PREDICTION OF EXTRACTIVES AND LIGNIN CONTENTS OF ANATOLIAN BLACK PINE (*Pinus nigra Arnold. var pallasiana*) AND TURKISH PINE (*Pinus brutia Ten.*) TREES USING INFRARED SPECTROSCOPY AND MULTIVARIATE CALIBRATION

A Thesis Submitted to
the Graduate School of Engineering and Science of
İzmir Institute of Technology
in Partial Fulfillment of the Requirements for the Degree of

MASTER OF SCIENCE

in Chemistry

by
İbrahim KARAMAN

July 2008
İZMİR

ACKNOWLEDGEMENTS

This thesis could not have been written without Dr. Durmuş ÖZDEMİR who not only served as my supervisor but also encouraged and challenged me throughout my academic program. He patiently guided me through the evaluation period of the thesis, never accepting less than my best efforts. I thank him so much.

Also, I would like to thank Betül ÖZTÜRK for her constructive comments on this thesis and for being an exceptional collaborator.

Special thanks to Dr. Birol ÜNER for participating as committee member and Hüseyin TANRIVERDİ both for their helps in wet chemistry analyses.

Thanks to Dr. Şerife YALÇIN for reviewing my work as committee member.

I am pleased to pay my thanks to Technological Research Council of Turkey (TÜBİTAK) for funding the project (105O524) and providing scholarship; İzmir Institute of Technology (İYTE) and Süleyman Demirel University for giving chance to work on this project.

Additionally, I would like to acknowledge all my colleagues in İYTE, Chemistry Department for their friendship over two years period of my academic life.

Thanks to Çağrı ÜZÜM and Caner ÜNLÜ who encouraged me to join in İYTE, Chemistry Department.

I would like to express my thanks to my lovely family for their motivation, continuous support and prayers. My father, Haydar, is the person who always supported me for my educational decisions since I was a child. My mother, Meral, is the one who raised me with her endless love. My sister, Nihal, thanks to you for being supportive and caring.

Finally, I want to express my appreciation to my beloved Ekin whose dedication, love and confidence in me have taken load off my shoulder. I owe her so much.

ABSTRACT

PREDICTION OF EXTRACTIVES AND LIGNIN CONTENTS OF ANATOLIAN BLACK PINE (*Pinus nigra Arnold. var pallasiana*) AND TURKISH PINE (*Pinus brutia Ten.*) TREES USING INFRARED SPECTROSCOPY AND MULTIVARIATE CALIBRATION

Determination of quality parameters such as extractives and lignin contents of wood by wet chemistry analyses takes long time. Near-infrared (NIR) and mid-infrared (MIR) spectroscopy coupled with multivariate calibration offer fast and nondestructive alternative to obtain reliable results. However, due to complexity of multi-wavelength spectra, wavelength selection is generally required. Turkish pine and Anatolian black pine are the most growing pine species in Turkey. Forest products industry has widely accepted use of these trees because of their ability to grow on a wide range of sites and their suitability to produce desirable products. Determination of extractives and lignin contents of wood provides information to tree breeders when to cut and on how much chemical is needed in pulping and bleaching process. In this study, 58 samples of Turkish pine and 51 samples of Anatolian black pine were collected to investigate the correlation between NIR and MIR spectra of these samples and their extractives and lignin contents which were determined with reference methods. Genetic inverse least squares (GILS) was used for multivariate calibration. Standard error of calibration (SEC) values were less than 1.86% (w/w) for lignin and 1.19% (w/w) for extractives whereas standard error of prediction (SEP) values were less than 3.81% (w/w) for lignin and 2.04% (w/w) for extractives. Resulting R^2 values for calibrations were larger than 0.8. Classification for Turkish pine and Anatolian black pine samples was performed by genetic algorithm based principal component analysis (GAPCA) and these two pine species were classified by using NIR and MIR spectra.

ÖZET

ANADOLU KARAÇAMI (*Pinus nigra Arnold. var pallasiana*) VE KIZILÇAM (*Pinus brutia Ten.*) AĞAÇLARINDAKİ EKSTRAKTİF MADDE VE LİGNİN MİKTARLARININ INFRARED SPEKTROSKOPİSİ VE ÇOK DEĞİŞKENLİ KALİBRASYON KULLANILARAK TAHMİNİ

Odun örneklerinin ekstraktif madde ve lignin miktarları gibi niteliksel parametrelerinin ıslak kimyasal analizlerle belirlenmesi uzun zaman almaktadır. Çok değişkenli kalibrasyonla birleştirilmiş yakın-infrared (NIR) ve orta-infrared (MIR) spektroskopisi güvenilir sonuçlar elde etmede hızlı ve tahribatsız bir alternatif sunar. Ancak, çok dalgaboylu spektrumların karmaşıklığından dolayı genelde dalgaboyu seçimi gerekir. Kızılçam ve Anadolu karaçamı Türkiye'de en çok yetişen çam türleridir. Orman ürünleri sanayisi bu ağaçların kullanımını geniş alanlarda yetişebilirliği ve arzu edilen ürünlerin üretilmesine uygunluğu nedeniyle geniş ölçüde kabul etmektedir. Ekstraktif madde ve lignin miktarlarının belirlenmesi ağaç yetiştiricilerine ne zaman kesim yapacakları, hamurlaştırma ve ağartma işlemlerinde ne kadar kimyasal gerektiği hakkında bilgi sağlar. Bu çalışmada, NIR ve MIR spektrumları ile referans yöntemlerle belirlenmiş olan ekstraktif madde ve lignin miktarları arasındaki bağıntıyı araştırmak için 58 kızılçam örneği ve 51 Anadolu karaçamı örneği toplanmıştır. Çok değişkenli kalibrasyon için genetik ters en küçük kareler (GILS) kullanılmıştır. Kalibrasyon standart hata (SEC) değerleri lignin için %1,86 (w/w)'dan ve ekstraktif madde için %1,19 (w/w)'dan az elde edilmiştir. Tahmin standart hata (SEP) değerleri lignin için %3,81 (w/w)'den ve ekstraktif madde için %2,04 (w/w)'ten az elde edilmiştir. Kalibrasyonlar sonucunda R^2 değerleri 0,8'den büyük belirlenmiştir. Kızılçam ve Anadolu karaçamı örneklerinin sınıflandırılmasında genetik algoritmalara dayalı temel bileşenler analizi uygulanmış ve bu iki çam türü NIR ve MIR spektrumları kullanılarak sınıflandırılmıştır.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

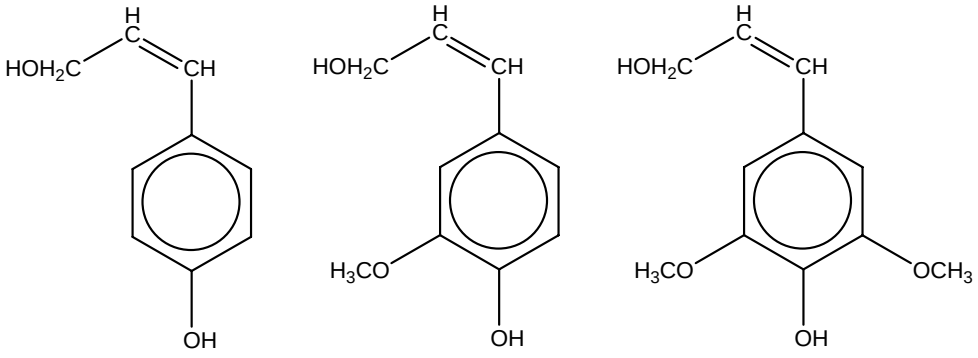
Table **Page**

CHAPTER 1

INTRODUCTION

Wood is a composite material and is composed of cellulose, lignin, hemicelluloses, extractives, and ash. Hence wood is also described as a lignocellulosic material. Cellulose is a linear polymer consisting of repeating (1→4)-β-D-glucopyranose units. It is the back-bone structure of the wood. Chemically it is highly strong against degradation due to hydrogen bonding between cellulose molecules. Hemicelluloses are polysaccharides made up of different carbohydrates including mannose, galactose, glucose, and 4-O-methyl-D-glucuronic acid, xylose, and arabinose. They fill the spaces in the wood fiber and enhance the strength of paper and pulp yield. Also they are more vulnerable than cellulose to chemical degradation (Biermann 1996).

Lignin is a highly branched complex polymer consisting of phenyl propane units. It has a high molecular weight and it cannot be measured easily. There are three basic lignin monomers shown in Figure 1.1, but not all the woods have all these three monomers. It depends on the nature of the wood. Only one or two of them may exist (Biermann 1996).



p-coumaryl alcohol | coniferyl alcohol | sinapyl alcohol

Figure 1.1. Basic lignin monomers or precursors for woods

Extractives are compounds which vary in molecular weights and are soluble in organic solvents and water. They contribute characteristic properties like color, odor,

taste and decay resistance to wood. Examples of extractives are terpenes (polymerized phosphate isoprene units), triglycerides, fatty acids, and phenolic compounds. Ash contains cations such as sodium, potassium, calcium and anions such as carbonate, phosphate, silicate, sulfate, chloride, etc. following combustion process of wood (Biermann 1996).

Trees are classified as hardwoods and softwoods. Woods of conifers like pine are called softwoods and woods of broad-leaved trees like oak are called hardwoods. The terms hard and soft do not come from the hardness of the wood, because a hardwood species can be softer than a softwood species. The difference comes from the cellular structure of the wood. For example, hardwood species have vessels that are used to carry sap from body through leaves even softwoods do not have those vessels. Softwoods have simpler structure than hardwoods (Wood Growth and Structure 2007). Chemical contents also differ with respect to being hardwood or softwood. Table 1.1 shows the comparison between North American hardwoods and softwoods with respect to chemical compositions (Biermann 1996).

Table 1.1. Chemical compositions of North American woods in weight percent unit (Source: Biermann 1996)

	Hardwoods (%w/w)	**Softwoods (%w/w)**
Cellulose	40-50	45-50
Hemicelluloses	17-35	25-35
Lignin	18-25	25-35
Extractives	1-5	3-8
Ash	0.4-0.8	0.2-0.5

As it is seen from Table 1.1, wood consists of cellulose and hemicelluloses in 60-80 weight percent unit range. The remaining part which mostly consists of lignin and extractives makes significance since contents of lignin and extractives are quite important in pulping and papermaking industry. High pulp yield requires high cellulose content but low extractives and lignin contents. During pulping process, lignin and extractives must be separated from cellulose fibers to obtain a high quality pulp (Poke, et al. 2005).

Determination of extractives and lignin contents provides information on how much chemicals are needed in pulping process. Besides, in tree breeding programs, sylvicultural treatments, which are applied for better tree growth, cause modifications and these trees may have different chemical properties than natural grown trees. Furthermore, wood samples obtained from the trees located in different regions show different properties in chemical compositions, morphology, etc. (Zobel and van Bujitenen 1989). Variations in chemical contents have consequences such as brightness of paper on the final products.

In Turkey, *Pinus brutia Ten.* (Turkish pine) and *Pinus nigra Arnold. var pallasiana* (Anatolian black pine) are the most growing pine species. Turkish pine has a rotation period around 60 – 80 years and Anatolian black pine has around 120 years. Because of fast rotation period, these pine species are widely accepted in forest products industry since they are very suitable for production of window door panels, floor coverings, etc. They are also used in papermaking and construction. Therefore, determination of extractives and lignin contents of Turkish pine and Anatolian black pine trees is important and currently used methods for this purpose are time consuming and costly processes. There are standard methods which Technical Association of the Pulp and Paper Industry (TAPPI) offers and they are based on wet chemistry. Rapid, inexpensive and nondestructive methods to measure extractives and lignin contents are the focus of researchers.

Recently, near-infrared (NIR) spectroscopy is being used for measuring chemical properties such as lignin, glucose, xylose, mannose, galactose, cellulose, extractives contents and mechanical properties such as annual ring widths, wood density, average fiber length, fiber length distributions, wood strength, stiffness, microfibril angle of wood species (Jones, et al. 2006, Poke and Raymond 2006, Hauksson, et al. 2001, Kelley, et al. 2004, Yeh, et al. 2004, Yeh, et al. 2005). These studies were performed using either diffuse reflectance or transmittance modes. Various multivariate calibration methods were used to analyze spectra and to construct calibration models. In some studies, mid-infrared (MIR) spectroscopy is used for rapid determination of chemical compositions of wood species (Schultz, et al. 1985, Nuopponen, et al. 2006, Dang, et al. 2007).

The major aim of this study is to construct multivariate calibration models for determining extractives and lignin contents of Turkish pine and Anatolian black pine trees by using diffuse reflectance near-infrared and mid-infrared spectroscopy.

Therefore, one can save time, effort and money by using this type of calibration models for different wood species. In addition, this study is also intended to construct a multivariate classification model which distinguishes Turkish pine and Anatolian black pine trees with respect to their spectra.

CHAPTER 2

INFRARED SPECTROSCOPY

2.1. Infrared Region

The infrared (IR) region of the electromagnetic spectrum lies over a wide wavelength range starting from around 780 nm to $1x10^6$ nm. This region is divided into three sub-regions due to varying applications and instrumentations. Table 2.1 shows the three distinctly different infrared regions.

Table 2.1. Infrared spectral regions
(Source: Sherman 1997)

Region	Wavelength Range, µm	Wavenumber, cm^{-1}
Near (NIR)	0.78 – 2.5	12,800 – 4,000
Middle (MIR)	2.5 – 50	4,000 – 200
Far (FIR)	50 – 1000	200 – 10

Mid-infrared (MIR) spectroscopy is widely used as a tool for both qualitative and quantitative analysis. The most common use of MIR spectroscopy is to identify organic, inorganic and biochemical species (Sherman 1997, Griffiths 1978, Koenig 1975). Especially the region around between 900 cm^{-1} and 1300 cm^{-1} which is called fingerprint region is highly specific to an individual compound. For instance, MIR spectra of 1-propanol and 2-propanol are very similar but show differences in fingerprint region. Far-infrared (FIR) spectroscopy is used for analysis of organic, inorganic, and organometallic compounds involving heavy atoms. It gives information about conformation and lattice dynamics of samples. Near-infrared (NIR) spectroscopy offers quantitative analysis of certain species such as water and hydrocarbons with low molecular weights without consumption or destruction of the sample. Therefore, NIR spectroscopy has become a popular method for simultaneous chemical analysis and is being studied widely in different fields such as process monitoring (DeThomas, et al.

1994), biotechnology (Arnold, et al. 2000), and pharmaceutical industry (Tran, et al. 2004).

Infrared radiation provides rotational and vibrational motion to a molecule. Since rotational motion has low energy, FIR region may be used for rotational spectroscopy. The MIR region is used to study fundamental vibrations (change in vibrational quantum number is ±1) and rotational – vibrational structures. NIR region in which radiation with higher energy lies is commonly used to study overtone (change in vibrational quantum number can be ±2, ±3, ±4…) and combination vibrations.

2.2. Infrared Instruments

An IR instrument contains a source of infrared radiation, a sample container which should be infrared transparent, a wavelength selecting device, a detector and a signal processor, consecutively. Nernst glower (ZrO_2+Y_2O_3), Nichrome wire (Ni+Cr), and Globar (SiC) can be used as IR sources depending on the type of application since they cover certain sub-regions of IR region. As a sample holder, mostly quartz cells are used in the NIR region and potassium bromide (KBr) is used in the MIR and FIR regions. Wavelength selecting devices will be discussed later because all IR instruments don't have wavelength selectors and they will be classified with respect to this. Thermal detectors are used in the IR region such as thermocouples and bolometers. A signal processor is an electronic device that amplifies the signal from the detector. The last component of an IR instrument is a readout device (Skoog, et al. 1998).

Commercially there are three types of instruments for infrared absorption measurements. These are dispersive instruments, multiplex instruments and non-dispersive instruments (Skoog, et al. 1998).

A dispersive instrument has a monochromator with a grating element to disperse the radiation coming from the source into its wavelengths and it is used as a wavelength selecting device. It is mostly designed double-beam, that is, incoming IR radiation is split into two beams in order to pass through the reference and sample materials. By this way signal is amplified and interferences of air during the analysis are prevented. A representative figure for a dispersive instrument is shown in Figure 2.1 (Smith 1996).

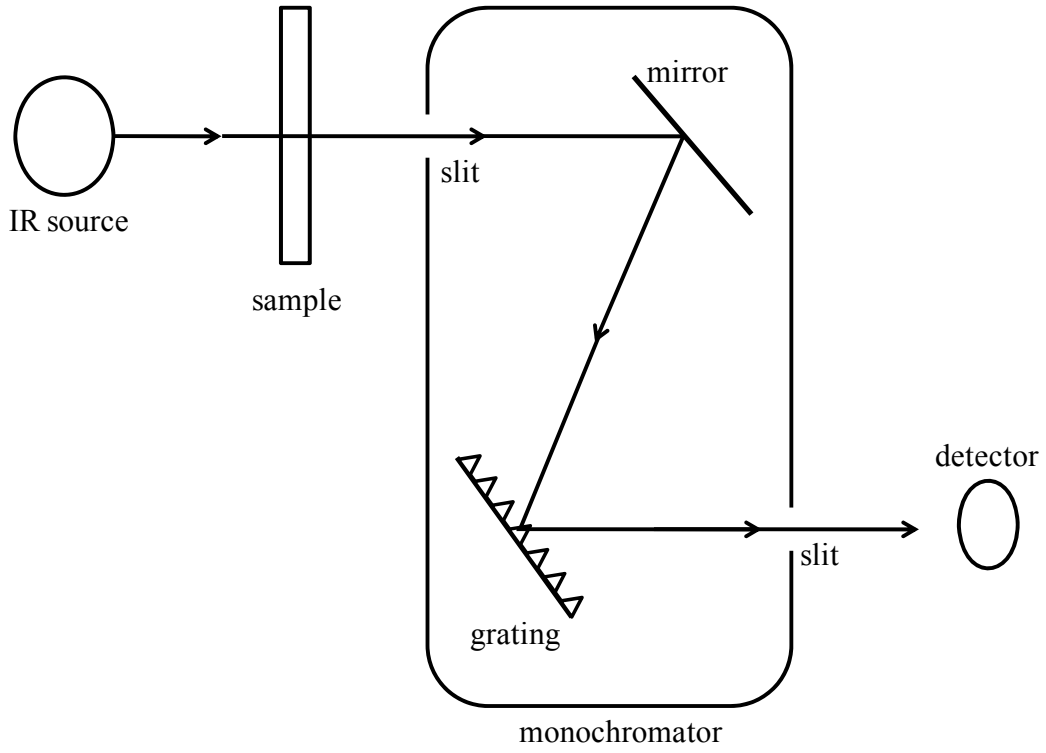


Figure 2.1. A schematic representation for a dispersive instrument

The most commonly used type of multiplex instruments is Fourier transform (FT) instruments. FT instruments don't have grating element to disperse the light. This feature allows high speed measurement and to obtain the full spectrum in one scan. In addition, increasing number of scans enhances signal-to-noise ratio of the spectrum. FT instruments are generally based on the Michelson interferometer. In Michelson interferometer, incoming radiation passes through a beam splitter and separated into two. One of the beams goes through a fixed mirror and the other one through a moving mirror by the help of beam splitter. The reflected beams from the mirrors combine constructively and destructively on the beam splitter. Then this radiation passes through the sample and goes to the detector. The detector measures the variation of light intensity of IR radiation with optical path difference as a sinusoidal wave. Light intensity versus optical path difference forms an interferogram. At the end, this interferogram is converted to a single beam spectrum by Fourier transformation. Figure 2.2 is the schematic representation of an FT instrument (Smith 1996).

