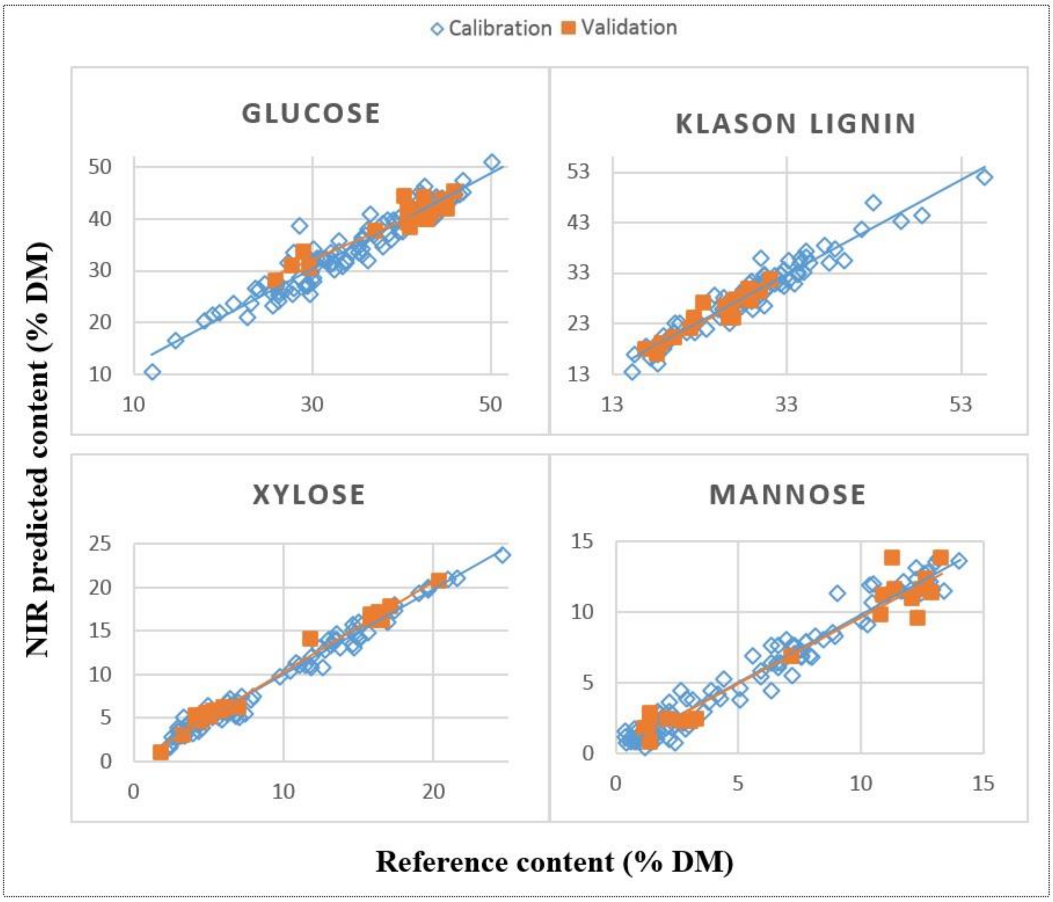


FIGURE 14. Predicted versus measured (reference analysis) values for wood-specific calibration set and external validation set.

In the following Table 5 the statistical values for local model are given. The upper part shows the statistics for the calibration set and the lower part for the validation set tested on this model. Constituents are ordered according to the mean concentrations from highest to lowest.

TABLE 5. Summary statistics for the calibration (upper) and validation (lower) of the wood-specific model.

Constituent	Samples	Factors	R^2_{cv}	RMSECV	RPD_{cv}	RER_{cv}
Glucose	105	6	0.89	2.65	3.05	14.76
Klason lignin	106	7	0.90	0.92	3.18	18.51
Xylose	110	8	0.97	1.08	5.76	23.99
Mannose	109	9	0.94	2.17	3.94	12.60

Constituent	Samples	Factors	R^2_{pred}	RMSEP	RPD_{pred}	RER_{pred}
Glucose	20	6	0.88	2.29	2.64	8.55
Klason lignin	20	7	0.89	1.37	3.01	10.17
Xylose	20	8	0.99	0.72	8.61	26.91
Mannose	20	9	0.95	1.15	4.34	10.34

4.2 Global NIRS models

There was a wide variety of different biomass sample types in the global model (Appendix II) and thus, not that even distribution between compositions of glucose, xylose and mannose as can be seen in the Table 6. Most probably due to the chromatographic detection difficulties in case of really low concentrations (Hayes 2012), zero value for the wet-chemical data can be seen in case of mannose.

TABLE 6. Descriptive statistics for the global calibration set (wet-chemical data).

	Glucose	Klason lignin	Xylose	Mannose
Min	3.77	0.07	0.45	0
Max	98.69	72.21	27.71	14.04
Range	94.92	72.14	27.26	14.04
STD	13.73	11.41	7.58	2.76
Average	36.37	22.67	14.53	1.74

The regression of model predicted to reference analysis values (in % of dry matter) for both calibration and validation sets for global model are shown for the four different constituents in the following Figure 15. It can be seen that the distribution between all the constituents for calibration set is approximately evenly distributed. Due to the large amount of different samples in the global model, the validation set of samples against them do not show that evenly distributed range for glucose or Klason, however even distribution can be seen for xylose and mannose.

