ORIGINAL

# Determination of lignin and extractive content of Turkish Pine (*Pinus brutia* Ten.) trees using near infrared spectroscopy and multivariate calibration

**B. Üner · İ. Karaman · H. Tanrıverdi · D. Özdemir**

**Abstract**   Determination of quality parameters such as lignin and extractive content of wood samples by wet chemistry analyses takes a long time. Near-infrared (NIR) spectroscopy coupled with multivariate calibration offers a fast and nondestructive alternative to obtain reliable results. However, due to the complexity of the NIR spectra, some wavelength selection is generally required to improve the predictive ability of multivariate calibration methods. *Pinus brutia* Ten. is the most growing pine species in Turkey. Its rotation period is around 80 years; the forest products industry has widely accepted the use of *Pinus brutia* Ten. because of its ability to grow on a wide range of sites and its suitability to produce desirable products. *Pinus brutia* Ten. is widely used in construction, window door panel, floor covering, etc. Determination of lignin and extractive content of wood provides information to tree breeders on when to cut and how much chemicals are needed for the pulping and bleaching process. In this study, 58 samples of *Pinus brutia* Ten. trees were collected in Isparta region of Turkey, and their lignin and extractive content were determined with standard reference (TAPPI) methods. Then, the same samples were scanned with near-infrared spectrometer between 1,000 and 2,500 nm in diffuse reflectance mode, and multivariate calibration models were built with genetic inverse least squares method for both lignin and extractive content using the concentration information obtained from wet standard reference method. Overall, standard error of calibration (SEC) and standard error of prediction (SEP) ranged between 0.35% (w/w) and 2.40% (w/w).

B. Üner (✉) · H. Tanrıverdi
Faculty of Forestry, Forest Products Engineering Department,
Suleyman Demirel University, 32260 Isparta, Turkey
e-mail: buner@orman.sdu.edu.tr

İ. Karaman · D. Özdemir
Faculty of Science, Department of Chemistry,
Izmir Institute of Technology, Gülbahce, 35430 Urla, Izmir, Turkey

## Introduction

Wood is composed of cellulose, lignin, hemicelluloses and extractive materials. Lignin and extractive content of wood are important constituents for the pulping and papermaking industry. They are unwanted materials and need to be removed from the pulp. The presence of extractive materials and lignin in paper cause discoloration, and the paper quality goes down. Elimination of these chemicals is carried out by chemical methods. After elimination, mainly long chain carbohydrates remain in the pulp (Poke and Raymond 2006; Sykes et al. 2005). If lignin and extractive content are known earlier, chemicals required during pulping and bleaching can be determined. In addition to that tree-breeding programs employ different silvicultural treatments to improve tree growth (Zobel and Talbert 1984). These trees may have different chemical properties to natural grown trees. Moreover, wood samples obtained from trees in different locations show different properties (chemical compositions, morphology, juvenile wood, mature wood, etc.) (Zobel and Talbert 1984; Zobel and van Buijtenen 1989). These differences play an important role in finished products. Determination of extractive and lignin content using traditional chemical methods is a costly and time-consuming process. Near-infrared reflectance analysis offers a low cost alternative for prediction of wood chemistry and quality.

Near-infrared (NIR) spectroscopy has become a popular method for simultaneous chemical analysis and is being studied extensively in a number of different fields such as process monitoring, biotechnology, and pharmaceutical industry because of its potential for on-line, nondestructive and noninvasive instrumentation (McClure 1994; Dethomas et al. 1994; Arnold et al. 2000; Tran et al. 2004). Traditionally, NIR spectroscopy has found its widest application area in agriculture and food industry (Delwiche 1998; Ferré and Rius 1996; Ferrioa et al. 2005; Hareland 1994; Kalivas 1997; McCaig 2002; Miralbés 2004; Punchwein and Eibelhuber 1989; Sorvaniemi et al. 1993). The NIR portion of the electromagnetic spectrum covers the range from 780 nm to 2,500 nm, and most of the absorption bands observed in this region are due to overtones and combinations of the fundamental mid-IR molecular vibrational bands. Although all the fundamental vibrational modes can have overtones, the most commonly observed bands arise from the C–H, O–H, and N–H bonds in the molecules. NIR spectroscopy is also used for the chemical and mechanical characterization of different wood species (Hedrick et al. 2007; So and Eberhardt 2006; Schimleck et al. 2006; Tsuchikawa et al. 2005; Cogdill et al. 2004; Jonsson et al. 2004).