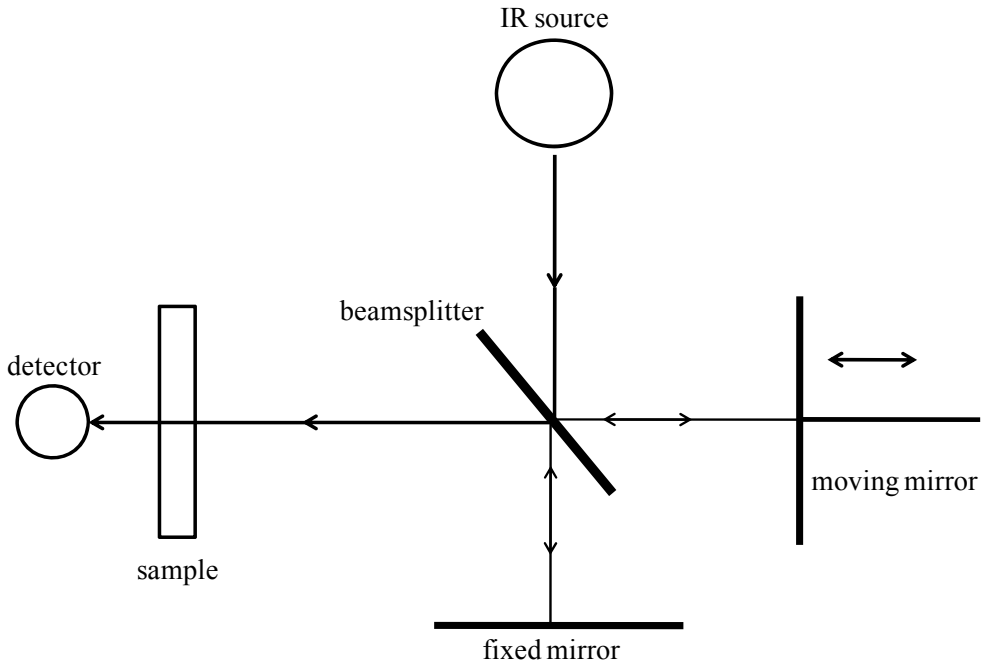


Figure 2.2. A schematic representation for an FT instrument

Non-dispersive instruments are filter or non-dispersive photometers. They are designed for quantitative analysis. Generally they are not complex, easy to use and not expensive compared to the instruments mentioned above (Skoog, et al. 1998).

2.3. Diffuse Reflectance Infrared Spectroscopy

Some of solid samples such as polymer films, food products, agriculture products and rubbers are inconvenient for absorption measurements most of the time due to sampling difficulties in transmission techniques. Diffuse reflectance infrared spectra are generally similar to that of corresponding infrared spectra and have the same chemical information. The differences are observed on the intensities of the peaks. Reflectance spectra can be used both qualitatively and quantitatively in MIR region. NIR region is often used quantitatively (Smith 1996).

There are mainly two types of reflection of radiation as illustrated in Figure 2.3. Specular reflection takes place when the angle of incident beam is equal to the angle of reflected beam. Diffuse reflection takes place when the incident beam coming with a constant angle is reflected through all directions. Thus one can say that diffuse reflection occurs on rough, specular reflection occurs on smooth surfaces (Smith 1996).

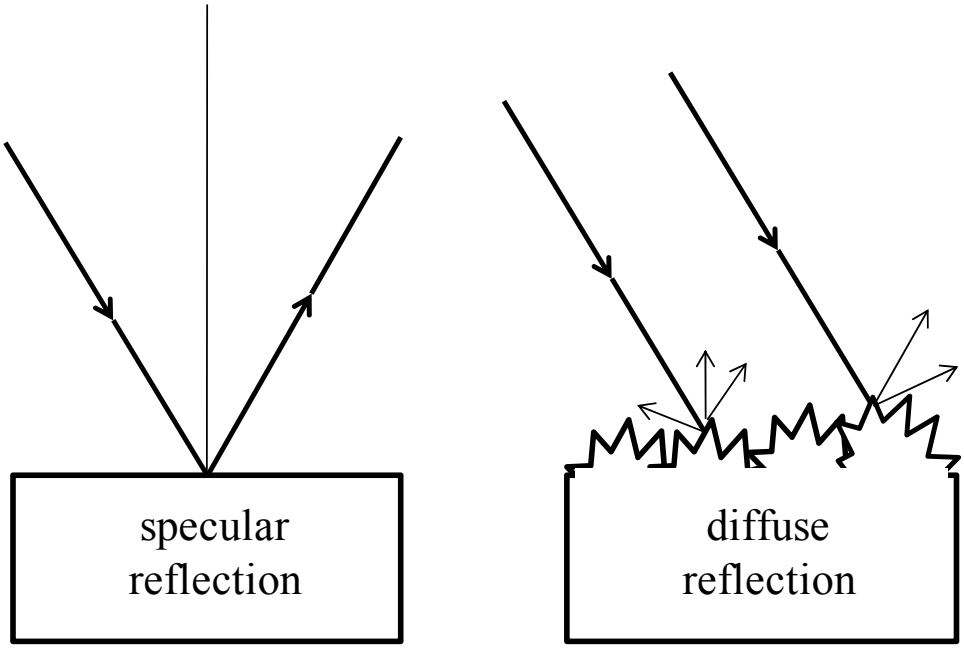


Figure 2.3. Schematic representation of specular and diffuse reflection processes

Since, in this study, diffuse reflectance is used, a detailed description of the method is given here. Diffuse reflectance (DR) spectroscopy is generally applied on powders and other solid samples. Generally, the sample preparation is at minimum and in some cases no sample preparation is required. It became more common after

development of FT instruments because signals reflected from the powder samples are very weak to be finely detected by a dispersive instrument. To obtain diffuse reflectance spectra, firstly diffuse reflectance accessory is attached to the sample compartment of the FT instrument (either MIR or NIR). The accessory consists of mirrors to focus the IR beam onto the sample. The sample is placed in a sample cup designed specifically on the focal point of the spherical focusing mirror. The radiation coming through the sample is diffusely reflected and collected by a second spherical mirror. Then, the collected light is focused on the detector. In Figure 2.4, optical diagram of a DR accessory is shown. The intensity of the diffusely reflected light is very sensitive on the packing density of the particles and the uniformity of the particle size. To overcome these effects, sample can be mixed with KBr during MIR measurements and can be meshed to get uniform particle size (Smith 1996).

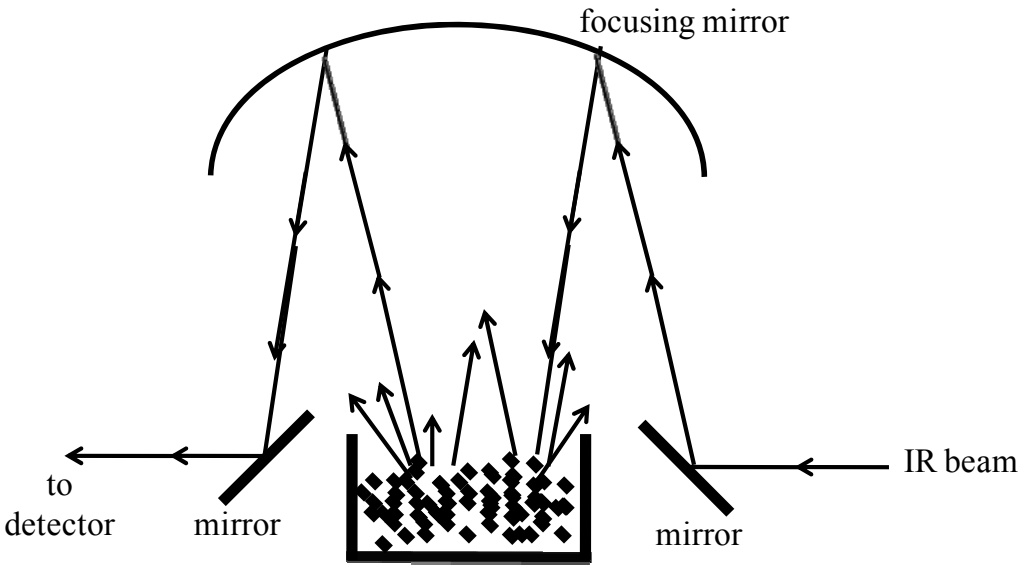


Figure 2.4. Schematic representation of a diffuse reflectance accessory

CHAPTER 3

MULTIVARIATE ANALYSIS METHODS

Chemometrics is the science of relating measurements made on a chemical system or process to the state of the system via application of mathematical or statistical methods according to The International Chemometrics Society (ICS). It is now recognized as a branch of analytical chemistry. There are several chemometric techniques for collecting good data (optimization of experimental parameters, design of experiments, calibration, signal processing) and for getting information from these data (statistics, pattern recognition, principal component analysis). The aim of using chemometrics is to combine the chemometric methods and their application in chemistry (Wikipedia contributors 2008). In this chapter, the focus is on calibration and classification techniques which are used in this study.

3.1. Calibration Methods

3.1.1. Overview

Calibration is a process that a model is constructed to obtain a relation between the output of an instrument and properties of samples. Prediction is a process that the constructed model is used to predict the properties of a sample which its instrument response is given. The model is constructed by measuring instrument responses and concentration levels of certain chemical contents of the samples. Then, this model is used to predict the concentration of an unknown content sample in the future (Beebe, et al. 1998). In this study, instrument responses refer to MIR and NIR spectra, and concentration levels refer to extractives and lignin contents of wood meal samples.

In many applications, one response is taken from an instrument and that response is related to the concentration of the chemical component of a sample. This method is called *univariate calibration* because number of instrumental response for each sample is just one. The process that relates multiple instrument responses to one or more properties of a sample is called *multivariate calibration*. The sample can be multi-

component and the aim is to predict the concentrations of the components from, for example, UV-Vis absorption measurements (Beebe, et al. 1998).

3.1.2. Univariate Calibration

This type of calibration has been widely used for years in chemical analysis. In an absorption or chromatography study, absorption at a wavelength or a peak area is taken and its relation to the concentration of a sample is then modeled. If the relation is considered as linear, there are two options.

- Classical calibration
- Inverse calibration

These models are based on Beer's law in which absorbance at a wavelength is directly proportional to the absorptivity coefficient, light path length and concentration.

3.1.2.1. Classical Calibration

This type of calibration considers absorbance at a spectroscopic wavelength of a chromatographic peak area as a function of concentration. The general formula of classical calibration is:

$$\mathbf{a} \approx \mathbf{c} \cdot s \tag{3.1}$$

where $\mathbf{a}$ is the vector of absorbances at one wavelength for a number of samples and $\mathbf{c}$ is the vector of corresponding concentrations. The scalar coefficient s is related with these parameters and can be determined by the following equation:

$$s \approx \left(\mathbf{c}' \cdot \mathbf{c}\right)^{-1} \cdot \mathbf{c}' \cdot \mathbf{a} \tag{3.2}$$

where the $\mathbf{c}'$ is the transpose of the concentration vector.

After determining s, the prediction model for an unknown is constructed as:

$$\hat{c} \approx \hat{a} / s \tag{3.3}$$

where the hat symbol for scalars a and c refer to prediction.

To check the prediction model's quality, residuals are calculated. Residuals or errors are the difference between the observed and predicted concentration values.

$$e = c - \hat{c} \tag{3.4}$$

The less the residuals mean the better the model (Brereton 2000).

3.1.2.2. Inverse Calibration

The aim of using calibration models is to predict concentration from a spectrum or a chromatogram. Errors in the classical calibration are due to instrumental response. However, developments in the reproducibility of instruments made the instruments reliable because concentration values are mostly determined gravimetrically or by dilutions, so source of error is larger than instrumental error in this case. More convenient approach can be that source of error is due to the concentration. In Figure 3.1, difference between errors due to instrument and concentration is represented.

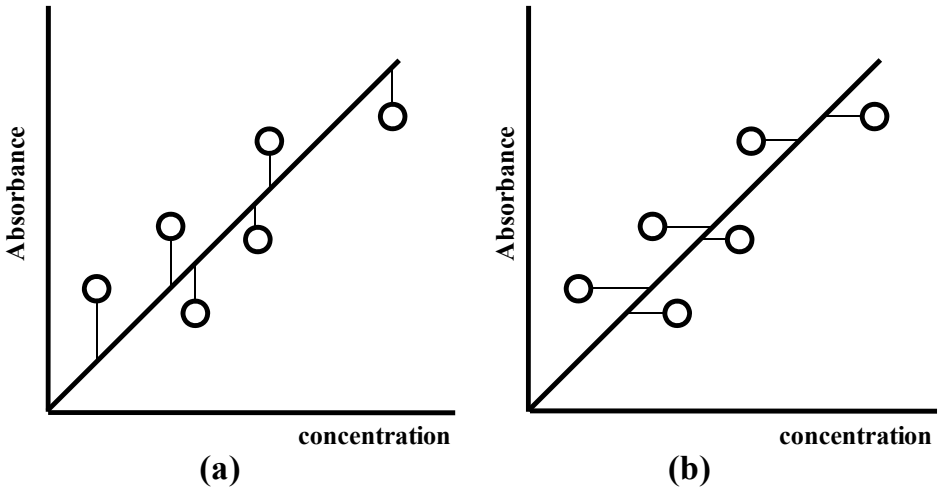


Figure 3.1. Error distributions in (a) classical and (b) inverse calibration models

Then, inverse calibration can be modeled as:

$$\mathbf{c} \approx \mathbf{a} \cdot b \tag{3.5}$$

where b is a scalar coefficient and is approximately inverse of s because each model makes assumptions on errors in a different way. b can be determined according to the following formula:

$$b \approx \left(\mathbf{a}' \cdot \mathbf{a}\right)^{-1} \cdot \mathbf{a}' \cdot \mathbf{c} \tag{3.6}$$

and prediction of an unknown sample can be performed easily by using b (Brereton 2003).

$$\hat{c} \approx \hat{a} \cdot b \tag{3.7}$$

Chemometricians prefer to use inverse models but traditional analytical chemistry books mostly represent classical models as calibration methods. For a good data set, both models should give comparable predictions. If not, other factors such as an intercept, non-linearities, outliers or noise in the spectra should be taken into consideration and the model must be modified (Brereton 2000).

3.1.3. Multivariate Calibration

Multivariate calibration is applicable to determination of major and also minor components of mixtures and for various instrument types. The necessity for sample preparation is reduced because selective input measurements are not needed any more. Actually output results must be selective. Therefore multivariate calibration can give rise to the development of new analytical instruments. In addition, it can enhance the analytical capacity and reliability of traditional instruments (Martens and Naes 1989).

Multivariate calibration has some advantages over univariate calibration.

1) Simultaneous analysis of multiple components in a sample is possible. By univariate method, there has to be one measurement for each component. So, spent time will be more (Beebe, et al. 1998).
2) Precision in the prediction can be enhanced by repeating a measurement and calculating the mean. This will cause consequence of reduction in the standard deviation of the mean. This is called signal averaging (Beebe, et al. 1998).
3) Multivariate calibration has fault-detection capabilities. That means unknown interferences in the sample can be overcome by multivariate calibration. In univariate calibration, the presence of interferences may cause wrong prediction of concentration of analyte. To avoid this problem, physical separation of analyte from interfering material or using selective measurements is needed and this means necessity of more effort. Figure 3.2

demonstrates how the calibration curve is affected by the interferences. By multivariate calibration, nonlinearities caused by the interferences can be reduced by selecting more variables and chance of obtaining better calibration curve can be increased. Therefore, time and effort spent to remove interferences physically is respectably decreased (Öztürk 2003).

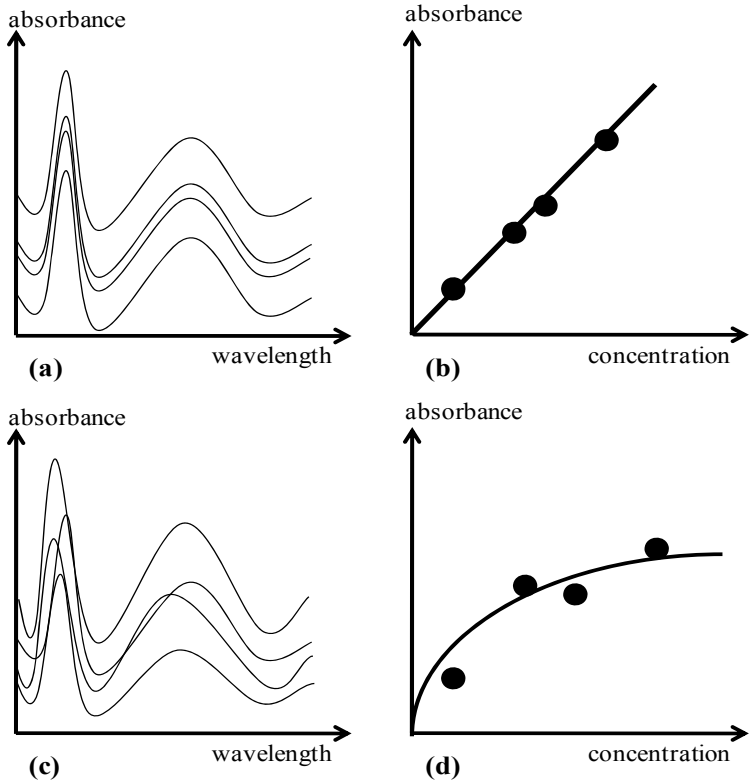


Figure 3.2. (a) Spectra of a sample in different concentrations which has no interference and its calibration curve (b) by univariate calibration; (c) spectra of a sample in different concentrations which has interfering materials and its calibration curve (d) by univariate calibration

In multivariate calibration, the equations can be developed in two ways. First one is that, as in the classical calibration case, absorbance is a function of concentration. Second one is that, as in the inverse calibration case, concentration is a function of absorbance. Difference from univariate calibration is the usage of absorbance values in the full spectrum of one sample. So, the absorbance vector in univariate calibration becomes a matrix. Also more than one component can be used and thus concentration vector becomes a matrix, too.

In this study, genetic inverse least squares method is used. Before discussing this method, it is necessary to explain classical least squares and inverse least squares methods as an introduction to the multivariate calibration methods.

3.1.3.1. Classical Least Squares (CLS)

Taking into consideration Beer's law, classical least squares method is modeled by the following equation:

$$\mathbf{A} = \mathbf{C} \times \mathbf{K} + \mathbf{E} \tag{3.8}$$

where $\mathbf{C}$ is the matrix which consists of concentrations of multi-component samples and $\mathbf{E}$ is the error matrix. If there is only one component, it is denoted as a vector $\mathbf{c}$. $\mathbf{A}$ is the matrix which consists of absorbance values of the samples at different wavelengths. Each row of $\mathbf{C}$ and $\mathbf{A}$ and correspond to one sample, each column represents different component and different absorption values, respectively. $\mathbf{K}$ is the matrix of absorptivity coefficients multiplied by path length. Each member of this matrix corresponds to absorptivity coefficient of an absorption value at a certain wavelength. $\mathbf{K}$ matrix can be determined by the following formula:

$$\mathbf{K} = \left(\mathbf{C}' \cdot \mathbf{C}\right)^{-1} \cdot \mathbf{C}' \cdot \mathbf{A} \tag{3.9}$$

To perform prediction, an unknown sample spectrum is measured ($\hat{\mathbf{a}}$). Given $\hat{\mathbf{a}}$ and $\mathbf{K}$, concentration can be predicted by using simple matrix algebra:

$$\hat{\mathbf{c}} = \hat{\mathbf{a}} \cdot \mathbf{K}' \cdot \left(\mathbf{K} \cdot \mathbf{K}'\right)^{-1} \tag{3.10}$$

Here, the notations of prediction elements are vector, not scalar as in the univariate calibration, because there are more than one component and there are more than one absorbance value in one unknown sample. The residual is the difference between the reference and predicted concentration values.

$$\mathbf{e} = \mathbf{c} - \hat{\mathbf{c}} \tag{3.11}$$

In summary, the CLS method can be applied to simple systems where all of the pure-component spectra can be measured. In order to construct the CLS model, the

pure-component spectra are measured for each analyte in the sample. These are utilized to form spectral matrix and the model is then constructed. This calibration model is used to predict the concentrations of components in unknown samples.

CLS method has advantages and disadvantages depending on the purpose. First of all, strict assumptions must be obeyed for the method to work well. This means that the measurements are linear with concentration, obligation of linear additivity, and all the components in the sample must be known. The simplicity to describe the model is an advantage. Only a small number of samples are needed to construct the calibration model. Since many variables are used, it is possible to overcome overlapping problems (Beebe, et al. 1998).

3.1.3.2. Inverse Least Squares (ILS)

In some cases, CLS may not work because the system of interest is not simple or it may not be possible to obtain the pure spectra of all the analytes in the unknown samples. Practically it is not clear that either inverse or classical method is optimal. There have been approaches on this purpose and they give some guidance (Haaland and Thomas 1988).