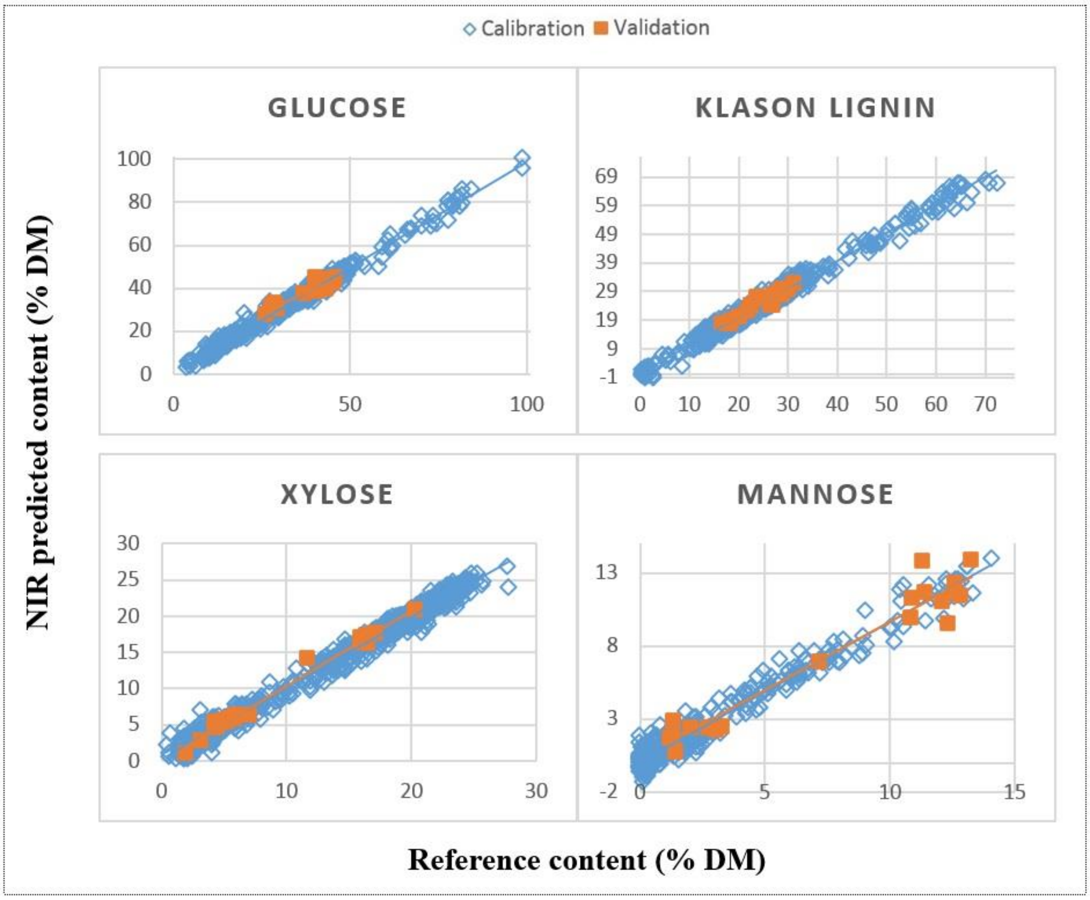


FIGURE 15. Predicted versus reference analysis measured values for global calibration set and external validation set.

In the following Table 7 the statistical values for global model are given. The upper part shows the statistics for the calibration set and the lower part for the validation set tested on this model. Constituents are ordered according to the mean concentrations from highest to lowest.

TABLE 7. Summary statistics for the calibration (upper) and validation (lower) sets of the global model.

Constituent	Samples	Factors	R^2_{cv}	RMSECV	RPD_{cv}	RER_{cv}
Glucose	732	27	0.98	2.10	6.53	45.11
Klason lignin	796	27	0.97	1.83	6.22	39.31
Xylose	737	25	0.98	1.08	7.02	25.24
Mannose	727	27	0.94	0.67	4.13	21.01

Constituent	Samples	Factors	R^2_{pred}	RMSEP	RPD_{pred}	RER_{pred}
Glucose	20	27	0.88	2.05	2.93	9.49
Klason lignin	20	27	0.83	1.76	2.40	8.12
Xylose	20	25	0.99	0.85	7.04	21.99
Mannose	20	27	0.95	1.10	4.61	11.00

4.3 Comparison between models

Since the calibration can be tight up to the samples used in the calibration set, by only modelling the calibration rarely gives a realistic demonstration on future performance. Thus, validation results were considered to be the most essential. However, it could be noted that both individual calibration models gave good R^2 values 0,89-0.97 and 0,94-0,98, for wood and global respectively. RMSEP was under 3% for both. Even though topic-specific evaluation has to be taken into account, some general guidelines for evaluating the model performance based on statistical results is shown in Table 8. R^2 and RER are adapted from Ward, Nielsen & Møller (2011, 4834) and RPD from Li-Chan, Chalmers & Griffiths (2011, 362).

TABLE 8. General guideline values for approximating model performance, the suitable applications mentioned in brackets.

Degree of calibration success	R^2	RPD	RER
Excellent (process control/quantification)	>0.95	>6.5	>20
Successful (quality control)	0.9-0.95	5-6.5	15-20
Moderately successful (screening)	0.8-0.9	3-5	10-15
Moderately useful (rougher screening)	0.7-0.8	2-3	8-10

4.3.1 Comparison between R^2, RPD and RER values

The following Table 9 shows statistics for prediction for all the constituents for both of the models. Both global and local model performed well in predicting the constituents in the 20 wood samples. Good values of R^2 were obtained for all the constituents, however xylose showed excellent results for R^2 for both of the models, local being slightly higher. No significant variation for R^2, expect the increase in accuracy for Klason lignin in local model, can be seen. Thus also biggest improvement in RMSEP value, 0.39% (from 1.76% to 1.37%, relative percentage decrease between values 22.2%) is shown for local model for Klason lignin. Otherwise the better values, higher R^2 and lower RMSEP, can be seen for glucose and mannose in global model compared to local.

RPD and RER values are consistent with the previous finding; the higher the result the more accurate the model. Compared to the Table 8 above, mannose results could be considered moderately successful and glucose and Klason lignin results moderately useful for screening purposes. Xylose again shows prime results; the highest RER and RPD values reflecting a stronger model for prediction, thus well suitable for process control and quantification.

TABLE 9. Validation statistics for both global (G) and local wood-specific (L) models. Coefficient of determination is shown in more accurate decimal precision in order to show the variations better. Bolded results are better when (G) and (L) for one constituent and one statistic term is compared.

	G	**L**	**G**	**L**	**G**	**L**	**G**	**L**
	R^2_{pred}		RMSEP		RPD_{pred}		RER_{pred}	
Glucose	**0.8836**	0.8785	**2.05**	2.29	**2.93**	2.64	**9.49**	8.55
Klason lignin	0.8278	**0.8917**	1.76	**1.37**	2.40	**3.01**	8.12	**10.17**
Xylose	0.9855	**0.9892**	0.85	**0.72**	7.04	**8.61**	21.99	**26.91**
Mannose	**0.9535**	0.9470	**1.10**	1.15	**4.61**	4.34	**11.00**	10.34

4.3.2 Comparison between SEP values

As presented in Chapter 2.5.6, Naes *et al.* gave the formula for testing the differences between two standard error of prediction (SEP) values between two models (Naes *et al.* 2007, 166-170). Because amount of samples in prediction set was 20, degrees of freedom was $(N_p - 2)$ = 18, α= 0.025 and thus, t = 2,101 according to Student's t-distribution. (Miller & Miller 2010, 266) All the intervals include 1, as can be seen in the Table 10 and Appendix IV for the calculations. This means based on this test there is no statistically significant difference at the 5% significance level.

TABLE 10. Test for differences between SEP values of global (G) and local (L) models.