*Pinus taeda* L. (loblolly pine) is a very important species in the Southern United States. Its wood is mainly utilized in pulping and construction industry. Therefore, many researches have been carried out on this tree species. Transmittance NIR spectra were used to determine chemical composition of solid wood of loblolly pine (Yeh et al. 2005). Strong correlation was found between wet chemistry and the NIR spectra. In the same study, sample preparation and quantity of sample were also discussed. It was found that the source of error is mainly coming from the wet chemistry. NIR spectra along with multivariate analysis were used to determine chemical and mechanical properties of loblolly pine (Kelley et al. 2004). Samples

were selected from different locations and height of the tree. The correlation coefficient for lignin and extractives was 0.80. These measurements were carried out between 500 nm and 2,400 nm. When spectral range was narrowed (650–1,150 nm), no change occurred in the correlation coefficient. The chemical composition of *Pinus taeda* L. was investigated with NIR spectroscopy (Jones et al. 2006). Diffuse reflectance NIR spectroscopy and radial strips from seven different locations were used. Strips represented juvenile wood, transition zone between mature and juvenile wood, and mature wood. Prediction errors were high. It was concluded that different origin of the samples was affecting the result.

Near-infrared spectroscopy coupled with multivariate calibration offers a fast and nondestructive method to determine chemical properties of lignocellulosic materials. However, due to the complexity of the NIR spectra, wavelength selection is generally required to improve the predictive ability of multivariate calibration methods (Özdemir 2006). Selecting the most suitable calibration method is very important in order to generate calibration models with high predictive ability for unknown samples. The major drawback of classical least squares (CLS) method is that all of the interfering species must be known and their concentrations included in the model. This need can be eliminated by using inverse least squares (ILS) method, which uses the inverse of Beer's Law. Genetic inverse least squares (GILS) is the modified version of original ILS method in which a small set of wavelengths is selected from a full spectral data matrix and evolved to an optimum solution using a genetic algorithm (GA) applied to a number of wavelength selection problems (Özdemir and Dinç 2004; Özdemir and Öztürk 2004; Özdemir 2005). GAs are nonlocal search and optimization methods that are based upon the principles of natural selection (Hibbert 1993; Paradkar and Williams 1997; Pizarro et al. 1998; Mosley and Williams 1998; Özdemir and Williams 1999).

GILS has some major advantages over 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. By selecting a subset of instrument responses, it is able to eliminate nonlinearities that might be present in the full spectral region.

The objective of this paper is to determine lignin and extractive content of *Pinus brutia* Ten. with NIR spectroscopy and multivariate calibration methods. *Pinus brutia* Ten. is the most common pine species in Turkey. Rotation period of this species is around 60–80 years. Therefore, it is considered to be the fastest growing pine species in Turkey and the forest products industry has widely accepted the use of Turkish pine (*Pinus brutia* Ten.) because of its ability to grow on a wide range of sites and its suitability for the manufacture of desirable products.