The relationship between the measurements and concentrations is modeled as in CLS but in this case the concentrations are treated as a function of absorbance values, as shown in the following equation:

$$\mathbf{C} = \mathbf{A} \times \mathbf{P} + \mathbf{E} \tag{3.12}$$

where $\mathbf{C}$ is the concentration matrix, $\mathbf{A}$ is the absorbance matrix and $\mathbf{E}$ is the error matrix as in CLS. The matrix $\mathbf{P}$ contains the model coefficients and can be determined by:

$$\mathbf{P} = \left(\mathbf{A}' \cdot \mathbf{A}\right)^{-1} \cdot \mathbf{A}' \cdot \mathbf{C} \tag{3.13}$$

A predicted concentration of a multi-component sample can be obtained by:

$$\hat{\mathbf{c}} = \hat{\mathbf{a}} \cdot \mathbf{P} \tag{3.14}$$

The residual is, as in the CLS model, the difference between the reference and predicted concentration values.

$$\mathbf{e} = \mathbf{c} - \hat{\mathbf{c}} \tag{3.15}$$

In summary, ILS can be used to construct accurate calibrations when just knowing the concentrations of analytes in the sample. That means there is no need to know all of the components in the sample. To use ILS, one must select as many variables (wavelengths in our case) as there are sources of variation in the system instead of using the full spectra. Weaknesses of ILS are that it has limited outlier detection and there is no efficient method for optimal wavelength selection for predictive models. Also collinearity between the absorbance values causes problems the validation of the model because it prevents stabilization of the predictions against noise in absorbances. So, it is very important to select the best set of wavelengths to use in the construction of calibration (Beebe, et al. 1998).

3.1.3.3. Genetic Inverse Least Squares (GILS)

This method is a modified version of ILS in which genetic algorithms (GA) are used as a tool for wavelength selection. GA are global search and optimization methods based on the principles of natural evolution and selection as developed by Darwin (Wang, et al. 1991). According to the Darwin's theory of evolution, individuals who fit better to the environment are more likely survive and breed, thus are able to pass their genetic information to their offspring. However, individuals who do not fit and unable to adapt will eventually be eliminated from the population. This process progresses slowly over a long period of time (or may never end) through generations and the species will evolve into better and fit forms. In the last couple of decades, scientists have been trying to take advantages of the natural evolutions as an improvement concept in the process of solving large-scale optimization problems. In the 1960's, biologists have begun to perform the simulation of genetic systems experiments with computer. The initial work in GA was done by Holland who developed a genetic algorithm in his research on adaptive systems in the early 1960's and is considered the father of the field (Gilbert, et al. 1997). Over the years, GA have attracted attention and have been applied to various global optimization problems in many areas including chemometrics (Fontain 1992, Cong and Li 1994, Wienke, et al. 1993, Hibbert 1993, Lucasius and Kateman 1991). In terms of calibration, there have been several applications of GA to wavelength selection (Lucasius, et al. 1994, Lucasius and

Kateman 1992, Paradkar and Williams 1997, Ozdemir, et al. 1998a, Ozdemir, et al. 1998b, Ozdemir and Williams 1999).

Computationally the implementation of a typical genetic algorithm is quite simple and consists of five basic steps including initialization of gene population, evolution of the population, selection of the parent genes for breeding and mating, crossover and mutation, and replacing the parents with their offspring. These steps have taken their names from the biological foundation of the algorithm. The implementation of a typical GA is shown in Figure 3.3.

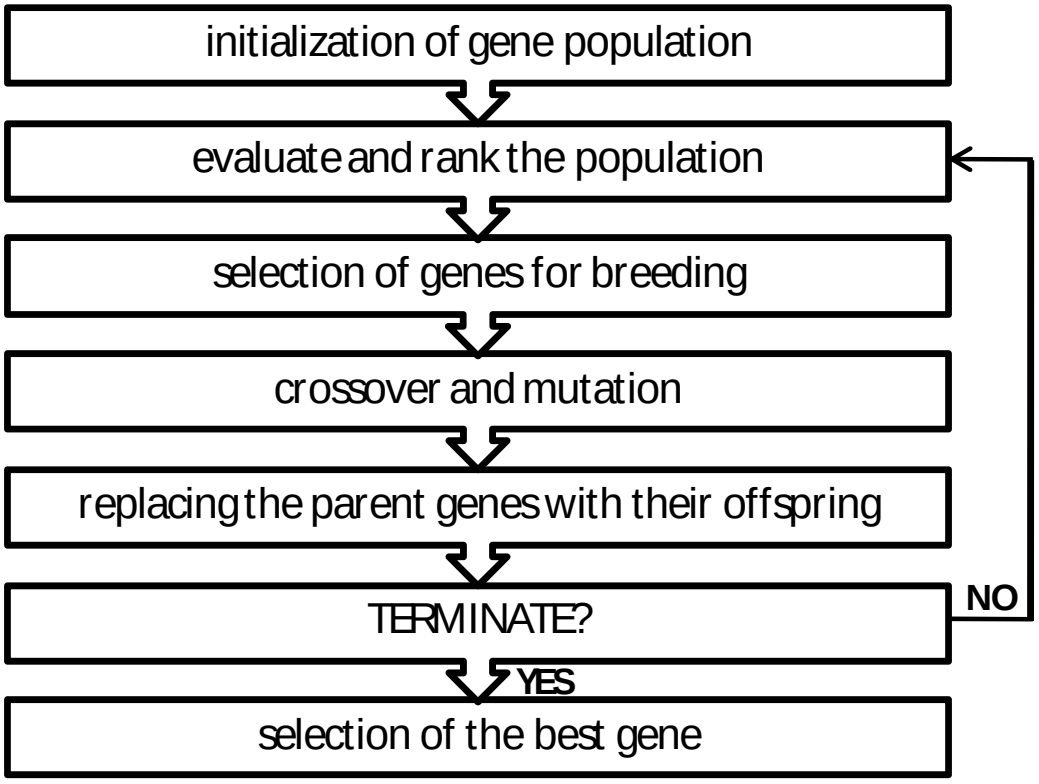


Figure 3.3. Flow chart of general genetic algorithm used in GILS

3.1.3.3.1. Initialization

A gene is defined as a potential solution to a given problem. The exact form of a gene may vary from application to application and depends upon the problem being investigated. The term population is used to describe the collection of individual genes in the current generation.

In the initial gene pool, a gene consists of absorbance values at randomly chosen wavelengths between a predefined lower and upper limit. An example of a gene is as the following:

$$S = [A_{8432} A_{6895} A_{5128}]$$

where S is so-called a gene, A is the absorbance measured at the indicated wavelength. The chosen absorbance value at one wavelength is a vector of samples. Concatenation of these vectors forms the new absorbance matrix which is the gene. Figure 3.4 shows the schematic representation of the gene for a wood sample. Then, the population is formed according to the number of genes initially entered as an input of the software.

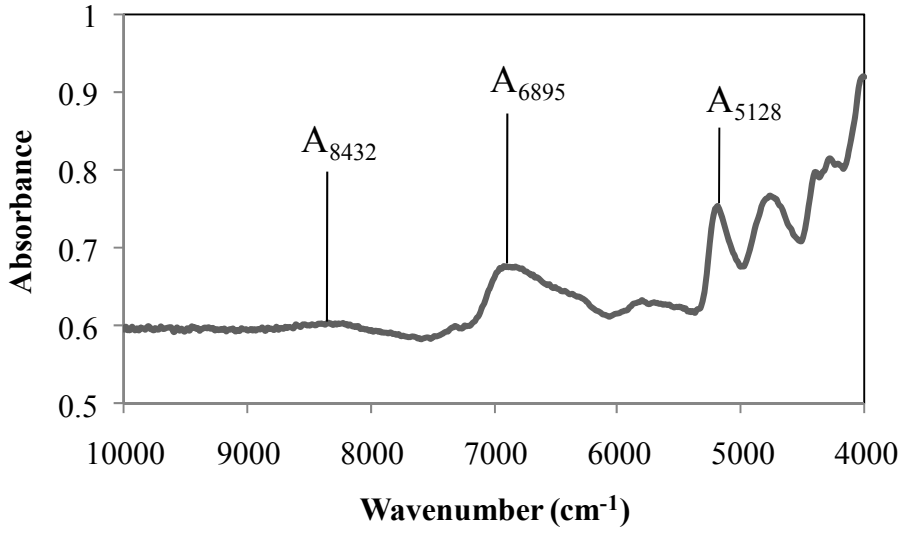


Figure 3.4. Illustration of a gene on a NIR spectrum of a wood sample

3.1.3.3.2. Evaluate and Rank the Population

In order to evaluate each gene's success in the prediction of analyte concentration, fitness function such as the reciprocal of standard error of calibration (SEC) is used. SEC is calculated from the ILS model in which absorbance values from the selected wavelengths are used to construct the model. SEC is calculated from the following equation:

$$SEC = \sqrt{\frac{\sum_{i=1}^{m} (c_i - \hat{c}_i)^2}{m-2}} \tag{3.16}$$

where c_i is the reference and $\hat{c}_i$ is the predicted values of concentration of i^{th} sample and m is the number of samples. Degrees of freedom is $m-2$ because when a linear model is assumed, there are only two parameters to be extracted which are the slope of the actual vs. reference concentration plot and the intercept. In each step, increase in the fitness value is targeted.

3.1.3.3.3. Selection of Genes for Breeding

This step involves the selection of the parent genes from the current population for breeding according to their fitness value. The goal is to give higher chance to those genes with higher fitness so that only the best performing members of the population will survive in the long run and will be able to transfer their information to next generations. Here, it is expected that the genes better suited for the problem will generate better off-springs. The genes with low fitness values will be given lower chance to breed and hence most of them will be unable to survive. There are number of selection methods that can be used for parent selection (Wang, et al. 1991). Top down selection is one of the simplest methods for parent selection. After genes are ranked in the current gene pool, they are allowed to mate in a way that the first gene mates with the second gene, third one with the forth one and so on. All the members of the current gene are given a chance to breed. Roulette wheel selection method, which is used in GILS, is the one where the chance of selecting gene is directly proportional to its fitness. In this method, each slot in the roulette wheel represents a gene. The gene with the highest fitness has the slot that has the largest area and the gene with the lowest fitness has the slot that has the smallest area. Therefore, when the wheel is rotated, there is a higher chance of selection for a gene with high fitness than for a gene with a low fitness. There will also be the genes which are selected multiple times and some of the genes will not be selected at all and will be thrown out from the gene pool. After all the parent genes are selected, they are allowed to mate top-down, whereby the first gene S_1 mates with the second gene S_2; S_3 with S_4 and so on until all the genes mate. Since no ranking is done for the roulette wheel selected genes, the genes with low fitness have a

chance to mate with better performing genes, thus resulting in an increased possibility of recombination.

3.1.3.3.4. Crossover and Mutation

The genetic algorithm does most of its work in the breeding/mating step. The step involves breaking the genes at random points and cross-coupling them as illustrated in the following example:

Consider S_1 and S_2 are parent genes which are to breed; S_3 and S_4 are their corresponding off-springs.

$$S_1 = [A_{4255} A_{5732} \oplus A_{9237} A_{4890}]$$
$$S_2 = [A_{5123} A_{8457} A_{9743} A_{7832} \oplus A_{8922}]$$
$$S_3 = [A_{4255} A_{5732} A_{8922}]$$
$$S_4 = [A_{5123} A_{8457} A_{9743} A_{7832} A_{9237} A_{4890}]$$

Here, the first part of S_1 is combined with the second part of S_2 to give S_3, likewise the second part of S_1 with the first part of S_2 to give S_4. This process is called single point crossover and it is the one used in GILS. The symbol $\oplus$ is used to indicate the separation of the genes and the place where crossover occurs. There are also other types of crossover methods such as two point crossover and uniform crossover, each having their advantages and disadvantages. In the uniform case, each gene is broken at every possible point and many combinations are possible in the mating step, thus resulting in more exploitation. However, it is more likely to destroy good genes. Single point crossover will not provide different off-spring if both parent genes are identical, which may happen in the roulette wheel selection, and broken at the same point. To avoid this problem, two points crossover, where each gene is broken in two points and recombined, can be used. Single point crossover generally does not disturb a good gene but it provides as many recombinations as other types of crossover schemes. Also mating can increase or decrease the number of base pairs in the off-spring.

Mutation, which introduces random deviations into the population, can be also introduced into the algorithm during the mating step at a rate of 1% as is typical in GA.

Replacing one of the wavelengths in an existing gene with a randomly generated new wavelength usually does this. However, it is not used in GILS in this study.

3.1.3.3.5. Replacing the Parent Genes by Their Off-springs

After crossover, the parent genes are replaced by their off-springs. The ranking process based on their fitness values follows the evolution step. Then the selection for breeding/mating starts again. This is repeated until a predefined number of iterations are reached.

At the end, the gene with the lowest SEC (highest fitness) is selected for model building. This model is used to predict the concentrations of component being analyzed in the validation set. The success of the model in the prediction of the validation set is evaluated using standard error of prediction (SEP) which is calculated as:

$$SEP = \sqrt{\frac{\sum_{i=1}^{m}\left(c_i - \hat{c}_i\right)^2}{m}} \tag{3.17}$$

where m is now, in this case, the number of validation samples.

3.1.3.3.6. Termination

The termination of the algorithm is done by setting predefined iteration number for the number of breeding/mating cycles. However no extensive statistical test has been done to optimize it, though it can also be optimized. Since the random processes are heavily involved in the GILS, the program is set to run predefined number of times for each component in a given multi-component mixture. The best run, i.e. the one generating the lowest SEC for the calibration set and at the same time obtained SEP for the validation set that is in the same range with SEC, is subsequently selected for evaluation and further analysis.

GILS has some major advantages over the classical univariate and multivariate calibration methods. First of all, it is quite simple in terms of the mathematics involved in the model building and prediction steps, but at the same time it has the advantages of the multivariate calibration methods with a reduced data set since it uses the full

spectrum to extract genes. By selecting a subset of instrument responses, it is able to eliminate nonlinearities that might be present in the full spectral region.

3.2. Classification and Clustering Techniques

Nowadays answering the question "Can a sample stated in a known class really belong to that class or not?" becomes valuable for the chemists who study in the area of qualitative analysis. Generally classification and clustering techniques are used to define the distribution of samples. Classification techniques are mainly divided into two different groups; supervised and unsupervised techniques. The extraction of useful data from analytical measurements and optimum information from analytical results are important objectives in terms of what we want as a result. Therefore to understand the differences between both techniques is important in the determination of which technique or techniques yields better result. If the classes or groups are known and the goal is to find in which class or group should be chosen for the investigated sample, supervised classification methods such as soft modeling of class analogy (SIMCA), linear discriminant analysis (LDA), and K-nearest neighbors (KNN) will be used. On the other hand, in unsupervised classification techniques, the chemical or physical variables of corresponding samples are not known clearly and therefore firstly the similarities of samples are found then the classes are developed. Principal component analysis (PCA) and hierarchical cluster analysis (HCA) are the most used techniques in unsupervised classification techniques (Beebe, et al. 1998).

General procedure in classification techniques are based on the variables of samples, then this variable data matrix is used to classify the samples according to their similarities or dissimilarities. After the developments in spectroscopy, the usage of spectral data matrix instead of variable data matrix becomes widespread. The critical point in this case is determining which part of spectrum contains the most useful information for interested samples. Therefore generally wavelength selection or optimization procedures are used to obtain necessary information. Genetic algorithms (GA) or moving window size are examples of the wavelength selection methods.

In this study principal component analysis was chosen as an unsupervised classification technique and genetic algorithm was imposed on the principal component analysis algorithm to select the wavelengths which have the most useful information.

Principal component analysis is a full spectral and soft modeling method which is based on the decomposition of data matrix into two separate and smaller matrices. These two kinds explain the relationships between the variables and the relationships between the objects. Also this division makes the dimensionality reduction for the large data matrix (Kowalski 1983). For instance, spectral data matrix contains hundreds of wavelengths with their corresponding absorbance values and it is really hard to visualize this data matrix in hundreds of dimensionality. As it is not possible to visualize dimensions larger than three, generally pictures or graphs that are used to explain the distributions of samples or variables should have three or less dimension in a space.

Singular value decomposition (SVD) and nonlinear iterative partial least squares (NIPALS) are most commonly used algorithms in PCA analysis. In this study, SVD based principal component analysis was used. In this algorithm the training set $\mathbf{A}$ with *m* samples and *n* variables is decomposed into the principal component scores ($\mathbf{U}$), matrix of singular values ($\mathbf{S}$), and $\mathbf{V}$ matrix whose rows are eigenvectors of $\mathbf{A}$. Equation (3.18) shows the mathematical expression of SVD. The singular values matrix of $\mathbf{S}$ is a diagonal matrix that has elements different from zero on diagonal. Eigenvalues of corresponding training set are calculated using the singular value matrix. The larger the eigenvalue means the more significant information. Generally the principal components (PC's) are calculated according to this significance.

$$\mathbf{A}_{\mathbf{mxn}} = \mathbf{U}_{\mathbf{mxm}}\mathbf{S}_{\mathbf{mxn}}\mathbf{V}^{\mathbf{T}}_{\mathbf{nxn}} \tag{3.18}$$

Often the Equation (3.18) is given in only two matrices in which it is shown in Equation (3.19).

$$\mathbf{A}_{\mathbf{mxn}} = \mathbf{T}_{\mathbf{mxn}}\mathbf{V}^{\mathbf{T}}_{\mathbf{nxn}} \tag{3.19}$$

where $\mathbf{T}$ $(= \mathbf{U}_{\mathbf{mxh}}\mathbf{S}_{\mathbf{hxn}})$ is the score matrix and proportional to the size of the training set contains the information about the objects, $\mathbf{V}^{\mathbf{T}}$ is the loading matrix that has the knowledge of variables. Each row of original data matrix is linear combinations of loading vectors. The first PC is generally the best straight line in multidimensional space and first two PC's are used to visualize the samples (Brereton 2003, Massart, et al. 1998). In Figure 3.5 three distinct groups can be seen by plotting first principal

component score versus second principal component score graph, which is called score plot.

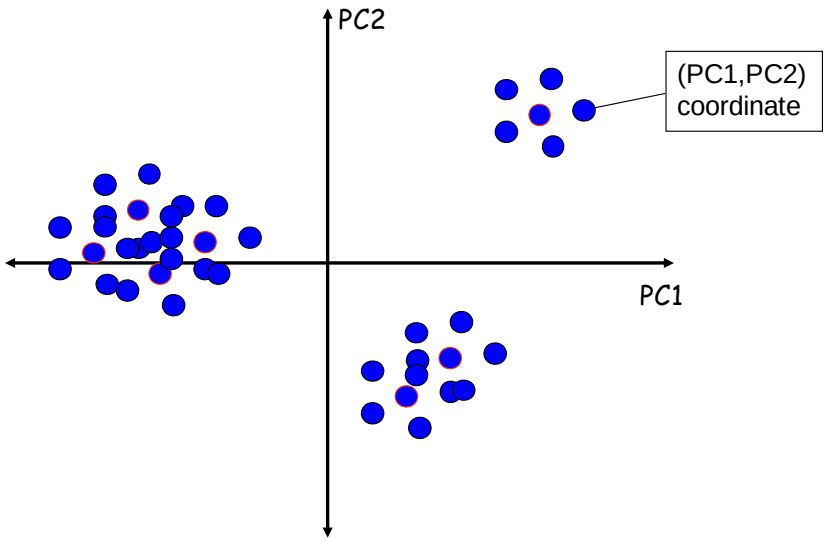


Figure 3.5. Score plot of a representative example which has three groups

As it is mentioned before, multidimensional data contains information of the variables of samples or objects. PCA generally uses not only all the wavelengths in the spectra but also the variables that are extracted from the spectral measurements. When data reduction term is used in PCA analysis, it means variable selection is done. On the other hand, all the wavelengths in the spectra are used in the explanation of the relationships between variables. For the best selection, one can also need a reduction in the number of wavelengths. As a result, the data interpretation of objects is done with the most useful wavelengths and their corresponding variables and the relationship between the samples can be observed clearly. GA are used for wavelength selection in this case as in the calibration part.

The algorithm is very similar to GILS method according to genetic algorithm steps, but there are some differences in the steps. Since the aim is optimizing a classification technique, which is in our case PCA, GA is imposed to PCA and distance which is defined as the distances between the groups of the sample set is taken as fitness function. The new algorithm was named as genetic algorithm based principal component analysis with distance fitness criterion (GAPCA-d). For instance, if the distance between two groups is increased by selection of certain wavelengths, and thus a clearer boundary between the groups is obtained, then classification is optimized.

The gene is defined as the randomly selected wavelengths with their corresponding instrumental responses in the whole spectra for a spectroscopic data set as in the GILS case.

The evaluation of the genes is done with a fitness function that measures the success of the population based on their ability to solve the given problem. Once a gene is selected, it is used to form the reduced data matrix at the points determined by the elements in that gene. This data matrix is used in PCA analysis where score and loading matrices of all PC's are determined. Also all the eigenvalues of the system are found and used to evaluate the systems. The summation of cumulative value of these first two eigenvalues generally explains about the 95% of the system. The total cumulative value of the first two eigenvalues is expected as near as 100%. According to the GAPCA algorithms, the first two eigenvalues are forced to be very significant for the explanation of the system. These first two PC's are used to calculate the distance between the classes or groups and the value of distance is chosen as large as possible for the system. After the calculation of the distance values for all the genes in the population, the genes are sorted from largest to smallest and the best one is reserved for the comparison with the best of the next generation. In Figure 3.6, it is schematically shown how fitness function, which is the summation of all distance values, is applied to the data.