	Lower	**Upper**
Glucose	0.76	1.61
Klason lignin	0.62	1.03
Xylose	0.56	1.19
Mannose	0.82	1.36

5 DISCUSSION

The main question behind this work was to see if there would be any significant improvement by restricting the models on solely wood samples. Wood samples were chosen for their relatively accurate reference method results, even distribution across the concentration range and availability of samples. As can be seen in Appendix I there were 64 softwood and 46 hardwood samples, and when considering their compositional differences (Table 1), there might not be an even distribution of constituents across the whole constituent concentration range. It could be noted that the excellent xylose prediction could be due to the bigger softwood sample amount and lower xylose concentration in this type of wood - thus smaller range and probably lower residual amount. Nevertheless, the division of validation set for softwood:hardwood was 13:7, approximately the same ratio than for the calibration set.

However, if and when in spite of slight unequal distribution between soft- and hardwood samples the glucose, Klason lignin, xylose and mannose concentrations were evenly represented, the differences with ranges between wood and global could be discussed. The amount of samples in global model was almost seven times higher than in the wood-specific model. According to Boysworth and Booksh written in the 10th chapter of Burns & Ciurzcak authored book, the distribution of variance across a large concentration range can result in global model being not the best approach to choose. (Burns & Ciurczak 2008, 218) However when observing the prediction of any extreme (low or high) value, with wider concentration range like in the case of the global model, these could be easier predicted than with the local model. Where, in contrast, these extreme concentrations might be shown as outliers and thus, even with one outlier the equation curve would be skewed influencing negatively on R^2 and RMSEP values. However, only consistent outliers in the both the global and local models between all constituents were removed from model developments (Hayes 2012). It was considered that even though species of paulownia for glucose concentration according to Grubb’s test was the only significant outlier in the local model, it was not excluded due to the fact that for the other constituents, this sample showed relatively accurate results (not anywhere near the critical value of Grubb’s test).

As the measured constituent seemed to be a factor of better prediction performance (xylose and mannose gave higher R^2 and lower RMSEP), the concentration of certain constituent could have an influence. For this both the reference method and variation between samples are important factors. For example the value considered to be Klason lignin (organic portion of the solid acid-insoluble residue gained from two-stage acid hydrolysis) can actually be influenced by other compounds. These can be sugar degradation products like furfural and hydroxymethyl furfural (HMF) which could have been condensed to solid products (Hayes 2011, 42). Hence, the concentration of these compounds based on the severity of the hydrolysis procedure (sugars more degraded), can alter the accuracy of the reference method results.

As well as the lower concentration, also the distribution range for the better predicted constituents was smaller: for local model mannose 0-14%, xylose 2-24%, Klason lignin 15-55%, glucose 12-51% and for global 0-14%, 0-28%, 0-72%, 4-99% respectively. Analogous ranges to that of local calibration set can be seen for the 20 wood sample validation set, even though Klason lignin's upper limit for this set was lower, 31%, and glucose had smaller range of 26-46%. Improved statistics could occur when filling these gaps in the concentration range by addition of more representable samples for the validation set (Hayes 2012). According to the indicative average values for compositions in Table 1, the whole lignin should be maximum around 31% of both soft- and hardwood (Sjöström 1981, 208). Klason lignin, as only partial of this (in addition to ASL), in local model exceeds this value prominently. Reasons can be the before-mentioned additive compounds falsely calculated as Klason lignin (AIR-AIA) or because acid-insoluble ash (AIA) is often difficult to predict (Hayes 2012). The lower minimum value in local model for glucose could be explained by the fact that any concentration value lower than 26% was from a bark sample. Bark, as mentioned has higher ash, extractive and lignin content than the stem, thus lower polysaccharide content (Häkkilä 1989, 148-159).

As shown in the Table 5 relatively small number of factors (<9) to adequately describe the amount of variance in local models were used, thus this estimates that these models were not overfitted and that chemical differences between samples were not large. However, in the case of global model a great amount of factors (25 or 27) were needed to explain the variations, which is due to the different types of samples and characteristics in the model. According to Downes, Medera & Harwood over 10 is typically considered a high value for factors. (Downes *et al.* 2011) However, this depends on the amount of samples in the calibration set. For example, over 10 factors for 50 samples is a lot but 25/27 for 750 is not. (Hayes 2010)

The result to the main objective was, that there was no clear significant increase in accuracy by using local model instead of global for this set of samples and validation set, as can be seen for the validation statistics. Similar results have been gained from other studies. Research paper on models based on different pine species from multiple regions against species-specific models on only two pine species, summarized the results in the following manner: “In summary, there is no evidence in this study to support the idea that predictions from species-specific calibration models will always be better than those from a robust global calibration model.” (Hodge & Woodbridge 2010) It was also noted in a study paper by the CEO of Celignis Daniel Hayes (2012) where models based on variety of Miscanthus (silvergrass) samples were compared to only Miscanthus giganteus samples that there was no consistent difference (Hayes 2012). Also three other studies: quality of hay forages (Abrams, Shenk, Westerhaus & Barton 1987), element concentration in needles (Gillon, Houssard & Joffre 1999) and cellulose content in eucalyptus woodmeal (Downes, Medera & Harwood 2011) all concluded broad-based global calibrations to be just as effective as calibrations based on smaller sample subsets.

Even by using the method for checking statistically significant differences between two calibration models (see Formulas 19-21, Appendix IV for values and results and Table 10 for summarized results), by SEP correlation method, none of the SEP variations between models were considered significant. Thus, both the local or global models could be seen as accurate and for future unknown wood samples either one could be used for prediction. The generalization concluding that the global and local models were alike in accuracy prediction could be spread to cover other types of samples, however it seems unlikely that e.g. concentrations of a sample set of municipal waste would be as evenly distributed as that of different woody feedstocks. However, lignocellulosic grasses could be thought to be a comparable set, as was also studied by Abrams *et al.* (1987).

When the accuracy of precision of global and local models are highly similar, it could be seen as profit for the company analysing the samples due to decrease in workload when feedstock-specific models for each material subset would not be necessary. Even though this was not the hypothesis, it can be considered as a valuable result. Based on this research there is no need for species-specific local models when multi-species global models with wide variety of samples gave adequately accurate results.

6 CONCLUSIONS

The use of near-infrared spectroscopy for quantitative analysis of biomass compositions seems promising: as well as being rapid and accurate, it is low laborious and non-destructive to the sample. Nevertheless, robust wet-chemical laboratory methods behind the model development are essential as well as is the representative sample collection. The variations and errors in the models are a combination of sampling and laboratory error, natural sample-specific variations, utilization of spectral treatments and applied multivariate methods. However, model performance based on this work for both wood-specific local and global models evaluated on an external wood validation set were acceptable for glucose and Klason lignin and excellent for mannose and xylose. Both of the models predicted the validation set with similar accuracies. Celignis Ltd can use these results when considering on developing new models.