## Genetic inverse least squares

The major drawback of classical least squares (CLS) method is that all of the interfering species must be known and their concentrations included in the model. This can be eliminated by using the inverse least squares (ILS) method, which uses

the inverse of Beer's Law. In the ILS method, concentration of a component is modeled as a function of absorbance measurements. Because modern spectroscopic instruments are very stable and provide excellent signal-to-noise (S/N) ratios, it is believed that the majority of errors lies in the reference values of the calibration sample and not in the measurement of their spectra. In fact, in many cases, the concentration data of calibration set is generated from another analytical technique that already has its inherent errors which might be higher than those of the spectrometer (for example, Kjeldahl protein analysis used to calibrate NIR spectra). The ILS model for $m$ calibration samples with $n$ wavelengths for each spectrum is described by:

$$\mathbf{C} = \mathbf{AP} + \mathbf{E_C} \tag{1}$$

where $\mathbf{C}$ is the $m \times l$ matrix of the component concentrations, $\mathbf{A}$ is the $m \times n$ matrix of the calibration spectra, $\mathbf{P}$ is the $n \times l$ matrix of the unknown calibration coefficients relating $l$ component concentrations to the spectral intensities, and $\mathbf{E_C}$ is the $m \times l$ matrix of errors in the concentrations that do not fit by the model. In the calibration step, ILS minimizes the squared sum of the residuals in the concentrations. The biggest advantage of ILS is that Eq. 1 can be reduced for the analysis of single component at a time since analysis is based on an ILS model which is invariant with respect to the number of chemical components included in the analysis. The reduced model is given as:

$$\mathbf{c} = \mathbf{Ap} + \mathbf{e_c} \tag{2}$$

where $\mathbf{c}$ is the $m \times 1$ vector of concentrations for the component that is analyzed, $\mathbf{p}$ is the $n \times 1$ vector of calibration coefficients, and $\mathbf{e_c}$ is the $m \times 1$ vector of concentration residuals. During the calibration step, the least squares estimate of $\mathbf{p}$ is:

$$\hat{\mathbf{p}} = (\mathbf{A'A})^{-1}\mathbf{A'} \cdot \mathbf{c} \tag{3}$$

where $\hat{\mathbf{p}}$ is the vector of estimated calibration coefficients. Once $\hat{\mathbf{p}}$ is calculated, concentration of the analyte of interest can be predicted with the equation below.

$$\hat{c} = \mathbf{a'} \cdot \hat{\mathbf{p}} \tag{4}$$

where $\hat{c}$ is the scalar estimated concentration, and $\mathbf{a}$ is the spectrum of the unknown sample. The ability to predict one component at a time without knowing the concentrations of interfering species has made ILS one of the most frequently used calibration method. The major disadvantage of ILS is that the number of wavelengths in the calibration spectra should not be more than the number of calibration samples. This is a big restriction since the number of wavelengths in a spectrum will generally be much more than the number of calibration samples and the selection of wavelengths that provide the best fit for the model is not a trivial process. Several wavelength selection strategies, such as stepwise wavelength selection and all possible combination searches, are available to build an ILS model which fits the data best. GAs are global search and optimization methods based upon the principles of natural evolution and selection as developed by Darwin. Computationally, the implementation of a typical GA is quite simple and consists of five basic steps including initialization of a gene population, evaluation

of the population, selection of the parent genes for breeding and mating, crossover and mutation, and replacing parents with their offspring. These steps have taken their names from the biological foundation of the algorithm. Genetic inverse least squares (GILS) is an implementation of a GA for selecting wavelengths to build multivariate calibration models with a reduced data set. GILS follows the same basic initialize/breed/mutate/evaluate algorithm like other GAs to select a subset of wavelengths but it is unique in the way it encodes genes. A gene is a potential solution to a given problem and the exact form may vary from application to application. Here, the term "gene" is used to describe the collection of instrumental response at the wavelength range given in the data set. The term "population" is used to describe the collection of individual genes in the current generation. In the initialization step, the first generation of genes is created randomly with a fixed population size. Although random initialization helps to minimize bias and maximize the number of possible recombinations, GILS is designed to select initial genes in a somewhat biased random fashion in order to start with genes better suited to the problem than those that would be randomly selected. Biasing is done with a correlation coefficient by plotting the predicted results of initial population against the actual component concentrations. The size of the gene pool is a user-defined even number in order to allow breeding of each gene in the population. It is important to note that the larger the population size, the longer the computation time. The number of instrumental responses in a gene is determined randomly between a fixed low limit and high limit. The lower limit was set to 2 in order to allow single-point crossover whereas the higher limit was set to eliminate overfitting problems and reduce the computation time. Once the initial gene population is created, the next step is to evaluate and rank the genes using a fitness function, which is the inverse of the standard error of calibration (SEC).