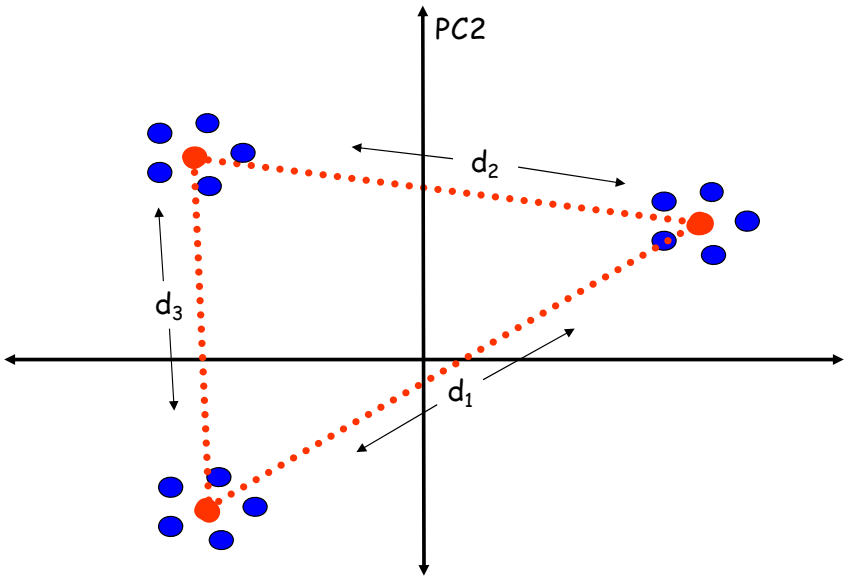


Figure 3.6. Schematic representation of evaluation of a gene according to fitness criterion in GAPCA method

The parent genes are selected in third step from the current population according to their fitness function using roulette wheel selection method. The genes that have higher fitness values shares larger portion of the wheel and those that have small fitness values have small areas on the wheel. The goal is to give a higher chance to those genes with fitness so that only the best performing members will survive in the long run and will pass their information to the next generations. Better suited for the problem will generate even better off-springs. The genes with the low fitness values will be given lower chance to breed hence most of them will be unable to survive.

After selection of the parent genes for the next generation, the genes were put the crossover by top down without resort the selected ones. After crossover, the parent genes are replaced by their off-springs and the off-springs are evaluated with SVD-PCA once again. Then, the whole roulette wheel selection and crossover operation are repeated.

This cycle continues until a predefined number of iteration is reached and the gene that has the highest classification power is selected to analyze the data at the final step. Because GAPCA is based on a lot of random processes, it is expected that whenever the algorithm is run, it will generate a different result. For this reason, the algorithm is designed to run multiple times for a given classification problem and it is possible to make a comparison among these runs in terms of the similarities and dissimilarities of the best genes of each run. By this way, it is possible to determine whether the GA based PCA is able to focus on the regions of the spectrum that contains the necessary information for accurate classification of the samples.

CHAPTER 4

EXPERIMENTATION & INSTRUMENTATION

4.1. Experimentation

Turkish pine and Anatolian black pine samples were collected from Isparta, Turkey. Turkish pine trees were sampled from terrains of 700 m elevation and Anatolian black pine trees from 1230 m. Average precipitation was 51.5 cm. Average maximum temperature in July was 30.3°C and in January 1.8°C. Age of Turkish pine trees were around 30 – 40 years and Anatolian black pine around 17 – 22 years. Tree selection was based on good form trees and eccentric piths were not used. Wood samples were taken from breast height section of the trees. There were total of 58 Turkish pine and 51 Anatolian black pine samples collected. Details about the samples are given in Table 4.1.

Table 4.1. Number of samples with respect to sylviculture terrains

	Turkish pine	**Anatolian black pine**
Control terrain	28	29
Thinning applied terrain	23	29
TOTAL	51	58

Thinning is the selective removal of trees to improve the growth rate or health of the remaining trees. From this definition, the term thinning applied terrain can be easily understood. Control terrain is where no thinning is applied to the trees.

Extractives and lignin contents of the samples were determined according to TAPPI standard methods T204 om-88 and T222 om-88. The obtained values were used as reference in the calibrations. First of all, wood meal samples were prepared with a Wiley mill and ground to pass various mesh screens. In order to determine the content of extractives, ethanol-benzene solution (1:2v/v) was used according to T204 om-88

method. Extraction was performed in a Soxhlet apparatus for 6 hours. After filtration of extract, the remaining solid was weighed. It is subtracted from initial mass and expressed as weight percentage. In order to determine acid insoluble lignin content, the carbohydrates in the wood meal sample hydrolyzed and dissolved in 72% (v/v) sulfuric acid according to T222 om-88 method. Then, acid insoluble lignin is filtered off, dried, and the content in the sample is measured as weight percent.

4.2. Instrumentation

Near-infrared spectroscopic analyses were performed with FTS-3000 NIR spectrometer (Bio-Rad, Excalibur, Cambridge, MA) and mid-infrared spectroscopic analyses were performed with Spectrum 100 FTIR spectrometer (Perkin Elmer, Waltham, MA). Configurations of the spectrometers are shown in Table 4.2. Three spectra were taken for each sample and the means of corresponding three spectra were used in multivariate analyses.

Table 4.2. Instrumental parameters used in the spectrometric analyses

	NIR spectrometer	**MIR spectrometer**
source	tungsten-halogen lamp	tungsten lamp
beam splitter	calcium fluoride	extended range KBr
detector	lead selenide	FR-DTS
resolution	16 cm^{-1}	4 cm^{-1}
# of scans	128	4
# of data points	780	3601
range	10,000 – 4,000 cm^{-1}	4,000 – 400 cm^{-1}

Initially, the wood meal samples were allowed to pass through 300 μm mesh screen to obtain uniform particle size. This was needed because non-uniform particle size might affect the absorbance measurements. Then, the samples were dried in an incubator for 24 hours before spectroscopic measurements to obtain uniform humidity. For both near-infrared and mid-infrared measurements, diffuse reflectance accessories (Pike Technologies, DiffusIR Accessory) were used. Wood meal samples were placed into micro sample cup (6.0 mm diameter, 1.6 mm deep) cautiously making the surface

as flat as possible to minimize absorbance changes due to the surface. For background correction, gold disk was used for near-infrared and mirror disk was used for mid-infrared measurements.

4.3. Data Analysis

The collected spectra were transferred in ASCII file format and were combined with Microsoft Excel program. Then, data files for multivariate analyses were prepared as text files. Genetic algorithm based calibration and classification methods were written in MATLAB programming language Version 7.0 (MathWorks Inc., Natick, MA).

CHAPTER 5

RESULTS AND DISCUSSION

5.1. Calibration Results

5.1.1. Near-Infrared Spectroscopy

Near infrared diffuse reflectance spectra of 10 samples from both Turkish pine and Anatolian black pine trees are shown in Figure 5.1 and Figure 5.2. It is evident that the samples yield high absorbances around 6800, 5150, and 4700 cm^{-1} wavelength regions. However, since they are all pine wood samples, their spectral characteristics are very much alike except the severe baseline differences among the samples. This type of baseline shifts in the absorbance scale is common in diffuse reflectance spectroscopy and part of it is due to composition differences and part of it is due to inhomogeneities. Since GILS method is a genetic algorithm based multivariate calibration technique, it was expected that it could select certain combination of wavelengths which had maximum correlation with extractives and lignin contents of the samples.

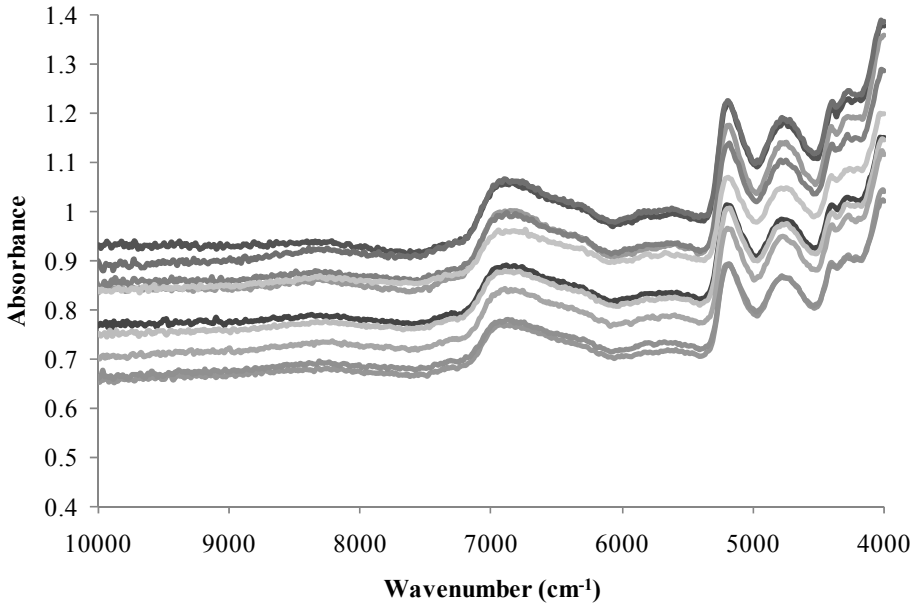


Figure 5.1. NIR diffuse reflectance spectra of 10 Turkish pine samples

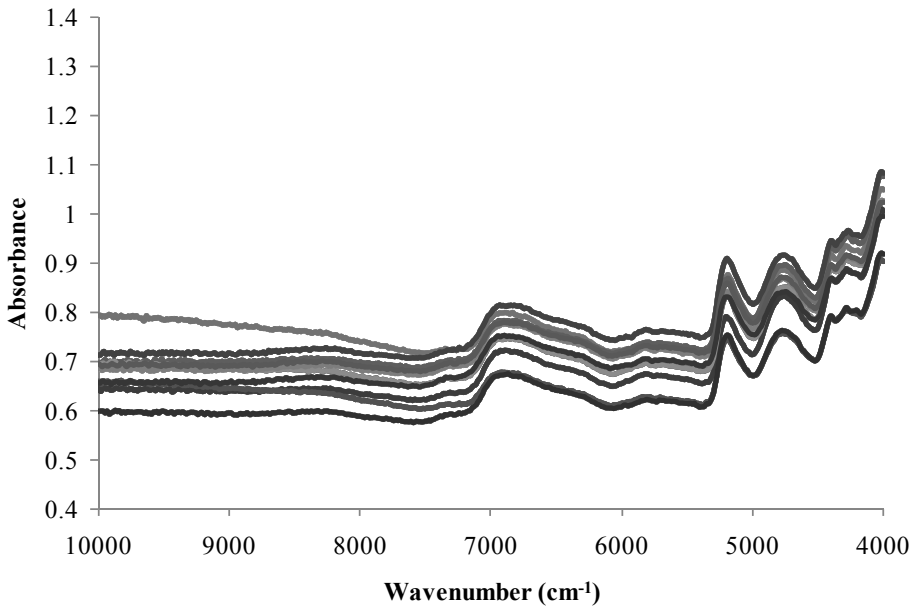


Figure 5.2. NIR diffuse reflectance spectra of 10 Anatolian black pine samples

5.1.1.1. Anatolian Black Pine

In order to construct NIR spectroscopic multivariate calibration models for extractives and lignin contents, three different calibration sets were prepared. The samples were received in two different dates, so calibration was constructed separately and for all samples. Reference extractives and lignin contents of Anatolian black pine samples are given in Table 5.1 and Table 5.2.

Table 5.1. Reference extractives and lignin contents of 1st party Anatolian black pine trees

sample no	lignin content (%w/w)	extractives content (%w/w)	sample no	lignin content (%w/w)	extractives content (%w/w)
1	19.46	4.60	**12**	26.46	13.00
2	16.87	7.30	**13**	17.53	10.60
3	20.99	7.40	**14**	23.03	8.30
4	22.16	6.80	**15**	17.96	6.90
5	22.90	10.00	**16**	14.30	8.80
6	19.44	5.30	**17**	20.71	4.60
7	20.80	6.30	**18**	18.00	5.10
8	24.23	6.60	**19**	23.88	7.80
9	21.60	8.80	**20**	19.19	7.60
10	23.59	7.60	**21**	20.28	4.30
11	24.61	9.40			

Table 5.2. Reference extractives and lignin contents of 2nd party Anatolian black pine trees

sample no	lignin content (%w/w)	extractives content (%w/w)	sample no	lignin content (%w/w)	extractives content (%w/w)
25	22.37	11.73	**43**	19.65	9.26
26	22.59	10.75	**44**	21.84	11.10
27	26.51	9.26	**46**	26.15	10.31
28	24.89	10.42	**47**	24.68	11.79
29	19.42	11.55	**48**	24.48	11.86
30	22.47	13.40	**49**	23.06	8.90
31	22.98	9.90	**50**	20.86	9.83
33	26.61	10.72	**51**	26.32	11.77
34	23.82	10.01	**52**	21.06	11.59
35	30.85	8.61	**53**	21.21	13.61
38	22.41	12.19	**54**	28.44	10.35
39	18.84	12.40	**55**	34.47	12.79
40	21.17	11.33	**56**	23.88	11.62
41	21.07	11.83	**57**	19.94	8.85
42	28.27	11.79	**58**	16.70	12.13

The first calibration set were generated from 1st party in which 14 of them were randomly selected with the samples having minimum and maximum extractives and lignin contents and these samples were assigned as calibration set. The remaining 7 samples were reserved for independent test samples. Reference extractives and lignin contents versus predicted values based on NIR spectra using GILS method are shown in Figure 5.3 for the first data set. Calibration models for lignin content determination gave standard error of calibration (SEC) and standard error of prediction (SEP) values as 0.51% (w/w) and 1.70% (w/w) for calibration and independent test sets, respectively. In the case of extractives content determination, the SEC and SEP values were 0.49% (w/w) and 1.12% (w/w) for calibration and prediction sets, respectively. The R^2 value of regression lines for lignin was 0.975 and that for extractives content was 0.961.

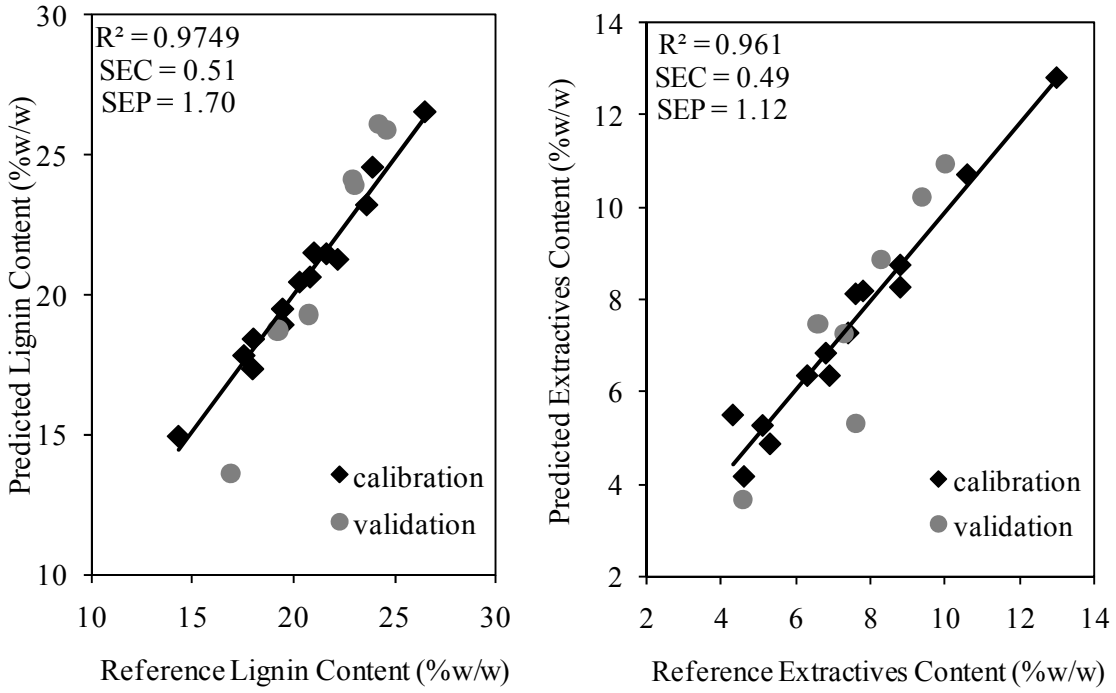


Figure 5.3. Reference vs. NIR predicted extractives and lignin contents for the first data set of Anatolian black pine trees

When these SEC and SEP values are examined, it is seen that the values for lignin content were comparable even though the SEP value is about the twice of the SEC. It must be realized that the GILS method is an iterative procedure due to the genetic algorithm used to select a subset of wavelengths from the whole spectral range. As mentioned above, NIR spectra of these samples suffer from somewhat large baseline fluctuation and this causes the GILS to model this effect while preparing calibration models even though the cross validation approach is used during model building step. Since independent test samples in the prediction set do not have same baseline trends as in the calibration set and therefore predictions result in larger SEP values. Yet, when the overall calibration performance of the models examined, it is possible to state that the NIR spectra do contain quantitative information that is correlated with extractives and lignin contents of the Anatolian black pine samples studied here.

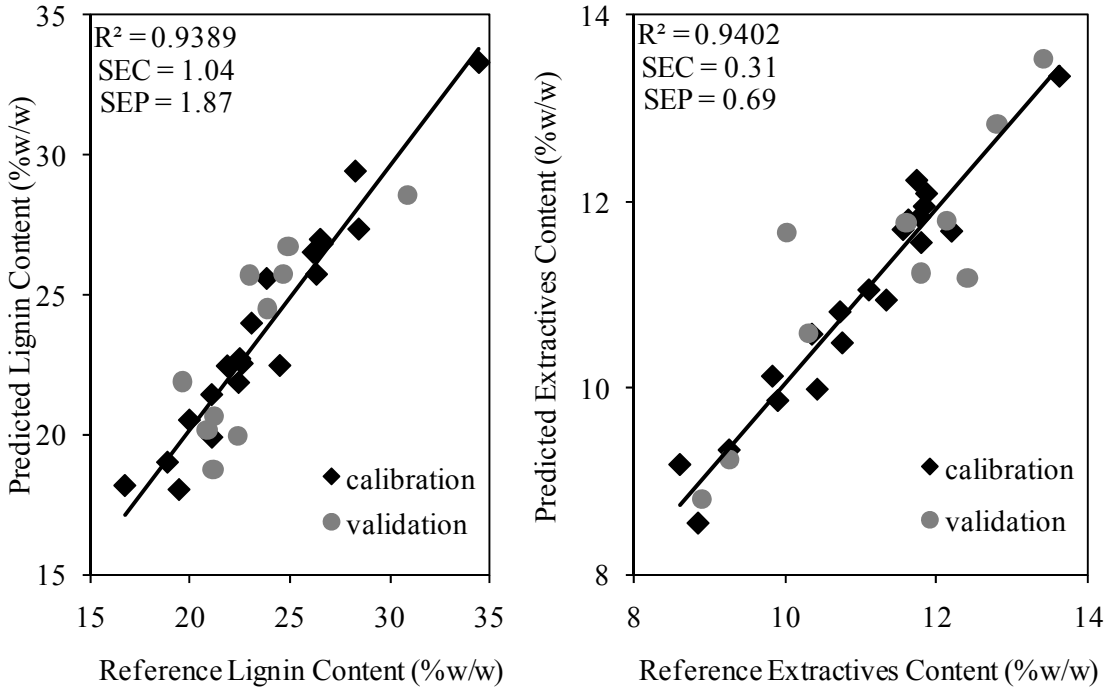


Figure 5.4. Reference vs. NIR predicted extractives and lignin contents for the second data set of Anatolian black pine trees

Figure 5.4 shows the reference extractives and lignin contents versus GILS predicted values for the second data set with 30 samples of which 20 of the used for model building in the calibration set and the remaining 10 samples were reserved for the prediction set. While, the concentrations of lignin content were ranging between 14% (w/w) and 27% (w/w) for the first data set, the upper level of lignin content in second data set was around 35% (w/w). On the other hand, the extractives content of the samples in the second data set were distributed in a narrower range between 5% (w/w) and 11% (w/w) when compared with the first data set. The SEC values for extractives and lignin contents were 0.31% (w/w) and 1.04% (w/w), respectively while the SEP values were ranged between 0.69% (w/w) and 1.87% (w/w) for extractives and lignin contents. The R^2 value of regression lines for lignin was 0.939 and that for extractives content was 0.940.