REFERENCES

Abrams, S., Shenk, S., Westerhaus, M. & Barton, F. 1987. Determination of Forage Quality by Near Infrared Reflectance Spectroscopy: Efficacy of Broad-Based Calibration Equations. Journal of Dairy Science 70(4), 806-813.

Amarasekara, A. 2014. Handbook of Cellulosic Ethanol. US, New Jersey: Wiley.

Anderson, R., Bendell, D. & Groundwater, P. 2004. Organic Spectroscopic Analysis. UK, Cambridge: The Royal Society of Chemistry (RSC).

Beebe, K., Pell, R., & Seasholtz M.B. 1998. Chemometrics: A Practical Guide. US, New York: Wiley.

Berg, J., Tymoczko, J & Stryer, L. 2002. Biochemistry. 5th edition. US, New York: W. H. Freeman.

Biedermann, F. & Obernberger, I. 2005. Ash-related Problems during Biomass Combustion and Possibilities for a Sustainable Ash Utilization. Austria, Graz: Austrian Bioenergy Centre.

Burns, D. & Ciurczak, E. 2008. Handbook of Near-Infrared Analysis. 3rd edition. USA, Boca Raton: Taylor and Francis Group, LLC.

Cabassi, G., Marino Gallina, P., Piombino, M., Orfeo, D. & Maffioli, G. 2005. Estimation of the properties of heterogeneous soils on a multi regional scale in Italy using near infrared spectroscopy and LOCAL calibration procedures. Conference paper. 12th International Conference of Near Infrared Spectroscopy, April 2005.

CAMO. 2015. The Unscrambler: All-in-one Multivariate Data Analysis and Design of Experiments software. Read 30th September 2015: http://www.camo.com/downloads/products/The_Unscrambler.pdf

Capareda, C. 2014. Introduction to Biomass Energy Conversions. 1st edition. USA, Boca Raton: Taylor and Francis Group, LLC.

Cheng, J. 2010. Biomass to Renewable Energy Process. 1st edition. USA, Boca Raton: Taylor and Francis Group, LLC.

Christopher, L. 2013. Integrated Forest Biorefineries. UK, Cambridge: The Royal Society of Chemistry.

Clarke, S. & Preto, F. 2015. Biomass Burn Characteristics. Ontario Ministry of Agriculture, Food and Rural Affairs. Read 30th September 2015. http://www.omafra.gov.on.ca/english/engineer/facts/11-033.htm#5

Davis M. 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 (2), 235–252. Read 5th September 2015.

http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.472.456&rep=rep1&type=pdf

Downes, G. Medera, R & Harwood, C. 2011. 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.

Ek, M., Gellerstedt, G. & Henriksson, G. 2009. Wood Chemistry and Biotechnology. Germany, Berlin: Walter de Gruyter.

Esbensen, K., Dominique, G., Westad, F. & Houmoller, L. 2002. Multivariate Data Analysis – in Practise: An Introduction to Multivariate Data Analysis and Experimental Design. 5th edition. Norway, Oslo: CAMO

Fearn, T. 2002. Assessing calibrations: SEP, RPD, RER and R2. NIR News 13(6): 12-14.

Frost, J. 2013. Multiple Regression Analysis: Use Adjusted R-Squared and Predicted R-Squared to Include the Correct Number of Variables. Minitab. Read 22nd October 2015. http://blog.minitab.com/blog/adventures-in-statistics/multiple-regession-analysis-use-adjusted-r-squared-and-predicted-r-squared-to-include-the-correct-number-of-variables

Gemperline, P. 2006. Practical Guide To Chemometrics. 2nd edition. US, Boca Raton: Taylor & Francis Group, LLC.

Gillon, D., Houssard, C. & Joffre, R. 1999. Using Near-Infrared Reflectance Spectroscopy to Predict Carbon, Nitrogen and Phosphorus Content in Heterogeneous Plant Material. Oecologia 118(2), 173-182.

Griffiths, P. & Haseth, J. 2007. Fourier Transform Infrared Spectrometry. 2nd edition. USA, New Jersey: John Wiley & Sons.

Grubbs, F. 1969. Procedures for detecting outlying observations in samples. Technometrics (11) 1–21. Read 16th November 2015. http://web.ipac.caltech.edu/staff/fmasci/home/astro_refs/OutlierProc_1969.pdf

Haaland, D. & Thomas, E. 1988. Partial least-squares methods for spectral analyses. 1. Relation to other quantitative calibration methods and the extraction of qualitative information. Analytical Chemistry 60(11), 1193-1202.

Hayes, D. 2012. Development of near infrared spectroscopy models for the quantitative prediction of the lignocellulosic components of wet Miscanthus samples. Bioresource Technology 119, 393–405.

Hayes, D. 2011. Analysis of Lignocellulosic Feedstocks for Biorefineries with a Focus on The Development of Near Infrared Spectroscopy as a Primary Analytical Tool. University of Limerick. Doctoral thesis.

Hayes, D. 2010. Rapid Biomass Analysis Methods. Lecture. DIBANET Summer School 14th December 2010. Brazil: Rio de Janeiro.

Hodge, G. & Woodbridge, W. 2010. Global near infrared models to predict lignin and cellulose content of pine wood. Journal of Near Infrared Spectroscopy 18, 367-380.

Häkkilä, P. 1989. Utilization of Residual Forest Biomass. 1st edition. Germany, Berlin: Springer-Verlag.

Jolliffe, I. 2002. Principal Component Analysis. 2nd edition. US, New York: Springer.

Kaur, H. 2009. Spectroscopy. 1st edition. India, Meerut: Pragati Prakashan.

Kelley, S. Professor and the Head of the Department of Forest Biomaterials at North Carolina State University. 2015. Biorefinery. Lecture. 23rd September 2015. Tampere University of Applied Sciences. Finland: Tampere.

Li-Chan, E., Chalmers, J. & Griffiths, P. 2011. Applications of Vibrational Spectroscopy in Food Science. 1st edition. UK, West Sussex: John Wiley & Sons, Ltd.

Loo, S. & Koppejan, J. 2008. The Handbook of Biomass Combustion and Co-firing. 2nd edition. UK, London: Earthscan.

Miller, J.N. & Miller, J.C. 2005. Statistics and Chemometrics for Analytical Chemistry.5th edition. UK, Essex: Pearson Education Limited. Read 2nd November 2015. http://197.14.51.10:81/pmb/CHIMIE/0273730428.pdf

Naes, T., Isaksson, T, Fearn, T. & Davies, T. 2007. A User-Friendly Guide to Multivariate Calibration and Classification. UK, Chichester: NIR Publications.