The third step is where the basic principle of natural evolution is put to work for GILS. This step involves the selection of the parent genes from the current population for breeding using a roulette wheel selection method according to their fitness values. The goal is to give a higher chance to those genes with high fitness so that only the best performing members of the population will survive in the long run and will be able to pass their information to the next generations. Because of the random nature of the roulette wheel selection method, however, genes with low fitness values will also have some chance to be selected. Also, there will be genes that are selected multiple times and some genes that will not be selected at all and will be thrown out of the gene pool. After the selection procedure is completed, the selected genes are allowed to mate top–down in pairs whereby the first gene mates with the second gene and the third one with the fourth one and so on as illustrated in the following example:

Parents

$$S_1 = (A_{1147}, A_{951}, \#A_{2179}, A_{2218}) \tag{5}$$

$$S_2 = (A_{1225}, A_{1478}, \#A_{1343}, A_{950}, A_{1451}, A_{2358}, A_{931}, A_{1158}). \tag{6}$$

The points where the genes are cut for mating are indicated by $\#$.

Offspring

$$S_3 = (A_{1147}, A_{951}, A_{1343}, A_{950}, A_{1451}, A_{2358}, A_{931}, A_{1158}) \tag{7}$$

$$S_4 = (A_{2179}, A_{2218}, A_{1225}, A_{1478}) \tag{8}$$

where $A_{1147}$ represents the instrument response at the wavelength given in subscript, $S_1$ and $S_2$ represent the first and second parent genes, and $S_3$ and $S_4$ are the corresponding genes for the offspring. Here, the first part of $S_1$ is combined with the second part of the $S_2$ to give the $S_3$; likewise the second part of the $S_1$ is combined with the first part of the $S_2$ to give $S_4$. This process is called single-point crossover and is common in GILS. Single-point crossover will not provide different offspring if both parent genes are identical, which may happen in roulette wheel selection, when both genes are broken at the same point. Also note that mating can increase or decrease the number of instrument responses in the offspring genes. After crossover, the parent genes are replaced by their offspring and the offspring are evaluated. The ranking process is based on their fitness values following the evaluation step. Then, the selection for breeding/mating starts all over again. This is repeated until a predefined number of iterations is reached.

Mutation which introduces random deviations into the population was also introduced into the GILS during the mating step at a rate of 1% which is typical in GAs. This is usually done by replacing one of the responses in an existing gene with a randomly selected new one. Mutation allows the GILS to explore the search space and incorporate new material into the genetic population. It helps keep the search moving and can eject GILS from a local minimum on the response surface. However, it is important not to set the mutation rate too high since it may keep the GA from being able to exploit the existing population. Also, the GILS method is an iterative algorithm and therefore there is a high possibility that the method may easily overfit the calibration data so that the predictions for independent sets might be poor. To eliminate possible overfitting problems, cross validation is used in which one spectrum is left out of the calibration set and the model is constructed with $m - 1$ sample. Then, this model is used to predict the concentration of the left out sample. This process is continued until all samples are left out at least once in each iteration. As long as the number of spectra in the calibration set is not too large, cross validation is an effective method of eliminating overfitting. If the number of calibration spectra is very large, then the GILS method has the option of half validation approach in which the half of the spectra in the calibration set is used to validate the model in each iteration.