When SEC and SEP values are examined in the second data set, it is seen that the agreement between these values are better than those obtained for the first data set even though lignin content interval is larger. One possible explanation of this improvement could be attributed to increased number of calibration and prediction

samples. On the other hand, the R^2 of calibration lines were now lower than those obtained for the first data set. This is also an expected outcome of calibration models with larger data set as variability increases with the increased number of sample in calibration set.

The third data set analyzed in this part of study was formed by combining the first and the second data sets into a single set. The calibration and prediction sets are formed by adding the corresponding spectra in the first data set to the data in the second data set. The calibration plots for extractives and lignin contents are given in Figure 5.5.

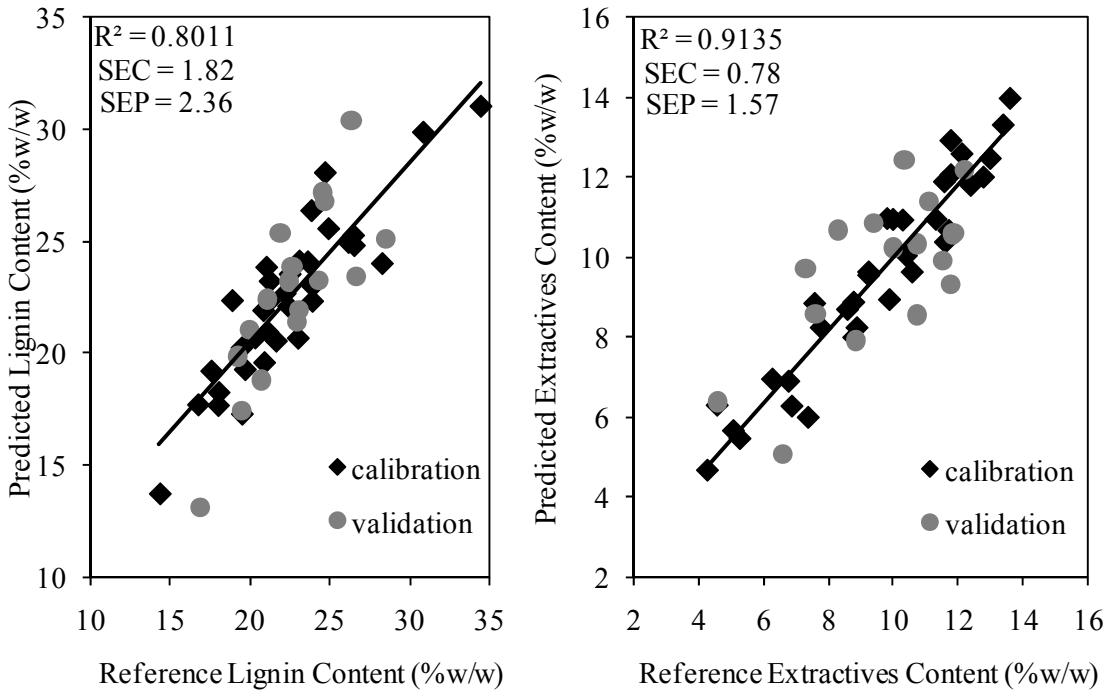


Figure 5.5. Reference vs. NIR predicted extractives and lignin contents for the third data set of Anatolian black pine trees

Since the samples in the first data set were received in different date than the samples in the second data set, both SEC and SEP values were somewhat higher in the third data set compared to the first and second data sets. For the determination of lignin content, SEC and SEP values were 1.82% (w/w) and 2.36% (w/w), respectively. In the case of extractive content determination similar results were obtained in which the SEC was 0.78% (w/w) and the SEP was 1.57% (w/w). These increases in calibration and

prediction results were also reflected in R^2 values of regression as the values went down to 0.801 for lignin and 0.913 for extractives.

Because GILS is a wavelength selection based method, it is interesting to observe the distribution of selected wavelengths in multiple runs over the entire full spectral region. Figure 5.6 illustrates the frequency distribution of selected wavelengths in 100 runs with 20 genes and 50 iterations for the third data set.

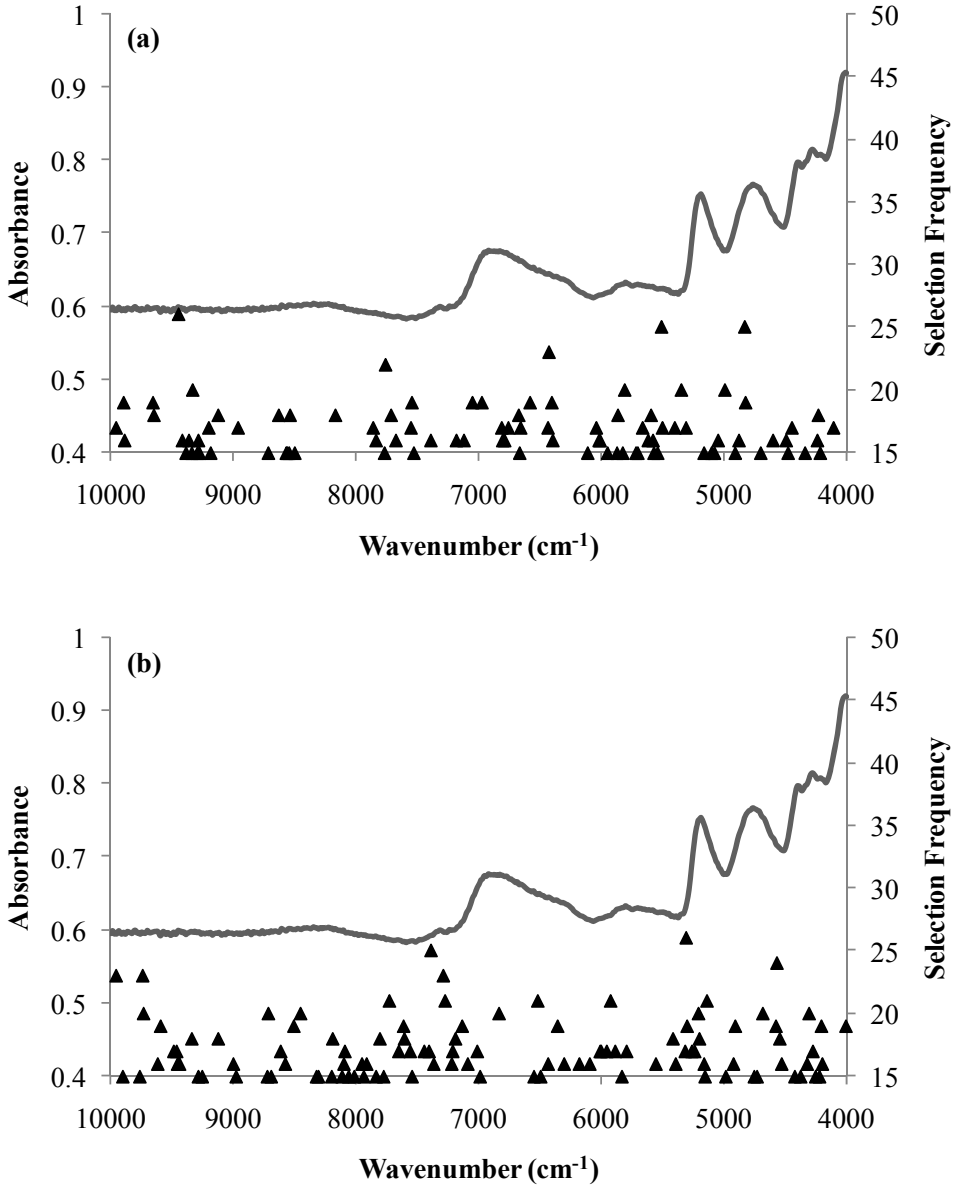


Figure 5.6. Frequency distribution of GILS selected NIR wavelengths for both lignin (a) and extractives (b) contents of Anatolian black pine samples in the third set

As can be seen from Figure 5.6 there are a number of regions where selection frequencies are very high compared to the rest of the spectrum. The wavelength region around 5500 and 9500 cm^{-1} for lignin content indicates a strong tendency for GILS method to select while for extractives content, around 5200 and 7300 cm^{-1} is the most frequently selected region.

5.1.1.2. Turkish Pine

As in the Anatolian black pine case, three different calibration sets were prepared. Reference extractives and lignin contents of Turkish pine samples are given in Table 5.3 and Table 5.4.

Table 5.3. Reference extractives and lignin contents of 1st party Turkish pine trees

sample no	lignin content (%w/w)	extractives content (%w/w)	sample no	lignin content (%w/w)	extractives content (%w/w)
1	28.80	6.00	**12**	28.67	7.89
2	29.27	5.84	**13a**	28.90	6.09
3	29.04	7.19	**13b**	29.07	6.09
4	28.93	6.78	**14**	29.16	6.86
5	28.63	6.46	**15**	28.82	6.72
6	29.10	6.27	**16**	28.79	6.69
7	28.96	7.51	**17**	29.20	7.02
8	28.69	6.03	**18**	28.44	7.53
9	29.00	5.73	**19**	29.08	6.43
10	28.96	5.85	**26**	28.93	6.31
11	29.02	6.59			

Table 5.4. Reference extractives and lignin contents of 2nd party Turkish pine trees

sample no	lignin content (%w/w)	extractives content (%w/w)	sample no	lignin content (%w/w)	extractives content (%w/w)
25	25.49	8.55	**45**	22.30	10.87
26	32.25	6.42	**46**	28.39	8.10
27	31.67	3.56	**47**	25.44	8.70
28	31.77	11.58	**48**	22.27	4.55
29	34.67	6.73	**49**	31.66	6.46
30	35.46	11.64	**50**	23.50	8.52
31	28.40	9.70	**51**	30.47	6.55
32	34.99	6.30	**52**	25.37	6.18
33	29.57	8.00	**53**	34.57	7.92
34	33.53	7.21	**54**	30.80	2.71
35	25.34	16.12	**55**	23.76	11.92
36	33.94	11.27	**56**	32.02	9.20
37	30.30	12.06	**57**	21.46	2.05
38	26.73	10.42	**58**	26.04	8.57
39	27.65	8.65	**59**	36.70	7.45
40	31.02	7.46	**61**	33.47	8.26
41	24.10	9.04	**64**	35.09	11.93
42	19.93	8.25	**66**	31.36	9.09
44	34.11	10.41			

The first calibration set were generated from 1st party in which 14 of them as calibration set and the remaining 7 samples as test samples. Reference extractives and lignin contents versus predicted values based on NIR spectra using GILS method are shown in Figure 5.7 for the first data set. Calibration models for lignin content determination gave standard error of calibration (SEC) and standard error of prediction (SEP) values as 0.03% (w/w) and 0.10% (w/w) for calibration and independent test sets, respectively. In the case of extractives content determination, the SEC and SEP values were 0.11% (w/w) and 0.27% (w/w) for calibration and prediction sets, respectively. The R^2 value of regression lines for lignin was 0.984 and that for extractives content was 0.964.

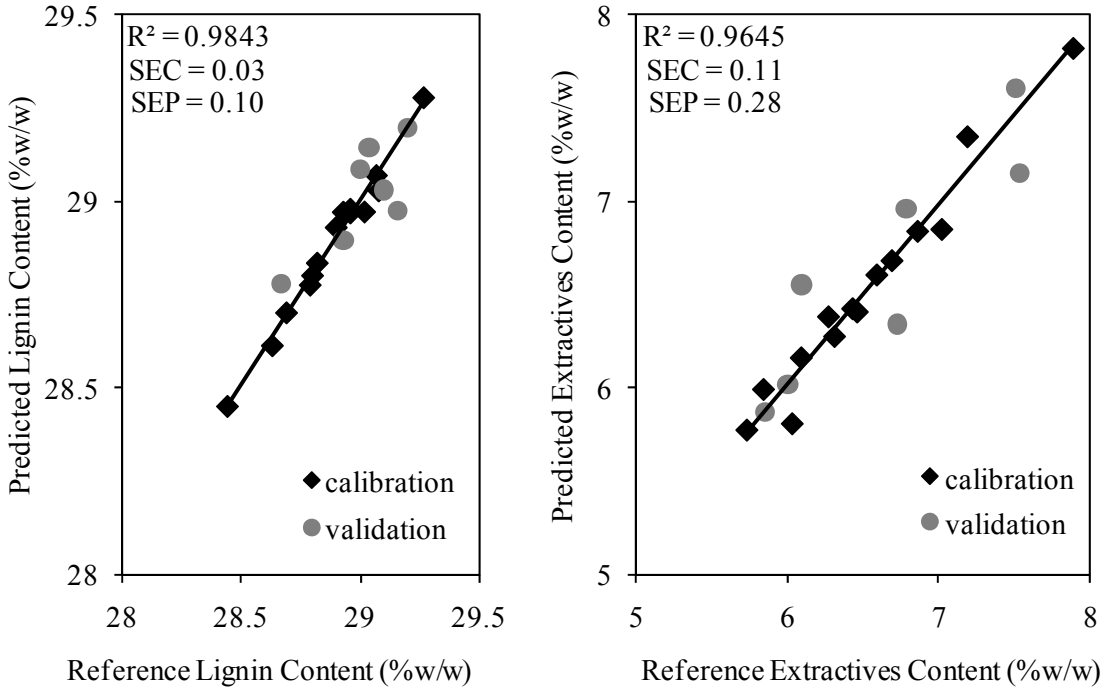


Figure 5.7. Reference vs. NIR predicted extractives and lignin contents for the first data set of Turkish pine trees

When these SEC and SEP values are examined, it is seen that the values are smaller than the Anatolian black pine case since reference values lay on narrower intervals. Similar regression coefficients show that NIR spectra of Turkish pine trees also contain information of extractives and lignin. NIR spectra of Turkish pine samples again suffer from baseline fluctuations and this causes GILS to model this effect while

preparing calibration models even though the cross validation approach is used during model building step. So, high SEP values of validation sets are caused by this.

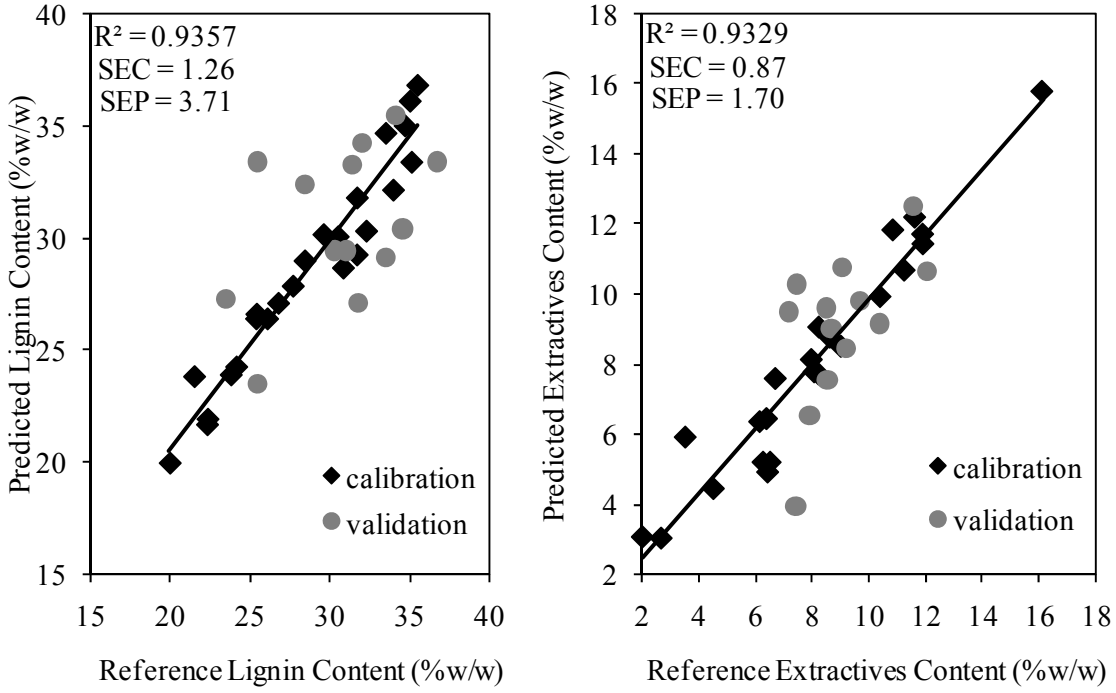


Figure 5.8. Reference vs. NIR predicted extractives and lignin contents for the second data set of Turkish pine trees

Figure 5.8 shows the reference extractives and lignin contents versus GILS predicted values for the second data set with 37 samples of which 24 of the used for model building in the calibration set and the remaining 13 samples were reserved for the prediction set. While, the concentrations of lignin content for Turkish pine trees were ranging between a very narrow interval for the first data set, lignin content range in second data set was much larger. Similarly, the extractives content of the samples in the second data set were distributed in a wider range between 3% (w/w) and 16% (w/w) when compared with the first data set. The SEC values for extractives and lignin contents were 0.87% (w/w) and 1.26% (w/w), respectively while the SEP values were ranged between 1.70% (w/w) and 3.71% (w/w) for extractives and lignin contents. The R^2 value of regression lines for lignin was 0.936 and that for extractives content was 0.933.

When SEC and SEP values are examined in the second data set, it is seen that the agreement between these values are worse than those obtained for the first data set. Actually, SEC values don't differ so much compared to Anatolian black pine calibration but SEP values are more than twice of SEC values. Explanation of this could be the increased number of calibration and prediction samples and larger data interval. On the other hand, the R^2 of calibration lines were now lower than those obtained for the first data set. This is also an expected outcome of calibration models with larger data set as variability increases with the increased number of sample in calibration set.

The third data set analyzed in this part of study is the same as the Anatolian black pine case. The calibration and prediction sets are formed by adding the corresponding spectra in the first data set to the data in the second data set. The calibration plots for extractives and lignin contents are given in Figure 5.9.

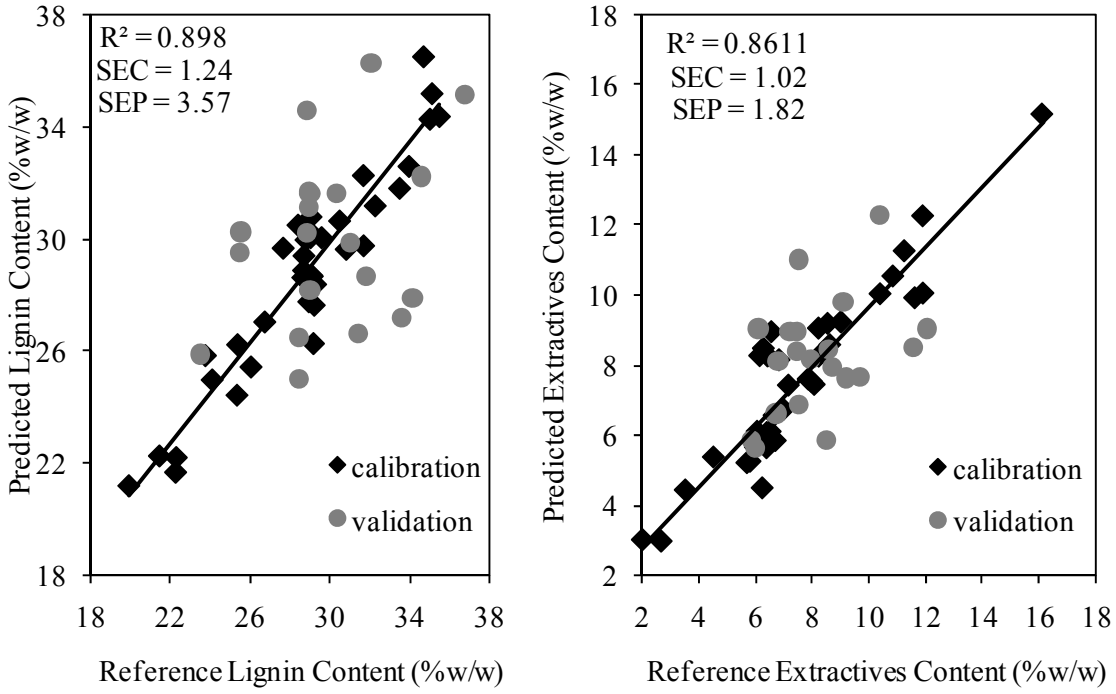


Figure 5.9. Reference vs. NIR predicted extractives and lignin contents for the third data set of Turkish pine trees

Both SEC and SEP values were higher in the third data set compared to the first and second data sets. The reason might be the time of reception of the samples and data interval differences. For the determination of lignin content, SEC and SEP values were

1.24% (w/w) and 3.57% (w/w), respectively. In the case of extractives content determination similar results were obtained in which the SEC was 1.02% (w/w) and the SEP was 1.82% (w/w). These increases in calibration and prediction results were also reflected in R^2 values of regression as the values went down to 0.898 for lignin and 0.861 for extractives.