Niederberger, J., Todt, B., Boca, A., Nitschke, R., Kohler, M., Kühn, P. & Bauhus, J. 2015. Use of near-infrared spectroscopy to assess phosphorus fractions of different plant availability in forest soils. Biogeosciences (12) 3415–3428. Read 22nd October 2015. http://www.biogeosciences.net/12/3415/2015/bg-12-3415-2015.pdf

Rowell, R. 2005. Handbook of Wood Chemistry and Wood Composites. 1st edition. USA, Boca Raton: Taylor and Francis Group.

Sariyildiz, T. & Anderson, J. 2005. Variation in the chemical composition of green leaves and leaf litters from three deciduous tree species growing on different soil types. Journal of Forest Ecology and Management, volume 210, 303–319.

Shenk, J. & Westerhaus, M. 1997. Investigation of a LOCAL calibration procedure for near infrared instruments. Journal of Near Infrared Spectroscopy 5(4), 223–232.

Sims, R., Taylor, M., Saddler, J. & Mabee, W. 2008. From 1st- to 2nd – Generation Biofuel Technologies. An overview of current industry and RD&D activities. France, Paris: OECD/IEA.

Shi, Y. & Weimer, P.J. 1996. Utilization of Individual Cellodextrins by Three Predominant Ruminal Cellulolytic Bacteria. Applied and Environmental Microbiology (Vol. 62, No. 3) p. 1084–1088.

Sjöström, E. 1993. Wood Chemistry: Fundamentals and Applications. 2nd edition. USA, San Diego: Academic Press.

Sluiter, A., Hames, B., Hyman, D., Payne, C., Ruiz, R., Scarlata, C., Sluiter, J., Templeton, D. & Wolfe J. 2008. Determination of Total Solids in Biomass and Total Dissolved Solids in Liquid Process Samples. Laboratory Analytical Procedure (LAP). Technical Report. NREL. Read 22nd October 2015. http://www.nrel.gov/docs/gen/fy08/42621.pdf

Sluiter, A., Hames, B., Ruiz, R., Scarlata, C., Sluiter, J., & Templeton, D. 2005. Determination of Ash in Biomass. Laboratory Analytical Procedure (LAP). Technical Report. NREL. Read 15th of September 2015. http://www.nrel.gov/docs/gen/fy08/42622.pdf

Sluiter, A., Hames, B., Ruiz, R., Scarlata, C., Sluiter, J., Templeton, D. & Crocker, D. 2008. Determination of Structural Carbohydrates and Lignin in Biomass. Laboratory Analytical Procedure (LAP). Technical Report. NREL. Read 15th of September 2015. http://www.nrel.gov/biomass/pdfs/42618.pdf

Sluiter, A., Ruiz, R., Scarlata, C., Sluiter, J. & Templeton, D. 2005. Determination of Extractives in Biomass. Laboratory Analytical Procedure (LAP). Technical Report. NREL. Read 13th September 2015. http://www.nrel.gov/docs/gen/fy08/42619.pdf

Sluiter, B., Ruiz, R., Scarlata, C., Sluiter, A. & Templeton, D. 2010. Compositional Analysis of Lignocellulosic Feedstocks. 1. Review and Description of Methods. Journal of Agricultural and Food Chemistry (2010 Aug 25) 58(16): 9043–9053. Read 22nd of October 2015. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2923870/#ref7

Sluiter, J. & Sluiter, A. 2011. Summative Mass Closure. Laboratory Analytical Procedure (LAP). Technical Report. NREL. Read 14th September 2015. http://www.nrel.gov/docs/gen/fy11/48087.pdf

Stenius, P. 2000. Forest Products Chemistry. Papermaking Science and Technology 3. 1st edition. Helsinki, Finland: Fapet OY.

US DOE. 2005. Genomics: GTL Roadmap, DOE/SC-0090. U.S. Department of Energy Office of Science, 204. Read 16th November 2015. http://genomicscience.energy.gov/roadmap/

Ward, A., Nielsen, A. & Møller, H. 2011. Rapid Assessment of Mineral Concentration in Meadow Grasses Sensors by Near Infrared Reflectance Spectroscopy. Sensors (11) 4830-4839.

Williams, P.& Norris, K. 1987. Variables Affecting Near-Infrared Reflectance Spectroscopic Analysis. Near-Infrared Technology in the Agriculture and Food Industries. US, Minnesota: American Association of Cereal Chemists, 143-167.

Yang, S-T. 2007. Bioprocesses for value-added products from renewable resources: new technologies and applications. 1st edition. UK, Oxford: Elsevier B.V.

APPENDICES

Appendix 1. Samples used for the wood-specific calibration model.

1(3)

The whole set of 110 samples was used for glucose model development. Excluded sample numbers for other constituents; Klason lignin 13, xylose 76;77;92;93 and mannose 40;76;77;92;93.

1	80001-DS-A	BNM WASTE WOOD PALLETS
2	80003-DJ-A	OAK
3	80005-DJ-A	EUROPEAN LARCH
4	80006-DJ-A	WESTERN RED CEDAR
5	80013-DJ-A	Species = Alder, Region = SE, Period = FLUSH, Partition = BARK
6	80013-DJ-E	Species = Alder, Region = SE, Period = FLUSH, Partition = BARK
7	80015-DJ-A	Species = Alder, Region = MID, Period = DORM, Partition = BARK
8	80015-DJ-E	Species = Alder, Region = MID, Period = DORM, Partition = BARK
9	80018-DJ-A	Species = Alder, Region = NW, Period = DORM, Partition = BARK
10	80018-DJ-E	Species = Alder, Region = NW, Period = DORM, Partition = BARK
11	80024-DJ-A	Species = Ash, Region = MID, Period = DORM, Partition = BARK
12	80024-DJ-E	Species = Ash, Region = MID, Period = DORM, Partition = BARK
13	80033-DJ-A	Species = Birch, Region = MID, Period = DORM, Partition = BARK
14	80037-DJ-A	Species = Birch, Region = MID, Period = DORM, Partition = BARK
15	80039-DJ-A	Species = Birch, Region = NW, Period = FLUSH, Partition = BARK
16	80039-DJ-E	Species = Birch, Region = NW, Period = FLUSH, Partition = BARK
17	80042-DJ-A	Species = Lodgepole Pine, Region = MID, Period = DORM, Partition = BARK
18	80042-DJ-E	Species = Lodgepole Pine, Region = MID, Period = DORM, Partition = BARK
19	80046-DJ-A	Species = Lodgepole Pine, Region = NW, Period = FLUSH, Partition = BARK
20	80046-DJ-E	Species = Lodgepole Pine, Region = NW, Period = FLUSH, Partition = BARK
21	80049-DJ-A	Species = Norway Spruce, Region = SE, Period = FLUSH, Partition = BARK
22	80049-DJ-E	Species = Norway Spruce, Region = SE, Period = FLUSH, Partition = BARK
23	80050-DJ-A	Species = Norway Spruce, Region = MID, Period = GROW, Partition = BARK
24	80050-DJ-E	Species = Norway Spruce, Region = MID, Period = GROW, Partition = BARK
25	80051-DJ-A	Species = Norway Spruce, Region = MID, Period = DORM, Partition = BARK
26	80051-DJ-E	Species = Norway Spruce, Region = MID, Period = DORM, Partition = BARK
27	80053-DJ-A	Species = Norway Spruce, Region = NW, Period = GROW, Partition = BARK
28	80055-DJ-A	Species = Norway Spruce, Region = NW, Period = FLUSH, Partition = BARK
29	80055-DJ-E	Species = Norway Spruce, Region = NW, Period = FLUSH, Partition = BARK
30	80062-DJ-A	Species = Sitka Spruce, Region = NW, Period = GROW, Partition = BARK
31	80062-DJ-E	Species = Sitka Spruce, Region = NW, Period = GROW, Partition = BARK
32	80064-DJ-A	Species = Sitka Spruce, Region = NW, Period = GROW, Partition = BARK
33	80064-DJ-E	Species = Sitka Spruce, Region = NW, Period = FLUSH, Partition = BARK
34	80201-DF-A	Scots pine, red deal, more than 212 microns. Paul Grogan Sample
35	80201-DS-A	Scots pine, red deal, more than 212 microns. Paul Grogan Sample
36	80201-DS-E	Scots pine, red deal, more than 212 microns. Paul Grogan Sample