In the end, the gene with the lowest SEC (highest fitness) is selected for the model building and this model is used to predict the concentrations of component being analyzed in the prediction (test) sets. The success of the model in the prediction of the test sets is evaluated using standard error of prediction (SEP). Because random processes are heavily involved in GILS as in all of the GAs, the program has been set to run several times for each component in this study. The best run (i.e., the one generating the lowest SEC for the calibration set and at the same time producing SEPs for prediction sets that are in the same range with the SEC) is subsequently selected for evaluation and further analysis. The termination of the

algorithm can be done in many ways. The easiest way is to set a predefined iteration number for the number of breeding/mating cycles.

## Materials and methods

Materials

Turkish pine (*Pinus brutia* Ten.) tree samples used for the study were collected from Ağlasun, Aşağıgökdere Isparta Turkey. The trees were sampled from plots ranged in the same elevation of 800 m with average precipitation of 0.515 m and average July maximum temperature of 30.3°C and average January temperature of 1.8°C. Trees aged between 17 and 25 were used in this study and their diameter ranged from 0.15 to 0.25 m. Tree selection was based on good form trees. Eccentric piths were not used. Samples were taken from the breast height section to determine chemical contents. A total of 58 trees were examined with two different sampling seasons. Among 58 samples, 21 of them were collected as the first party, and initial wet chemical analyses were performed on these samples. Then, the remaining 37 were analyzed as a second party in a later period. All tree samples wait in bed for 3 years, then they were planted in field. Twelve years later, thinning was applied.

Methods

Extractive and lignin content of wood were determined according to TAPPI standard test method T204 om-88 and T222 om-88. Wood meal samples were prepared using a Wiley mill and ground to pass various mesh screens. In order to determine extractive content, ethanol–benzene solution (1:2 v/v) was used. Extraction was carried out in a Soxhlet apparatus for 6 h. In order to determine acid insoluble lignin content of wood, the carbohydrates in wood were hydrolyzed and solubilized with sulfuric acid according to TAPPI standard test method. Acid insoluble lignin is filtered off, dried and the content was measured.

Near-infrared spectroscopic measurements of the samples were carried out in diffuse reflectance mode between 1,000 and 2,500 nm wavelength interval. Background correction is performed by a gold-coated disk. Near-infrared spectroscopic analyses were performed with FTS-3000 NIR spectrometer (Bio-Rad, Excalibur, Cambridge, MA). Resolution was 16 cm$^{-1}$, and number of scans is 128 for each spectrum. Three spectra were taken for each sample and the means of three corresponding spectra were used in multivariate analyses. Initially, the wood meal samples were allowed to pass through 300-μm mesh screen to obtain uniform particle size. This was needed because nonuniform particle size might affect the absorbance measurements. Then, the samples were dried in an incubator for 24 h to obtain uniform humidity. For the measurements, a diffuse reflectance accessory (Pike Technologies, DiffusIR Accessory) was used. Wood meal samples were placed into micro sample cup (6.0 mm diameter, 1.6 mm depth) cautiously making the surface as flat as possible to minimize absorbance changes due to the surface.

Software

After collection of raw spectra, all of the data were transferred into a Microsoft Excel worksheet where text files of calibration sets and independent validation sets were prepared. Multivariate calibration models were built with GILS method for extractive and lignin content for first 21 samples and the remaining 37 samples separately and also for the combination of both the data sets. The GILS method was implemented in MATLAB programming language Version 7.0 (MathWorks Inc, Natick, MA).

## Results and discussion

Details of the data sets used to construct and verify the GILS models are demonstrated in Table 1. Wood is an organic living material and its properties change from tree to tree depending on growth condition. This change can be seen in Table 1 and average lignin content was $29.19 \pm 3.16\%$ (w/w) and extractive content was $7.76 \pm 2.43\%$ (w/w). Using wet chemical methods to determine the chemical contents of samples, a lot of chemicals need to be utilized and it takes a long time to determine. For that reason, a rapid alternative analysis method would be advantageous in order to give a general idea about extractive and lignin content of the unknown samples for routine analysis. NIR spectroscopy is a good candidate for that purpose when combined with chemometric methods. On the other hand, the success of NIR analysis heavily relies on accurate wet chemical determinations for the training process of the multivariate calibration models.