Figure 5.10 illustrates the frequency distribution of selected wavelengths in 100 runs with 20 genes and 50 iterations for the third data set.

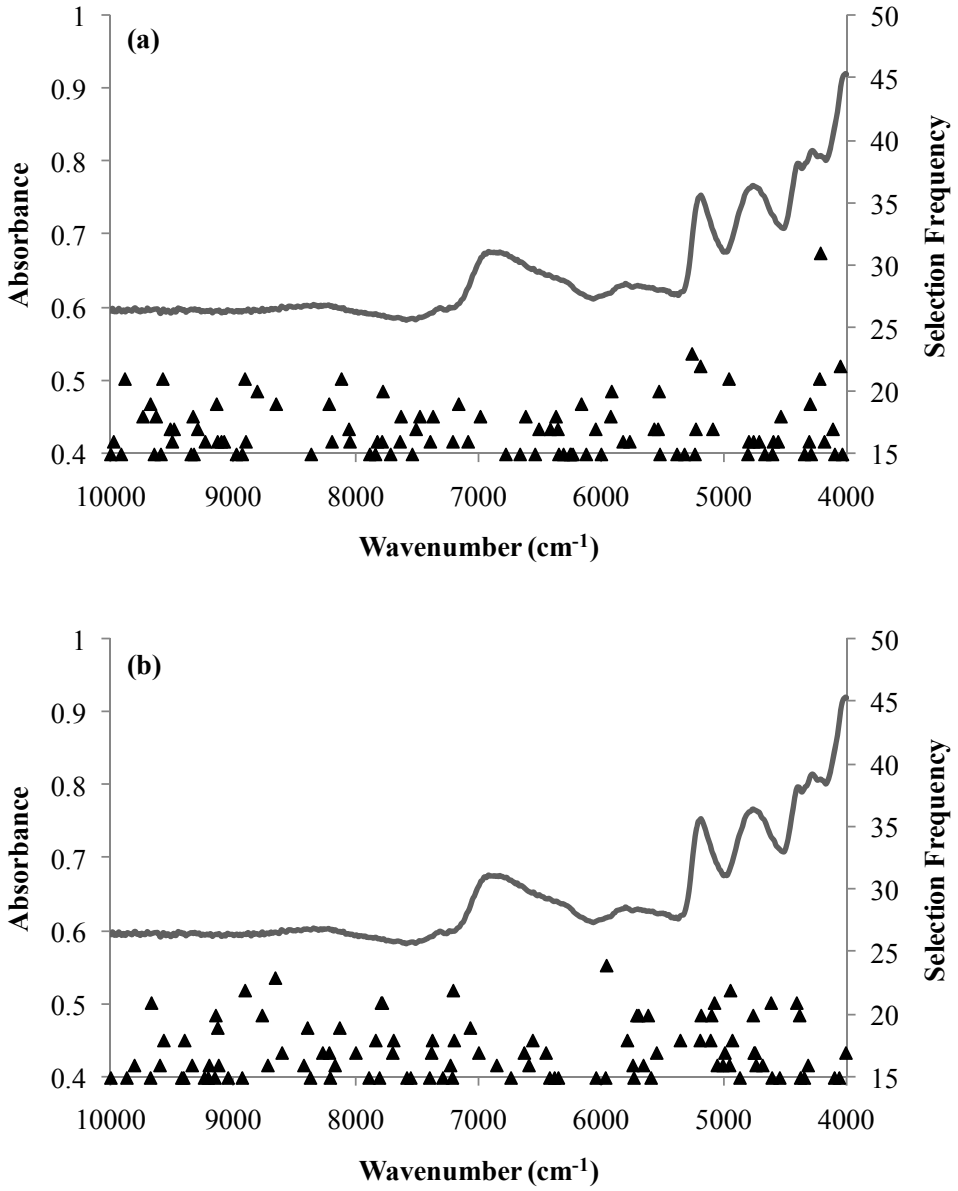


Figure 5.10. Frequency distribution of GILS selected NIR wavelengths for both lignin (a) and extractives (b) contents of Turkish pine samples in the third set

As can be seen from Figure 5.10 there are a number of regions where selection frequencies are very high compared to the rest of the spectrum. The wavelength region around 4200 and 5300 cm^{-1} for lignin content indicates a strong tendency for GILS method to select while for extractives content, around 6000 and 8500 cm^{-1} is the most frequently selected region.

5.1.2. Mid-Infrared Spectroscopy

Mid-infrared diffuse reflectance spectra of 10 wood samples from both Turkish pine and Anatolian black pine trees are shown in Figure 5.11 and Figure 5.12, separately. It is evident that the samples yield high absorbance values around 3400, 2900, and between the range 1750 and 1000 cm^{-1} wavelengths. Also there is a peak around 2150 cm^{-1}. From the spectra, similarities are obviously seen but there are still baseline differences as mentioned in near infrared spectroscopy case. However using the GILS method will decrease the effect of baseline shifts because it can select certain combination of wavelengths which have maximum correlation with extractives and lignin contents of the samples.

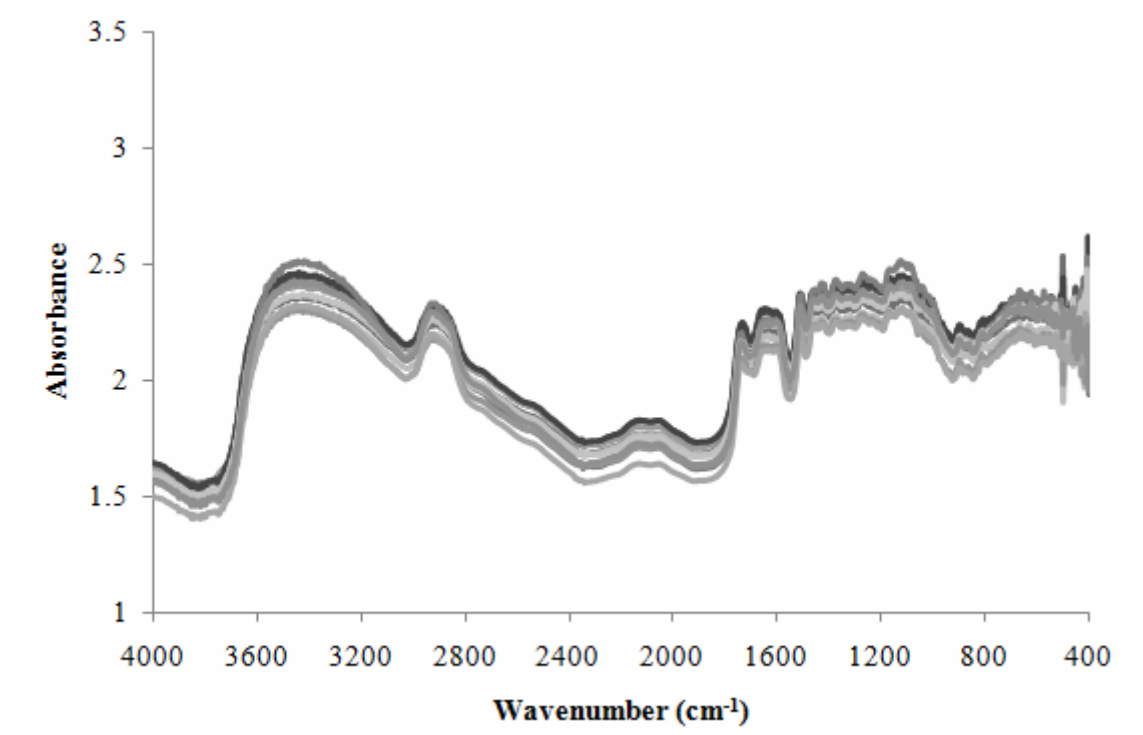


Figure 5.11. MIR diffuse reflectance spectra of 10 Turkish pine samples

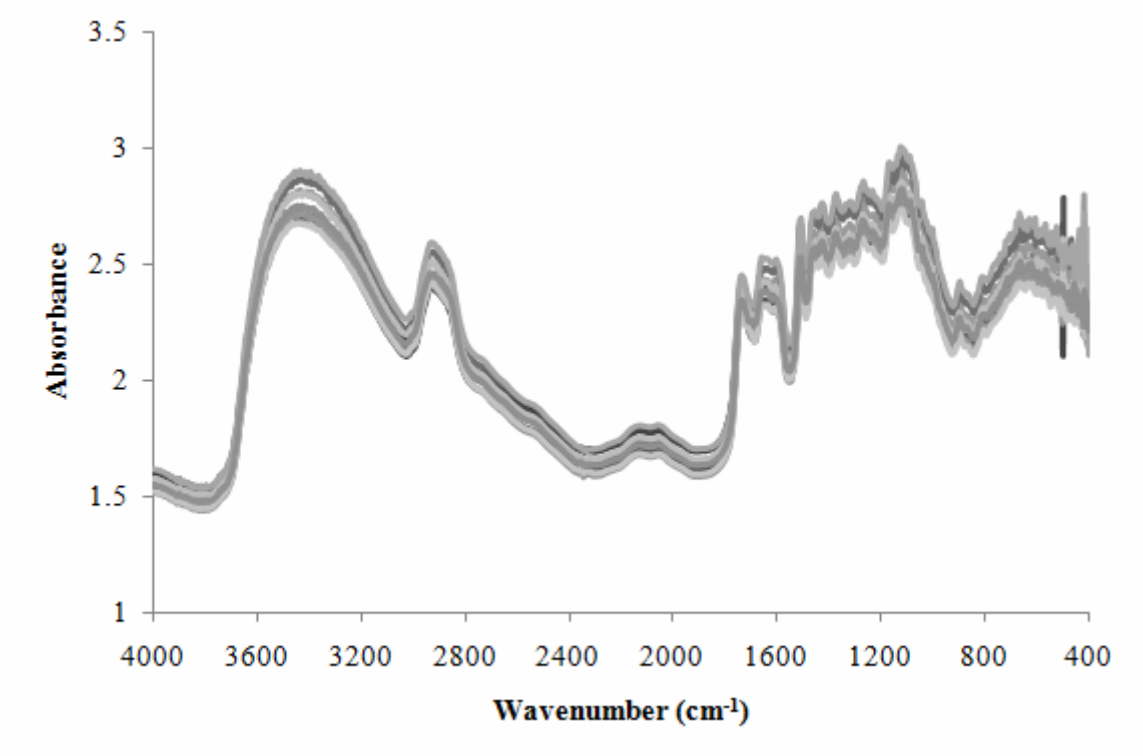


Figure 5.12. MIR diffuse reflectance spectra of 10 Anatolian black pine samples

5.1.2.1. Anatolian Black Pine

In order to construct MIR spectroscopic multivariate calibration models for extractives and lignin contents for Anatolian black pine, the procedure followed in the NIR calibration is again used, i.e., three different calibration sets were used again but NIR spectra were replaced with the MIR spectra.

The first calibration set were generated from 1st party in which 14 of them as calibration set and the remaining 7 samples as test samples. Reference extractives and lignin contents versus predicted values based on MIR spectra using GILS method are shown in Figure 5.13 for the first data set. Calibration models for lignin content determination gave standard error of calibration (SEC) and standard error of prediction (SEP) values as 0.62% (w/w) and 1.66% (w/w) for calibration and independent test sets, respectively. In the case of extractives content determination, the SEC and SEP values were 0.51% (w/w) and 1.13% (w/w) for calibration and prediction sets, respectively. The R^2 value of regression lines for lignin was 0.961 and that for extractives content was 0.958.

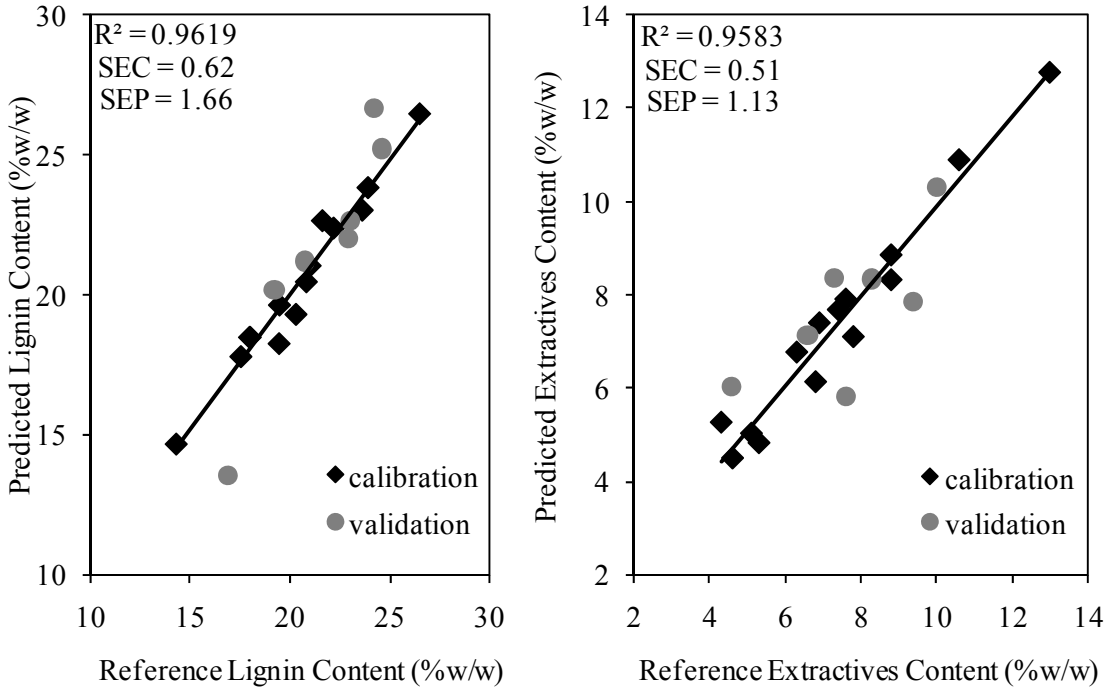


Figure 5.13. Reference vs. MIR predicted extractives and lignin contents for the first data set of Anatolian black pine trees

It is seen from SEC and SEP values that for extractives and lignin contents they are still comparable. But the values become higher than corresponding NIR results. Again it must be realized that the GILS method is an iterative procedure due to the genetic algorithm used to select a subset of wavelengths from the whole spectral range. The effect of baseline fluctuation will be more since MIR region is very sensitive for quantitative analysis because absorbance changes become more than it becomes in NIR case. The reason can be that fundamental vibrations have more probability to be observed than overtones. Yet, when the overall calibration performance of the models examined, it is possible to state that the MIR spectra do contain quantitative information that is correlated with extractives and lignin contents of the Anatolian black pine samples studied here.

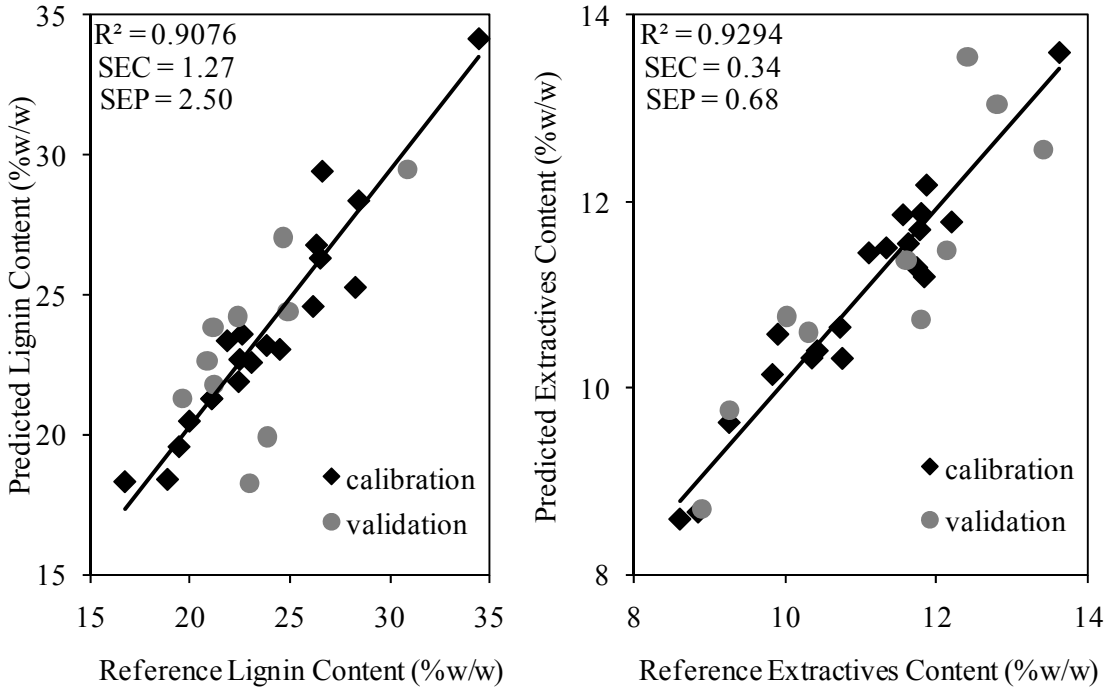


Figure 5.14. Reference vs. MIR predicted extractives and lignin contents for the second data set of Anatolian black pine trees

Figure 5.14 shows the reference extractives and lignin contents versus predicted values by GILS for the second data set. The SEC values for extractives and lignin contents were 0.34% (w/w) and 1.27% (w/w), respectively while the SEP values were ranged between 0.68% (w/w) and 2.50% (w/w) for extractives and lignin contents. The R^2 value of regression lines for lignin was 0.908 and that for extractives content was 0.929.

When compared with the first data set and also the NIR results, SEC and SEP values became higher and thus regression became smaller. One possible explanation of this improvement could be attributed to increased number of calibration and prediction samples. On the other hand, the R^2 of calibration lines were now lower than those obtained for the first data set. This is also an expected outcome of calibration models with larger data set as variability increases with the increased number of sample in calibration set. However the correlation between chemical contents and MIR spectra of Anatolian black pine samples are still seen.

The third data set analyzed in this part of study was formed by combining the first and the second data sets into a single set. The calibration and prediction sets are

formed by adding the corresponding spectra in the first data set to the data in the second data set. The calibration plots for extractives and lignin contents are given in Figure 5.15.

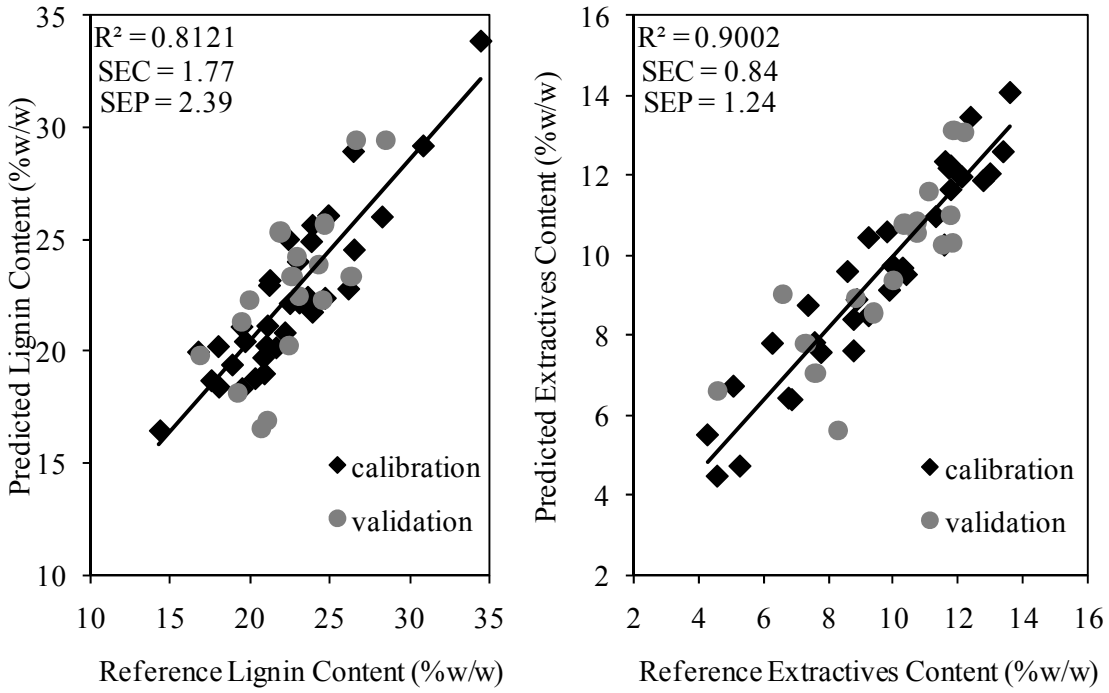


Figure 5.15. Reference vs. MIR predicted extractives and lignin contents for the third data set of Anatolian black pine trees

Since the samples in the first data set were received in different date than the samples in the second data set, both SEC and SEP values were somewhat higher in the third data set compared to the first and second data sets. For the determination of lignin content, SEC and SEP values were 1.77% (w/w) and 2.38% (w/w), respectively. In the case of extractives content determination similar results were obtained in which the SEC was 0.84% (w/w) and the SEP was 1.24% (w/w). These increases in calibration and prediction results were also reflected in R^2 values of regression as the values went down to 0.812 for lignin and 0.900 for extractives.