(continues)

2(3)

37	80201-DF-E	Scots pine, red deal, more than 212 microns. Paul Grogan Sample
38	80501-DJ-A	Sitka spruce, less than 125 microns. From Glasgow.
39	80503-DJ-A	Species = Sitka Spruce, Region = SE, Period = GROW, Partition = BRANCH
40	80504-DJ-A	Species = Sitka Spruce, Region = SE, Period = DORM, Partition = BRANCH
41	80504-DJ-E	Species = Sitka Spruce, Region = SE, Period = DORM, Partition = BRANCH
42	80505-DJ-A	Species = Sitka Spruce, Region = SE, Period = FLUSH, Partition = BRANCH
43	80508-DJ-A	Species = Sitka Spruce, Region = MID, Period = FLUSH, Partition = BRANCH
44	80514-DJ-A	Species = Sitka Spruce, Region = SE, Period = FLUSH, Partition = TOP
45	80523-DJ-A	Species = Sitka Spruce, Region = SE, Period = FLUSH, Partition = WOOD
46	80525-DJ-A	Species = Sitka Spruce, Region = MID, Period = DORM, Partition = WOOD
47	80532-DJ-A	Species = Sitka Spruce, Region = SE, Period = FLUSH, Partition = STEM
48	80539-DJ-A	'Real World Sample', WIT Sample ID = B3, Location = Abbyfeale, Species = SS, Type = Energywood, Chipper = Truck, Date chipped = 09_08
49	80542-DJ-A	'Real World Sample', WIT Sample ID = B6, Location = Woodberry, Species = SS, Type = Roundwood, Chipper = Musmax, Date chipped = AUTUMN_07
50	80548-DJ-A	'Real World Sample', WIT Sample ID = B12, Location = Bweeng, Species = SS, Type = Roundwood, Chipper = , Date chipped = 09_07
51	80551-DJ-A	'Real World Sample', WIT Sample ID = B15, Location = Storage Trial Bin 2 Refill, Species = SS, Type = Roundwood, Chipper = Musmax, Date chipped = DEC_07
52	80552-DJ-A	'Real World Sample', WIT Sample ID = B16, Location = Storage Trial Bin 2 , Species = SS, Type = Roundwood, Chipper = Musmax, Date chipped = August_08
53	80554-DJ-A	'Real World Sample', WIT Sample ID = B18, Location = Storage Trial Bin 3, Species = SS, Type = Roundwood, Chipper = Musmax, Date chipped = August_08
54	80555-DJ-A	'Real World Sample', WIT Sample ID = B19, Location = Storage Trial Bin 4, Species = SS, Type = Roundwood, Chipper = Musmax, Date chipped = August_08
55	80562-DJ-A	'Real World Sample', WIT Sample ID = B26, Location = Abbyfeale, Species = SS, Type = Energywood, Chipper = Silvatec, Date chipped = 09_08
56	80563-DJ-A	'Real World Sample', WIT Sample ID = B27, Location = Abbyfeale, Species = SS, Type = Firewood, Chipper = , Date chipped = 09_08
57	80566-DJ-A	Site = INISTIOGE Condition = GREEN Plot Number = 3 Material Type = SS LOGGING RESIDUES
58	81001-DJ-A	Wild poplar, less than 125 microns. From Glasgow.
59	81501-DJ-A	Eucalyptus, Unograndis, from Brazil
60	81502-DJ-A	Eucalyptus grandis, from Brazil
61	82002-DJ-A	Species = Lodgepole Pine, Region = SE, Period = DORM, Partition = BRANCH
62	82003-DJ-A	Species = Lodgepole Pine, Region = SE, Period = FLUSH, Partition = BRANCH
63	82003-DJ-E	Species = Lodgepole Pine, Region = SE, Period = FLUSH, Partition = BRANCH
64	82004-DJ-A	Species = Lodgepole Pine, Region = MID, Period = GROW, Partition = BRANCH
65	82004-DJ-E	Species = Lodgepole Pine, Region = MID, Period = GROW, Partition = BRANCH
66	82006-DJ-A	Species = Lodgepole Pine, Region = MID, Period = FLUSH, Partition = BRANCH
67	82006-DJ-E	Species = Lodgepole Pine, Region = MID, Period = FLUSH, Partition = BRANCH
68	82009-DJ-A	Species = Lodgepole Pine, Region = NW, Period = FLUSH, Partition = BRANCH
69	82009-DJ-E	Species = Lodgepole Pine, Region = NW, Period = FLUSH, Partition = BRANCH
70	82010-DJ-A	Species = Lodgepole Pine, Region = SE, Period = GROW, Partition = TOP
71	82020-DJ-A	Species = Lodgepole Pine, Region = SE, Period = DORM, Partition = WOOD
72	82028-DJ-A	Species = Lodgepole Pine, Region = SE, Period = GROW, Partition = STEM
73	82029-DJ-A	Species = Lodgepole Pine, Region = SE, Period = DORM, Partition = STEM