**Table 1** Summary of lignin and extractive contents data of Turkish pine (*Pinus brutia* Ten.) tree samples used for calibrations and validations of all three sets

|  | All of the samples (58) | First set | | Second set | | Third set | |
|---|---|---|---|---|---|---|---|
|  |  | Calibration set (14) | Validation set (7) | Calibration set (24) | Validation set (13) | Calibration set (38) | Validation set (20) |
| **Lignin % (w/w)** | | | | | | | |
| Mean | 29.19 | 28.88 | 29.01 | 28.64 | 30.63 | 28.76 | 30.00 |
| SD | 3.16 | 0.21 | 0.18 | 4.85 | 3.94 | 3.83 | 3.25 |
| Min | 19.93 | 28.44 | 28.67 | 19.93 | 23.50 | 19.93 | 23.50 |
| Max | 36.70 | 29.27 | 29.20 | 36.70 | 35.46 | 36.70 | 35.46 |
| **Extractive % (w/w)** | | | | | | | |
| Mean | 7.76 | 6.53 | 6.64 | 8.11 | 9.07 | 7.53 | 8.22 |
| SD | 2.43 | 0.58 | 0.70 | 3.27 | 1.53 | 2.71 | 1.74 |
| Min | 2.05 | 5.73 | 5.85 | 2.05 | 7.21 | 2.05 | 5.85 |
| Max | 16.12 | 7.89 | 7.53 | 16.12 | 12.06 | 16.12 | 12.06 |

Number of samples for each data set is in parentheses
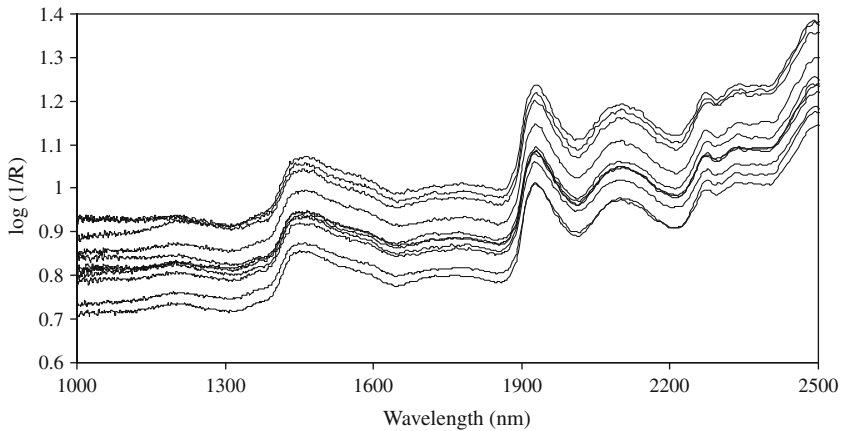
**Fig. 1** Near-infrared diffuse reflectance spectra of 10 representative Turkish pine (*Pinus brutia* Ten.) samples for the first and second set between 1,000 and 2,500 nm

Near-infrared diffuse reflectance spectra of 10 wood samples among the 58 samples are shown in Fig. 1. It is evident that the samples yield high absorbance values around 1,450, 1,900, and 2,100 nm wavelength regions. Among these bands, 1,450 and 2,100 nm correspond to the characteristic absorption region of lignin and cellulose, respectively. Both lignin and cellulose give rise to absorption around 1,900 nm. However, since they are all pine wood samples, their spectral characteristics are very much alike except the baseline differences among the samples. This type of baseline shifts in the absorbance scale is quite common in diffuse reflectance spectroscopy and part of it is due to sample composition differences and part of it is due to sample inhomogeneities. Because the GILS method used here is a genetic algorithm–based multivariate calibration technique, it was expected that it can select certain combinations of wavelengths which have maximum correlation with extractive and lignin content of the samples.