Figure 5.16 illustrates the frequency distribution of selected wavelengths in 100 runs with 20 genes and 50 iterations for the third data set.

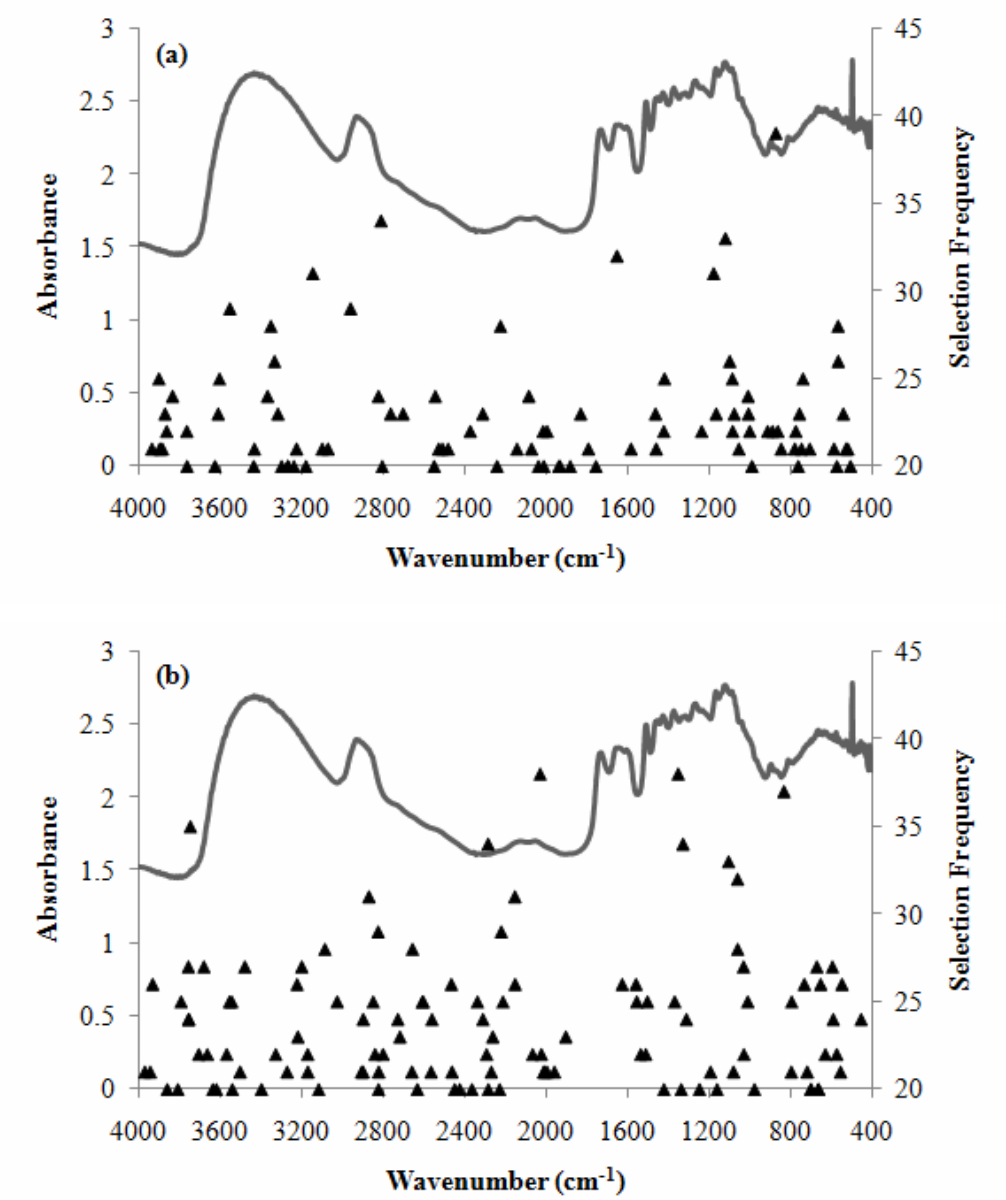


Figure 5.16. Frequency distribution of GILS selected MIR wavelengths for both lignin (a) and extractives (b) contents of Anatolian black pine samples in the third set

As can be seen from Figure 5.16 there are a number of regions where selection frequencies are very high compared to the rest of the spectrum. The wavelength region around 1000 cm^{-1} for lignin content indicates a strong tendency for GILS method to select while for extractives content, around 1200 and 2200 cm^{-1} is the most frequently selected region.

5.1.2.2. Turkish Pine

In order to construct MIR spectroscopic multivariate calibration models for extractives and lignin contents for Turkish pine, the procedure followed in the NIR calibration is again used, i.e., three different calibration sets were used again but NIR spectra were replaced with the MIR spectra.

Reference extractives and lignin contents versus predicted values based on MIR spectra using GILS method are shown in Figure 5.17 for the first data set. Calibration models for lignin content determination gave standard error of calibration (SEC) and standard error of prediction (SEP) values as 0.03% (w/w) and 0.11% (w/w) for calibration and independent test sets, respectively. In the case of extractives content determination, the SEC and SEP values were 0.14% (w/w) and 0.29% (w/w) for calibration and prediction sets, respectively. The R^2 value of regression lines for lignin was 0.981 and that for extractives content was 0.946.

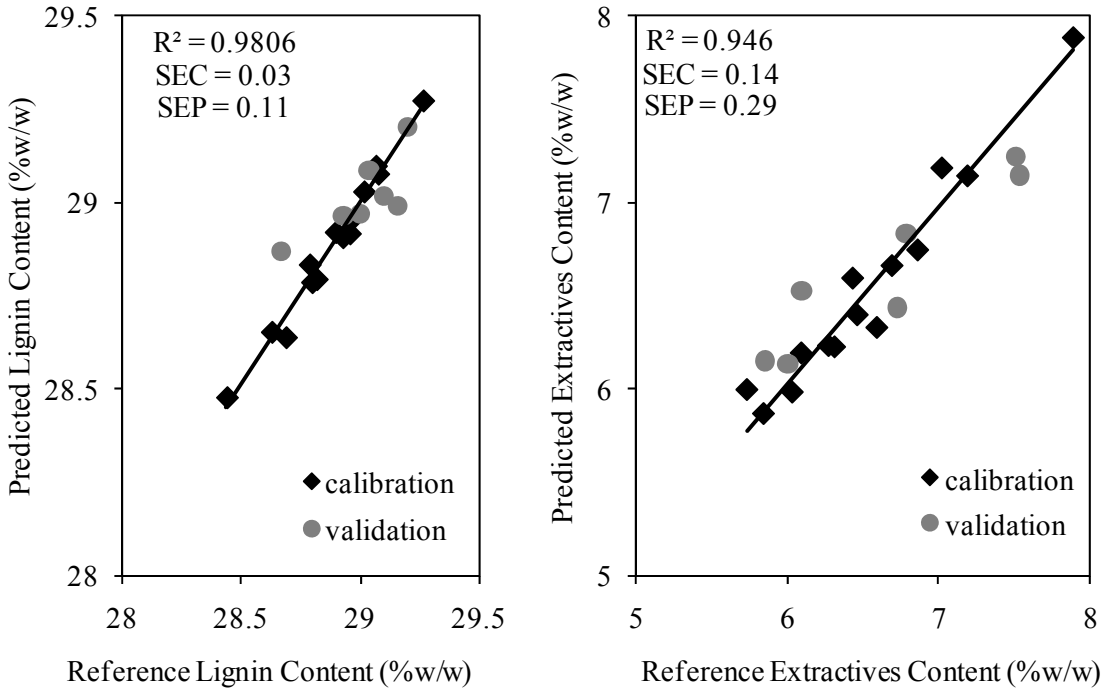


Figure 5.17. Reference vs. MIR predicted extractives and lignin contents for the first data set of Turkish pine trees

From SEC and SEP values, the results are very comparable with NIR results but SEP values are somewhat higher. Similar regression coefficients show that MIR spectra of Turkish pine trees also contain information of extractives and lignin.

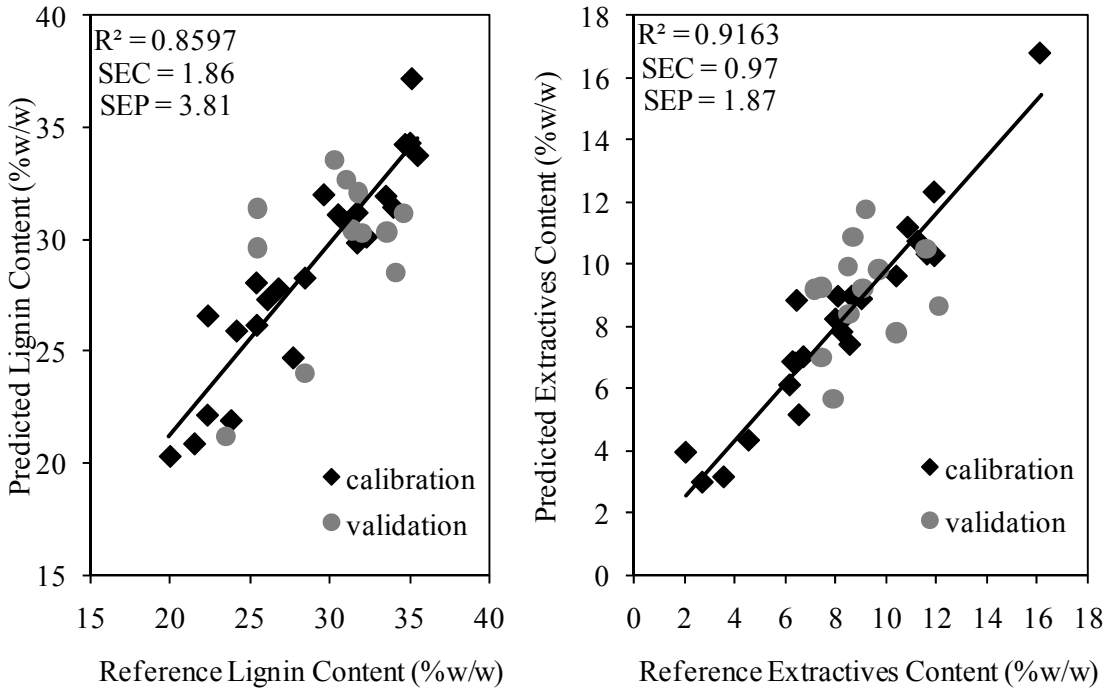


Figure 5.18. Reference vs. MIR predicted extractives and lignin contents for the second data set of Turkish pine trees

Figure 5.18 shows the reference extractives and lignin contents versus GILS predicted values for the second data set. The procedure was same again; MIR spectra were replaced with the NIR spectra before the calibration process. The SEC values for extractives and lignin contents were 0.97% (w/w) and 1.86% (w/w), respectively while the SEP values were ranged between 1.87% (w/w) and 3.81% (w/w) for extractives and lignin contents. The R^2 value of regression lines for lignin was 0.860 and that for extractives content was 0.916. SEC and SEP values in the second data set are higher than the corresponding NIR values. Also, regression coefficients became smaller.

The third data set is formed by adding the corresponding spectra in the first data set to the data in the second data set. The calibration plots for extractives and lignin contents are given in Figure 5.19.

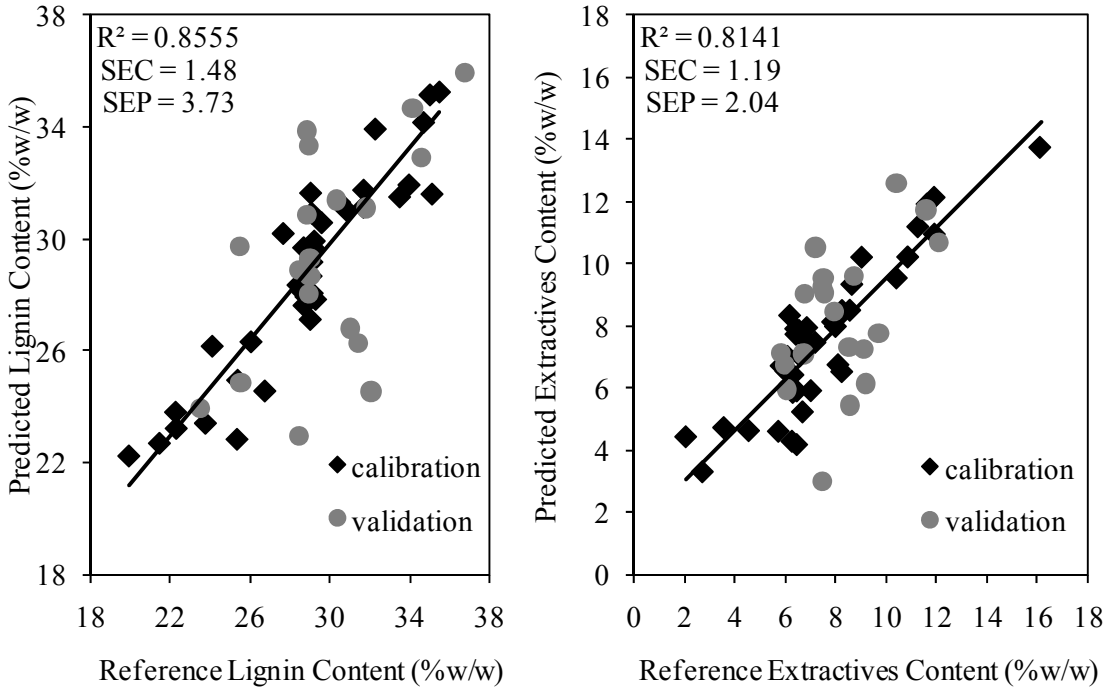


Figure 5.19. Reference vs. MIR predicted extractives and lignin contents for the third data set of Turkish pine trees

Both SEC and SEP values were higher in the third data set compared to the first and second data sets. The reason might be the time of reception of the samples and data interval differences. For the determination of lignin content, SEC and SEP values were 1.48% (w/w) and 3.73% (w/w), respectively. In the case of extractives content determination similar results were obtained in which the SEC was 1.19% (w/w) and the SEP was 2.04% (w/w). These increases in calibration and prediction results were also reflected in R^2 values of regression as the values went down to 0.855 for lignin and 0.814 for extractives.

Figure 5.20 illustrates the frequency distribution of selected wavelengths in 100 runs with 20 genes and 50 iterations for the third data set.

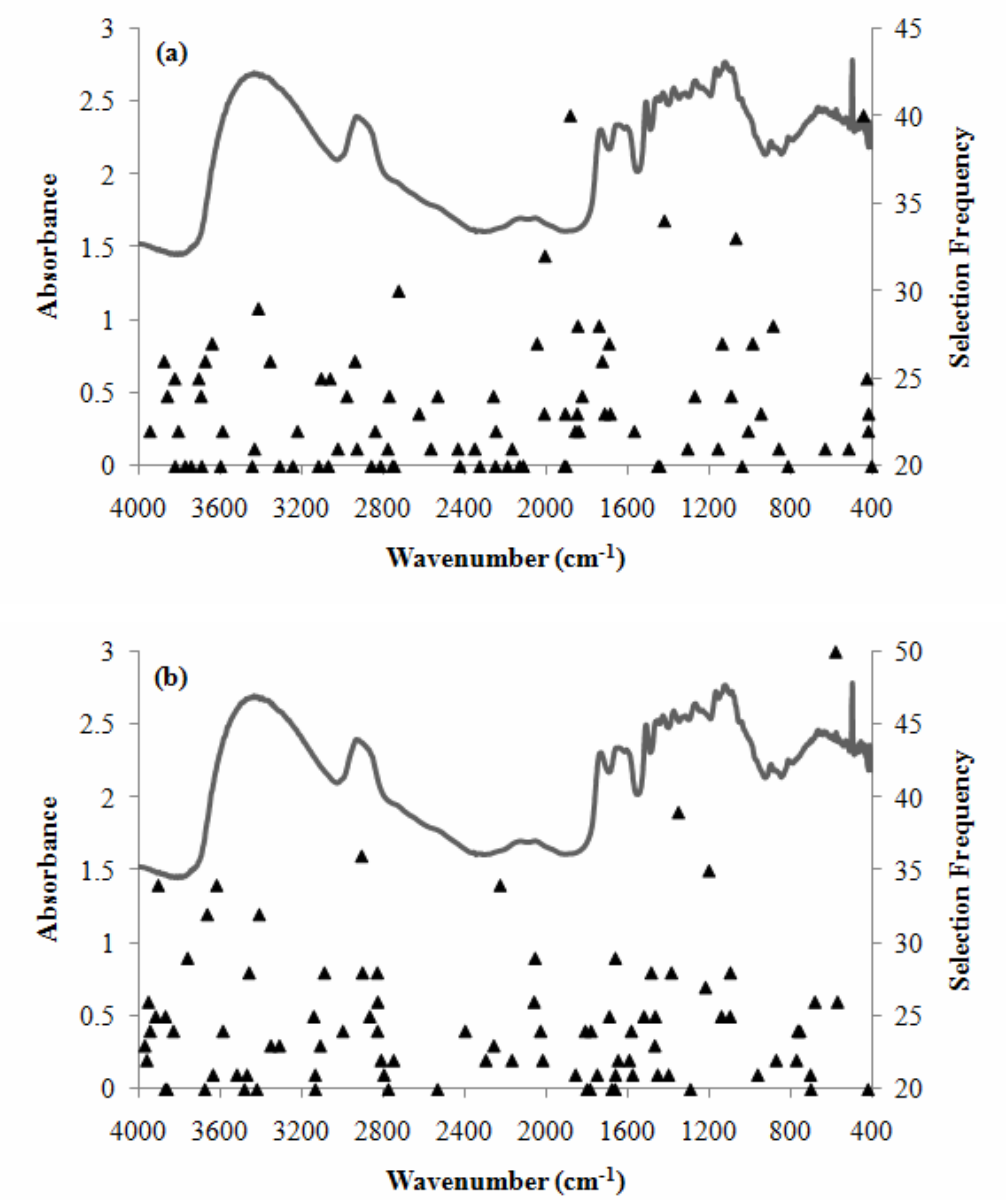


Figure 5.20. Frequency distribution of GILS selected MIR wavelengths for both lignin (a) and extractives (b) contents of Turkish pine samples in the third set

As can be seen from Figure 5.20 there are a number of regions where selection frequencies are very high compared to the rest of the spectrum. The wavelength region around 2000 cm^{-1} for lignin content indicates a strong tendency for GILS method to select while for extractives content, around 1300 and 2800 cm^{-1} is the most frequently selected region.

5.1.3. Calibration Summary

To make a comparison between NIR spectroscopy and MIR spectroscopy; SEC, SEP and R^2 values are given for all three data sets in Table 5.5, Table 5.6 and Table 5.7. In the tables, **ABP** stands for Anatolian black pine and **TP** stands for Turkish pine.

Table 5.5. Calibration summary for the first data set

1st data set	**LIGNIN**			**EXTRACTIVES**		
	SEC (%w/w)	**SEP (%w/w)**	**R^2**	**SEC (%w/w)**	**SEP (%w/w)**	**R^2**
NIR – ABP	0.51	1.70	0.975	0.49	1.12	0.961
MIR – ABP	0.62	1.66	0.961	0.51	1.13	0.958
NIR – TP	0.03	0.10	0.984	0.11	0.27	0.964
MIR – TP	0.03	0.11	0.981	0.14	0.29	0.946

For the first data set, the results are very comparable but NIR results seem to have somewhat better R^2 values for calibration models.

Table 5.6. Calibration summary for the second data set

2nd data set	**LIGNIN**			**EXTRACTIVES**		
	SEC (%w/w)	**SEP (%w/w)**	**R^2**	**SEC (%w/w)**	**SEP (%w/w)**	**R^2**
NIR – ABP	1.04	1.87	0.939	0.31	0.69	0.940
MIR – ABP	1.27	2.50	0.908	0.34	0.68	0.929
NIR – TP	1.26	3.71	0.936	0.87	1.70	0.933
MIR – TP	1.86	3.81	0.855	0.97	1.87	0.814

For the second data set, it is obviously seen that values for NIR models are better than values for MIR models. High R^2 values show that calibration models constructed by NIR spectra of the samples of second set are more successful in this case. However, it is necessary to take a look at the third data set which is the combination of first and second data sets. Since samples from first and second sets were received in different time periods, making a general comment according to the second data set would be insufficient.