(continues)

74	82031-DJ-A	Species = Lodgepole Pine, Region = MID, Period = GROW, Partition = STEM
75	82505-DJ-A	Species = Norway Spruce, Region = MID, Period = DORM, Partition = BRANCH
76	82507-DJ-A	Species = Norway Spruce, Region = NW, Period = GROW, Partition = BRANCH
77	82507-DJ-E	Species = Norway Spruce, Region = NW, Period = GROW, Partition = BRANCH
78	82509-DJ-A	Species = Norway Spruce, Region = NW, Period = FLUSH, Partition = BRANCH
79	82512-DJ-A	Species = Norway Spruce, Region = SE, Period = FLUSH, Partition = TOP
80	82521-DJ-A	Species = Norway Spruce, Region = SE, Period = FLUSH, Partition = WOOD
81	82522-DJ-A	Species = Norway Spruce, Region = MID, Period = GROW, Partition = WOOD
82	82530-DJ-A	Species = Norway Spruce, Region = SE, Period = FLUSH, Partition = STEM
83	83001-DJ-A	SPRUCE150
84	85001-DJ-A	Paulowina, Elongata x Fortunes, B4R6
85	85001-DJ-E	Paulowina, Elongata x Fortunes, B4R6
86	85002-DJ-A	Paulowina, Fortunei, B2R9
87	85002-DJ-E	Paulowina, Fortunei, B2R9
88	85503-DJ-A	Species = Ash, Region = SE, Period = FLUSH, Partition = BRANCH
89	85503-DJ-E	Species = Ash, Region = SE, Period = FLUSH, Partition = BRANCH
90	85505-DJ-A	Species = Ash, Region = MID, Period = DORM, Partition = BRANCH
91	85505-DJ-E	Species = Ash, Region = MID, Period = DORM, Partition = BRANCH
92	85509-DJ-A	Species = Ash, Region = NW, Period = FLUSH, Partition = BRANCH
93	85509-DJ-E	Species = Ash, Region = NW, Period = FLUSH, Partition = BRANCH
94	85512-DJ-A	Species = Ash, Region = SE, Period = FLUSH, Partition = TOP
95	85514-DJ-A	Species = Ash, Region = MID, Period = DORM, Partition = TOP
96	85521-DJ-A	Species = Ash, Region = SE, Period = FLUSH, Partition = WOOD
97	85523-DJ-A	Species = Ash, Region = MID, Period = DORM, Partition = WOOD
98	85524-DJ-A	Species = Ash, Region = MID, Period = FLUSH, Partition = WOOD
99	85530-DJ-A	Species = Ash, Region = SE, Period = FLUSH, Partition = STEM
100	86004-DJ-A	Species = Alder, Region = MID, Period = GROW, Partition = BRANCH
101	86004-DJ-E	Species = Alder, Region = MID, Period = GROW, Partition = BRANCH
102	86007-DJ-A	Species = Alder, Region = NW, Period = GROW, Partition = BRANCH
103	86026-DJ-A	Species = Alder, Region = NW, Period = DORM, Partition = WOOD
104	86030-DJ-A	Species = Alder, Region = SE, Period = FLUSH, Partition = STEM
105	86512-DJ-A	Species = Birch, Region = SE, Period = FLUSH, Partition = TOP
106	86519-DJ-A	Species = Birch, Region = SE, Period = GROW, Partition = WOOD
107	86521-DJ-A	Species = Birch, Region = SE, Period = FLUSH, Partition = WOOD
108	86525-DJ-A	Species = Birch, Region = NW, Period = GROW, Partition = WOOD
109	86530-DJ-A	Species = Birch, Region = SE, Period = FLUSH, Partition = STEM
110	86534-DJ-A	Species = Birch, Region = NW, Period = GROW, Partition = STEM

Appendix 2. Samples used for the global model.

Agricultural Residues and Wastes
Straw
Sugarcane Bagasse
Corn Stover
Spent Mushroom Compost
Animal Manures
Poultry Litter
Biorefinery products
Pretreated Biomass
Hydrolysis Residues
Torrified Biomass
Energy crops
Miscanthus
Switchgrass
Coppices
Willow
Reed Canary Grass
Hemp
Grass
Industrial Residues and Wastes
Forest Residues
Sawmill Residues
Hardwood
Softwood
Bark
Pulp
Foliage
Municipal wastes
Municipal solid waste
Compost
Grass
Paper and Cardboard
Foliage

Appendix 3. Samples used for the validation set.

1	80054-DJ-A	Species = Norway Spruce, Region = NW, Period = DORM, Partition = BARK
2	80061-DJ-A	Species = Sitka Spruce, Region = MID, Period = FLUSH, Partition = BARK
3	80513-DJ-A	Species = Sitka Spruce, Region = SE, Period = DORM, Partition = TOP
4	80516-DJ-A	Species = Sitka Spruce, Region = MID, Period = DORM, Partition = TOP
5	80528-DJ-A	Species = Sitka Spruce, Region = NW, Period = DORM, Partition = WOOD
6	80531-DJ-A	Species = Sitka Spruce, Region = SE, Period = DORM, Partition = STEM
7	80534-DJ-A	Species = Sitka Spruce, Region = MID, Period = DORM, Partition = STEM
8	80550-DJ-A	'Real World Sample', WIT Sample ID = B14, Location = Storage Trial Bin 6, Species = SS, Type = Roundwood, Chipper = Musmax, Date chipped = 39661
9	80558-DJ-A	'Real World Sample', WIT Sample ID = B22, Location = Storage Trial Bin 1, Species = SS, Type = Roundwood, Chipper = Musmax, Date chipped = 04_07
10	82035-DJ-A	Species = Lodgepole Pine, Region = NW, Period = DORM, Partition = STEM
11	82501-DJ-A	Species = Norway Spruce, Region = SE, Period = GROW, Partition = BRANCH
12	82516-DJ-A	Species = Norway Spruce, Region = NW, Period = GROW, Partition = TOP
13	82535-DJ-A	Species = Norway Spruce, Region = NW, Period = DORM, Partition = STEM
14	85501-DJ-A	Species = Ash, Region = SE, Period = GROW, Partition = BRANCH
15	85518-DJ-A	Species = Ash, Region = NW, Period = FLUSH, Partition = TOP
16	85531-DJ-A	Species = Ash, Region = MID, Period = GROW, Partition = STEM
17	85536-DJ-A	Species = Ash, Region = NW, Period = FLUSH, Partition = STEM
18	86031-DJ-A	Species = Alder, Region = MID, Period = GROW, Partition = STEM
19	86036-DJ-A	Species = Alder, Region = NW, Period = FLUSH, Partition = STEM
20	86517-DJ-A	Species = Birch, Region = NW, Period = DORM, Partition = TOP

Appendix 4. Test for significant differences between two models: SEP values

1(3)

In the following tables the values for the validation set samples (wet-chemical and NIRS predicted) and the test presented in Chapter 2.5.6 according to Formulas (19-21) are shown for each of the four constituents; glucose, Klason lignin, xylose, mannose.