In order to construct NIR spectroscopic multivariate calibration models for extractive and lignin content, three different calibration sets were prepared. The first calibration set was generated from the above mentioned 21 samples. This data set is called the first data set in which 14 of them were randomly selected for calibration set including the samples having minimum and maximum extractive and lignin contents. The remaining seven samples were reserved for independent test samples. As can be seen from Table 1, minimum and maximum extractive and lignin contents were not the same samples, and, therefore, different samples were used in both calibration and validation sets for extractive and lignin content. Reference extractive and lignin contents versus predicted values based on NIR spectra using GILS method are shown in Fig. 2 for the first data set. Calibration models for lignin content determination gave standard error of calibration (SEC) and standard error of prediction (SEP) values of 0.07% (w/w) and 0.10% (w/w) for calibration and independent test sets, respectively. In the case of extractive content determination, the SEC and SEP values were 0.09% (w/w) and 0.26% (w/w) for calibration and
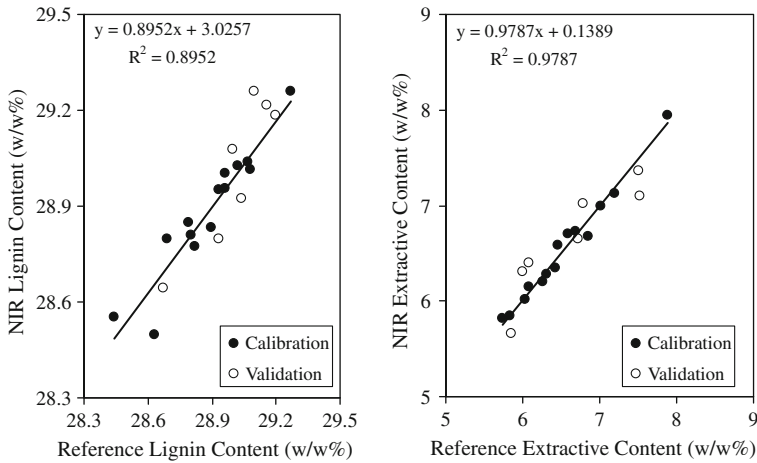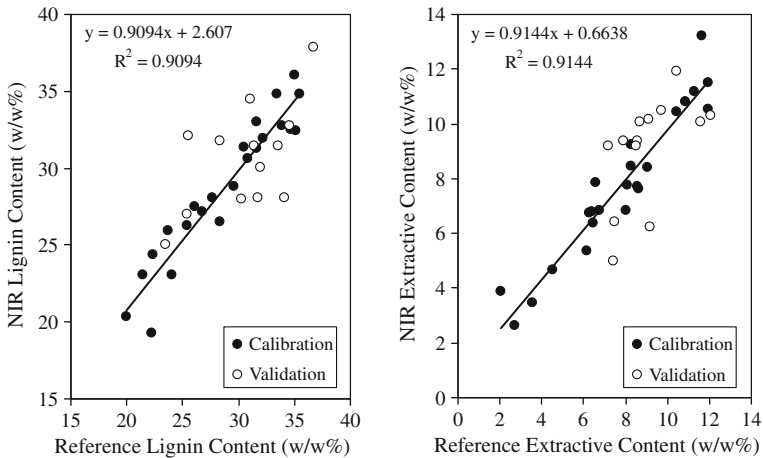
**Fig. 2** Reference versus NIR-predicted lignin and extractive content of Turkish pine (*Pinus brutia* Ten.) samples for the first data set

prediction sets, respectively. The $R^2$ value of regression lines for lignin was 0.89 and that for extractive content was 0.98.