Table 5.7. Calibration summary for the third data set

3rd data set	**LIGNIN**			**EXTRACTIVES**		
	SEC (%w/w)	**SEP (%w/w)**	**R^2**	**SEC (%w/w)**	**SEP (%w/w)**	**R^2**
NIR – ABP	1.82	2.36	0.801	0.78	1.57	0.913
MIR – ABP	1.77	2.38	0.812	0.84	1.24	0.900
NIR – TP	1.24	3.57	0.898	1.02	1.82	0.861
MIR – TP	1.48	3.73	0.855	1.19	2.04	0.814

For the third data set, the R^2 values are lower than the ones in the first two data sets. When NIR and MIR results are compared, NIR results seem to be better except in the case of lignin prediction for Anatolian black pine.

As a result, both NIR and MIR spectroscopy can be used to construct calibration models for lignin and extractives determination.

5.2. Classification Results

Classification techniques were applied to NIR and MIR spectral measurements of Turkish pine (TP) and Anatolian black pine (ABP) trees. Each pine species contain 51 wood samples with their corresponding spectral data. Two different classification modeling techniques were chosen; singular value decomposition based principal component analysis (SVD–PCA) and genetic algorithm as a wavelength selection in

principal component analysis with distance criterion (GAPCA–d). Then results were compared to visualize both the power of genetic algorithm and the construction of classes.

5.2.1. Near-Infrared Spectroscopy

NIR spectral measurements were considered as data matrix at the beginning of analysis, firstly SVD-PCA was performed and principal components were found. The first two PC's were taken in the construction of score and loading plots. As it is mentioned previously, PCA generally forces the first two principal components for explanation of system in 95%. Figure 5.21 shows the loading and score plots of Turkish pine and Anatolian black pine data matrix.

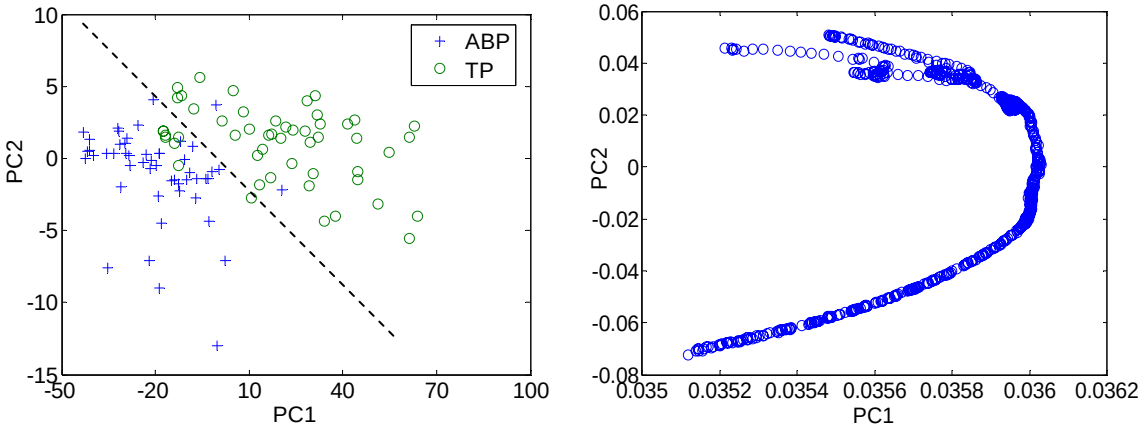


Figure 5.21. Results of SVD–PCA method, a) score and b) loading plot of pine samples measured with NIR

SVD–PCA takes the all variables in the determination of principal components. Therefore it can be seen complicated, because NIR spectrum contains spectral overlap and absorption at similar wavelengths. The homogenous distributions of wavelengths seen from the loading plot prove that there is no unexpected data point in the spectra. Only the upper left side of plot gives differences due to the baseline shift in the spectral measurements. On the other hand, score plot which indicates the classes of samples distinguishes the sample set with a few overlap samples. It can be the result of the baseline shift in NIR spectra. To reduce the spectral effects in data matrix, genetic

algorithm is used as a wavelength selection and then principal component analysis with distance criterion was applied to data set. In the result with a few wavelengths as it shown in Figure 5.22 Turkish pine and Anatolian black pine trees were distinguished much better. In this regression analysis, gene population is predefined as 20 genes with 10 iterations. In the beginning of the classification, the whole spectrum was taken and the algorithm selected the wavelengths by itself.

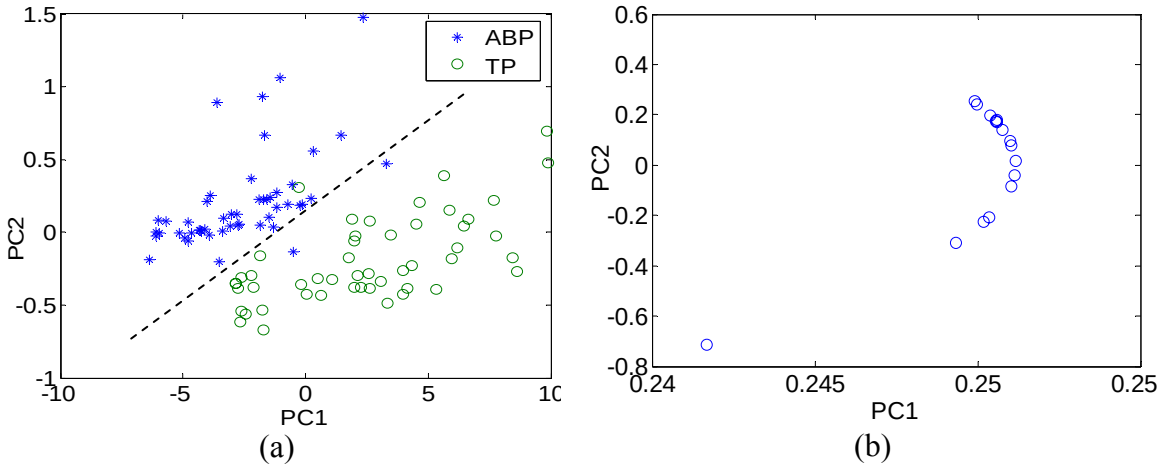


Figure 5.22. Results of GAPCA-d method, a) score and b) loading plots of pine samples measured with NIR

5.2.2. Mid-Infrared Spectroscopy

Same studies were also performed to the MIR measurements of Turkish pine and Anatolian black pine tree samples. Two separate classes were obtained from both SVD-PCA and GAPCA-d. The difference was only the number wavelengths with their corresponding absorbance values in the principal components.

In Figure 5.23, the distribution of loading plot obtained from SVD-PCA shows scattered wavelength points at the left bottom of the plot. This was due to the absorbance values in the region of 400 and 1400 cm^{-1}. As it was shown before MIR spectra of samples, in that region the fluctuation in absorbance spectra causes the heterogeneity in the loading plot.

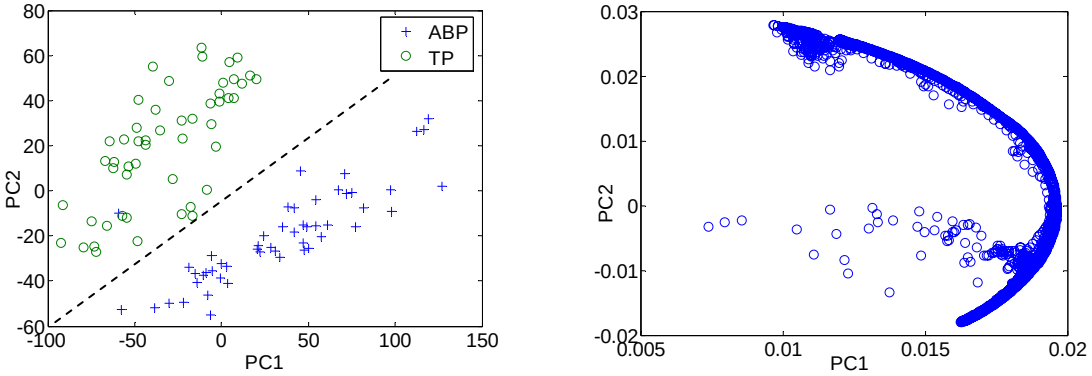


Figure 5.23. Results of SVD–PCA method, a) score and b) loading plot of pine samples measured with MIR

Loading plot of principal components especially is used in the explanation of which variables are most influential or most correlated in the determination of the classes. All the relationships of variables are shown in the same graph. MIR measurements of tree samples show that classes can be constructed with a few wavelengths. Figure 5.24 indicates the selected wavelengths in MIR spectrum and descriptive ability of GAPCA-d. Two different tree classes were obtained without any overlap samples. In the beginning of regression analysis only 6 genes with 10 iterations was utilized and full spectrum was used in the analysis. The whole spectrum contains 3601 wavelengths. SVD-PCA used all the wavelengths in the construction of classes whereas GAPCA-d only selected 30 wavelengths.

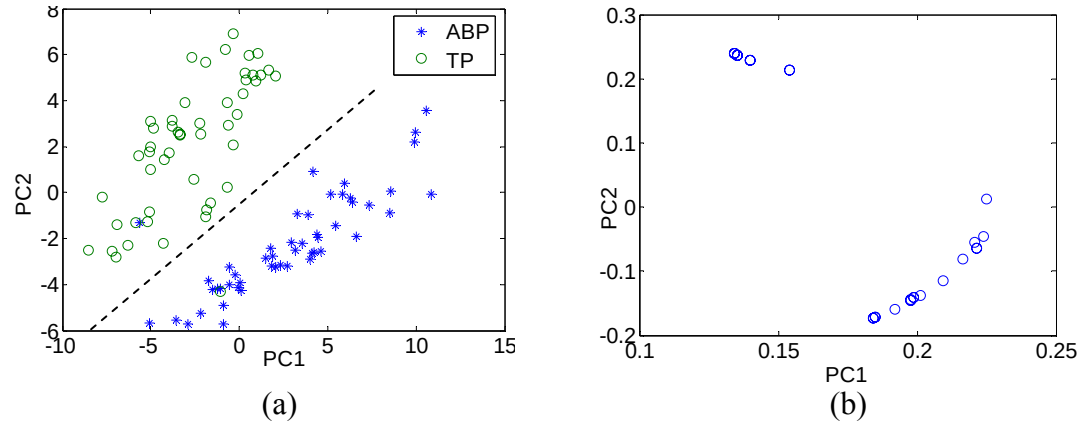


Figure 5.24. Results of GAPCA-d method, a) score and b) loading plot of pine samples measured with MIR

As a result, we can conclude that the Turkish pine and Anatolian black pine tree samples with MIR spectroscopic techniques much better than NIR measurements in the classification. On the other hand, GAPCA-d does not only classify the samples but also select a few wavelengths which contains the necessary information. In the future, the algorithm can be improved by adding validation steps after construction of models.

CHAPTER 6

CONCLUSION

In this study, calibration models were developed for extractives and lignin contents of Turkish pine and Anatolian black pine trees by coupling infrared spectroscopy and multivariate calibration. Samples were analyzed in both near-infrared and mid-infrared regions by using diffuse reflectance measurements. To construct calibration models, GILS was used as a multivariate calibration method. Reliability of the calibration models was determined by SEC and SEP values as well as with the R^2 values from the reference vs. predicted content plots. From the results, it is seen that successful calibration models can be constructed by using the methods mentioned to provide fast and nondestructive determination of extractives and lignin contents. This might give rise to improvements in the forest industry in economical manner. In addition, by wavelength selection feature of GILS method, the wavelengths which carry information of extractives and lignin contents could be determined in order to develop case specific analysis models.

Classification of pine samples were performed by SVD-PCA and GAPCA. Wavelength selection feature of GAPCA method enhances the success of classification. The results show that eliminating wavelengths without information make the classification more successful. The application of this part of the study might be that an unknown pine sample could be predicted from its NIR or MIR spectrum whether it belongs to Turkish pine species or Anatolian black pine species.

REFERENCES

Arnold, S.A., Crowley, J., Vaidyanathan, S., Matheson, L., Mohan, P., Hall, J.W., Harvey, L.M., McNeil, B. 2000. At-line monitoring of a submerged filamentous bacterial cultivation using near infrared spectroscopy. *Enzyme and Microbial Technology* 27:691-697.

Beebe K.R., Pell. R.J., and Seasholtz M.B. 1998. *Chemometrics, a practical guide.* Wiley-Interscience: John Wiley & Sons, Inc.

Biermann, C.J. 1996. *Handbook of pulping and papermaking (second edition).* Academic Press.

Brereton R.G. 2000. Introduction to multivariate calibration in analytical chemistry. *The Analyst* 125:2125-2154.

Brereton R.G. 2003. *Chemometrics, data analysis for the laboratory and chemical plant.* Wiley: John Wiley & Sons Ltd.

Cong, P. and Li, T. 1994. Numeric genetic algorithm part I. theory, algorithm and simulated experiments. *Analytica Chimica Acta* 293:191-203.

Dang, V.Q., Bhardwaj, N.K., Hoang, V., Nguyen, K.L. 2007. Determination of lignin content in high-yield kraft pulps using photoacoustic rapid scan Fourier transform infrared spectroscopy. *Carbohydrate Polymers* 68:489-494.

DeThomas, F.A., Hall, J.W., Monfre, S.L. 1994. Real-time monitoring of polyurethane production using near infrared spectroscopy. *Talanta* 41:425-431.

Fontain, E. 1992. The problem of atom-to-atom mapping. An application of genetic algorithms. *Analytica Chimica Acta* 265:227-232.

Gilbert R. J., Goodacre, R., Woward, A. N., and Kell, D.B. 1997. Genetic programming: a novel method for the quantitative analysis of pyrolysis mass spectral data. *Analytical Chemistry 69*:4381-4389.

Griffiths, P.R. 1978. *Chemical infrared Fourier transform spectroscopy*, New York: Wiley.

Haaland, D.M. and Thomas, E.V. 1988. Partial least-squares methods for spectral analyses. 1. relation to other quantitative calibration methods and the extraction of qualitative information. *Analytical Chemistry* 60:1193-1202.

Hauksson, J.B., Bergqvist, G., Bergsten, U., Sjöström, M., Edlund, U. 2001. Prediction of basic wood properties for Norway spruce. interpretation of near infrared spectroscopy data using partial least squares regression. *Wood Science and Technology* 35:475-485.

Hibbert, D.B. 1993. Genetic algorithms in chemistry. *Chemometrics and Intelligent Laboratory Systems* 19:277-293.

Jones, P.D., Schimleck, L.R., Peter, G.F., Daniels, R.F., Clark III, A. 2006. Nondestructive estimation of wood chemical composition of sections of radial strips by diffuse reflectance near infrared spectroscopy. *Wood Science and Technology* 40:709-720.

Kelley, S.S., Rials, T.G., Snell, R., Groom, L.H., Sluiter, A. 2004. Use of near infrared spectroscopy to measure the chemical and mechanical properties of solid wood. *Wood Science Technology* 38:257-276.

Koenig, J.L. 1975. Application of Fourier transform infrared spectroscopy to chemical systems. *Applied Spectroscopy* 29:293-308.

Kowalski, B.R. 1984. *Chemometrics. Mathematics and statistics in chemistry*. D. Reidel Publishing.

Li, T., Lucasius, C.B., and Kateman, G. 1992. Optimization of calibration data with the dynamic genetic algorithm. *Analytica Chimica Acta* 268:123-134.

Lucasius, C.B. and Kateman, G. 1991. Genetic algorithms for large-scale optimization in chemometrics: an application. *Trends in Analytical Chemistry* 10:254-261.

Lucasius, C.B., Beckers, M.L.M., and Kateman, G. 1994. Genetic algorithms in wavelength selection: a comparative study. *Analytica Chimica Acta* 286:135-153.

Martens, H., and Naes, T. 1989. *Multivariate calibration.* Wiley: John Wiley & Sons Ltd.

Massart, D.L., Vandeginste, B.G.M., Buydens, L.M.C., De Jong, S., Lewi, P.J., Smeyers-Verbeke, J. 1998. *Part B: Handbook of chemometrics and qualimetrics*. Elsevier.

Nuopponen, M.H., Birch, G.M., Sykes, R.J., Lee, S.J., Stewart, D. 2006. Estimation of wood density and chemical composition by means of diffuse reflectance mid-infrared Fourier transform (DRIFT-MIR) spectroscopy. *Journal of Agricultural Food Chemistry.* 54:34-40.

Ozdemir, D., Mosley, R.M., and Williams, R.R. 1998a. Hybrid calibration models an alternative to calibration transfer. *Applied Spectroscopy* 52:599-603(5).

Ozdemir, D., Mosley, R.M., and Williams, R.R. 1998b. Effect of wavelength drift on single- and multi-instrument calibration using genetic regression. *Applied Spectroscopy* 52: 1203-1209(7).

Ozdemir, D. and Williams, R.R. 1999. Multi-instrument calibration with genetic regression in UV-visible spectroscopy. *Applied Spectroscopy* 53:210-217(8).

Öztürk, B. 2003. Monitoring the esterification reactions of carboxylic acids with alcohols using near infrared spectroscopy and multivariate calibration methods. *Izmir Institute of Technology thesis of M.Sc.*

Paradkar, R.P. and Williams, R.R. 1996. Genetic regression as a calibration technique for solid-phase extraction of dithizone-metal chelates. *Applied Spectroscopy* 50:753-758(6).

Paradkar, R.P. and Williams, R.R. 1997. Correcting fluctuating baselines and spectral overlap with genetic regression. *Applied Spectroscopy* 51:92-100(9).

Poke; F.S., Wright, J.K., Raymond, C.A. 2005. Predicting extractives and lignin contents in *Eucalyptus globulus* using near infrared reflectance analysis. *Journal of Wood Chemistry and Technology* 24:55-67.

Poke, S.P., Raymond, C.A. 2006. Predicting extractives, lignin, and cellulose contents using near infrared spectroscopy on solid wood in *Eucalyptus globulus*. *Journal of Wood Chemistry and Technology* 26:187-199.

Schultz, T.P., Templeton, M.C., McGinnis, G.D. 1985. Rapid determination of lignocellulose by diffuse reflectance Fourier transform infrared spectrometry. *Analytical Chemistry* 57:2867-2869.

Sherman, C.P. 1997. *in "Handbook of instrumental techniques for analytical chemistry" (F. Settle, ed.)*, Chap. 15. New Jersey: Prentice Hall.

Skoog, D.A., Holler, F.J., Nieman, T.A. 1998. *Principles of instrumental analysis – fifth edition*. Philadelphia: Saunders College Publishing, Harcourt Brace College Publishers.

Smith, B.C. 1996. *Fundamentals of Fourier transform infrared spectroscopy*. New York: CRC Press.

Tran, C.D., Oliveira, D., Grishko, V.I. 2004. Determination of enantiomeric compositions of pharmaceutical products by near-infrared spectrometry. *Analytical Biochemistry* 325:206-214.

Wang, Y., Veltkamp, D. J., and Kowalski, B. R. 1991. Multivariate instrument standardization. *Analytical Chemistry* 63:2750-2756.

Wienke, D., Lucasius, C. B., Ehrlich, M., and Kateman, G. 1993. Multicriteria target vector optimization of analytical procedures using a genetic algorithm: part II. polyoptimization of the phothometric calibration graph of dry glucose sensors for quantitative clinical analysis. *Analytica Chimica Acta* 271:253-268.

Wikipedia contributors, "Chemometrics". 2008. Wikipedia, The Free Encyclopedia. http://en. wikipedia.org/wiki/Chemometrics (accessed May 27, 2008).

Wood Growth and Structure. 2007. Farm Forest Line. http://www.farmforestline.com.au/pages/2.1.2.1_wood.html (accessed May 23, 2008).

Yeh, T., Chang, H., Kadla, J.F. 2004. Rapid prediction of solid wood lignin content using transmittance near-infrared spectroscopy. *Journal of Agricultural Food Chemistry* 52:1435-1439.

Yeh, T., Yamada, T., Capanema, E., Chang, H., Chiang, V., Kadla, J.F. 2005. Rapid screening of wood chemical component variations using transmittance near-infrared spectroscopy. *Journal of Agricultural Food Chemistry* 53:3328-3332.

Zobel, B.J., van Buijtenen, J.P. 1989. *Wood variation: its causes and control*. Springer-Verlag, Berlin.