Glucose	**LOCAL**		**GLOBAL**			
Y	Y pred	Residual	Y pred	Residual	r =	0,634919
26,04	27,83955	1,79955	24,53103	-1,50897	k =	1,292749
27,83	30,94984	3,11984	29,94366	2,11366	L =	1,453278
29,08	33,33533	4,25533	32,50001	3,42001		
40,89	41,90682	1,01682	44,04613	3,15613	SEP1 (L)	2,3328
41,88	40,81377	-1,06623	43,14746	1,26746	SEP2 (G)	2,1004
43,08	39,74939	-3,33061	41,11983	-1,96017		
42,72	40,0431	-2,6769	41,19152	-1,52848	lower	0,764235
37,14	37,65136	0,51136	36,43509	-0,70491	upper	1,614077
40,94	38,33677	-2,60323	40,47204	-0,46796		
45,98	44,95698	-1,02302	45,85929	-0,12071		
40,99	40,54563	-0,44437	39,62081	-1,36919		
44,78	43,61987	-1,16013	47,0112	2,2312		
42,63	43,98372	1,35372	45,99971	3,36971		
43,37	41,30656	-2,06344	43,48566	0,11566		
43,38	41,06472	-2,31528	39,63922	-3,74078		
45,22	41,76852	-3,45148	41,75423	-3,46577		
40,33	44,38277	4,05277	40,73977	0,40977		
29,64	30,63897	0,99897	31,30668	1,66668		
40,7	39,13723	-1,56277	40,63401	-0,06599		
42,79	42,16544	-0,62456	42,54984	-0,24016		

Klason lignin	**LOCAL**		**GLOBAL**			
Y	Y pred	Residual	Y pred	Residual	r =	0,851449
31,24	31,7534	0,5134	32,07492	0,83492	k =	1,134895
23,68	27,09809	3,41809	27,72097	4,04097	L =	1,29288
16,92	17,62732	0,70732	18,3592	1,4392		
20,19	20,01996	-0,17004	19,43064	-0,75936	SEP1 (L)	1,408
18,3	17,16975	-1,13025	17,25471	-1,04529	SEP2 (G)	1,7646
22,54	23,87348	1,33348	24,70555	2,16555		
22,54	23,52278	0,98278	23,99847	1,45847	lower	0,61716
22,02	22,07721	0,05721	22,37699	0,35699	upper	1,031608
18,76	18,87364	0,11364	18,0812	-0,6788		
26,12	26,30115	0,18115	25,71017	-0,40983		
26,58	23,90944	-2,67056	25,05585	-1,52415		
28,9	27,35684	-1,54316	26,23117	-2,66883		
28,67	28,97161	0,30161	26,66698	-2,00302		

(continues)

27,01	24,03856	-2,97144	23,73267	-3,27733
28,81	29,63059	0,82059	29,54617	0,73617
28,98	28,50486	-0,47514	28,22717	-0,75283
26,47	25,67698	-0,79302	24,51583	-1,95417
30,13	29,30947	-0,82053	28,75338	-1,37662
26,83	27,58985	0,75985	25,54134	-1,28866
28,04	28,05769	0,01769	26,95704	-1,08296

Xylose	**LOCAL**		**GLOBAL**			
Y	Y pred	Residual	Y pred	Residual	r =	0,640012
3,25	2,91992	-0,33008	2,36548	-0,88452	k =	1,289564
1,88	0,9742398	-0,9057602	-0,03847	-1,91847	L =	1,450447
11,81	13,99785	2,18785	13,10818	1,29818		
15,91	16,95892	1,04892	16,7298	0,8198	SEP1 (L)	0,6865
17,22	17,69779	0,47779	17,90212	0,68212	SEP2 (G)	0,8401
16,62	16,22562	-0,39438	16,25802	-0,36198		
15,93	16,28135	0,35135	15,88011	-0,04989	lower	0,563388
20,35	20,79568	0,44568	19,36052	-0,98948	upper	1,185254
16,42	17,21106	0,79106	17,31313	0,89313		
4,44	4,634931	0,194931	2,78139	-1,65861		
5,42	5,65487	0,23487	4,65887	-0,76113		
4,27	5,207102	0,937102	3,95819	-0,31181		
4,25	4,477834	0,227834	3,51508	-0,73492		
4,91	5,460495	0,550495	4,93933	0,02933		
4,46	4,725264	0,265264	4,99272	0,53272		
5,15	5,103525	-0,046475	4,62117	-0,52883		
7,13	6,256569	-0,873431	6,98799	-0,14201		
5,12	5,37406	0,25406	5,65716	0,53716		
6,04	6,203019	0,163019	5,57045	-0,46955		
5,88	5,724775	-0,155225	5,71173	-0,16827		

Mannose	**LOCAL**		**GLOBAL**			
Y	Y pred	Residual	Y pred	Residual	r =	0,856038
3,32	2,452991	-0,867009	4,20524	0,88524	k =	1,131052
1,36	1,674047	0,314047	2,58613	1,22613	L =	1,288224
1,19	1,673038	0,483038	1,52184	0,33184		
2,09	2,362242	0,272242	2,05857	-0,03143	SEP1 (L)	1,1619
3,05	2,239027	-0,810973	2,01118	-1,03882	SEP2 (G)	1,0995
1,38	2,711361	1,331361	1,75354	0,37354		
1,38	2,788861	1,408861	1,55179	0,17179	lower	0,820318
1,43	0,6965432	-0,7334568	0,88127	-0,54873	upper	1,361334
2,77	2,275058	-0,494942	2,1513	-0,6187		
12,65	12,35244	-0,29756	12,0715	-0,5785		
12,35	9,502433	-2,847567	9,71557	-2,63443		
13,28	13,83537	0,55537	13,96161	0,68161		

(continues)

11,29	13,75408	2,46408	13,59212	2,30212
12,07	10,92613	-1,14387	10,87904	-1,19096
12,88	11,37184	-1,50816	11,04568	-1,83432
12,52	11,61459	-0,90541	11,59572	-0,92428
10,93	11,20363	0,27363	10,73389	-0,19611
7,18	6,792759	-0,387241	7,03342	-0,14658
10,87	9,872519	-0,997481	9,84443	-1,02557
11,38	11,6472	0,2672	11,37951	-0,00049