When these SEC and SEP errors are examined, it can be seen that the values for lignin content were comparable even though the SEP value is slightly larger than 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 earlier, NIR spectra of these samples suffer from somewhat large baseline fluctuations 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 is examined, it is possible to state that the NIR spectra do contain quantitative information that is correlated with extractive and lignin content of the wood samples studied here. It is also worth to mention here that the results of GILS can only be as good as the result of wet chemical analysis since the model building is based on the reference values obtained from it.

Figure 3 shows the reference lignin and extractive contents versus GILS-predicted values for the second data set with 37 samples of which 25 were 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 ranged between 28.44 and 29.27% (w/w) for the first data set, the lower and upper level of lignin content in second data set ranged between 19.93 and 36.70% (w/w) resulting in a much wider concentration range. In addition, the extractive content of the samples in the first data set was distributed between 5.73 and 7.89% (w/w) which is also quite narrow when compared with the second data set in which the extractive content of the samples ranged between 2.05 and 12.06% (w/w). It is clear that there
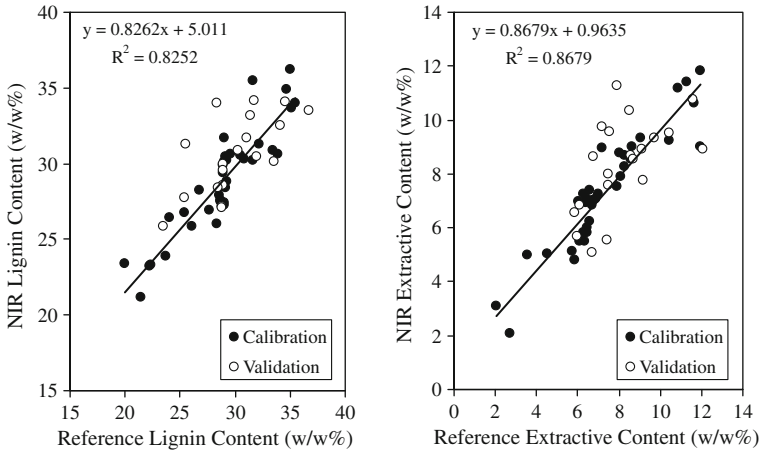
**Fig. 3** Reference versus NIR-predicted lignin and extractive content of Turkish pine (*Pinus brutia* Ten.) samples for the data second set

is a serious difference in the concentration ranges of lignin and extractive contents of the first and second data sets. These could be due to samples origin and/or season as the samples were collected from different areas and seasons in the first party and the second party. The SEC values for lignin and extractive content were 1.51 and 0.85% (w/w), respectively, while the SEP values ranged between 3.28 and 1.62% (w/w) for lignin and extractive content, respectively. When SEC and SEP values were examined in the second data set, it can be seen that the values are larger than the ones obtained for the first data set; however, it should be noted that the concentration ranges in the second data set are much larger among the samples especially in the upper limit and therefore larger SEC and SEP values are expected. The $R^2$ values of regression lines for both lignin and extractive content were 0.91.

The third data set analyzed in this study was formed by combining the first and the second data sets into a single set as given in Table 1. 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 lignin and extractive content are given in Fig. 4.

When compared with the analysis of the first data set, the results of SEC and SEP of the third data set appears somewhat larger. As explained earlier, both lignin and extractive contents of the samples in the first data set were narrower than the second set. As a result, smaller SEC and SEP values were expected. On the other hand, these results were very similar to the second data set. For the determination of lignin content, SEC and SEP values were 1.60 and 2.94% (w/w), respectively. In the case of extractive content determination, the SEC was 0.86% (w/w) and the SEP was 1.85% (w/w).

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 illustrates the frequency distribution of selected wavelengths in 100 runs with 30 genes and 50 iterations for the third data set.

**Fig. 4** Reference versus NIR-predicted lignin and extractive content of Turkish pine (*Pinus brutia* Ten.) samples for the third data set
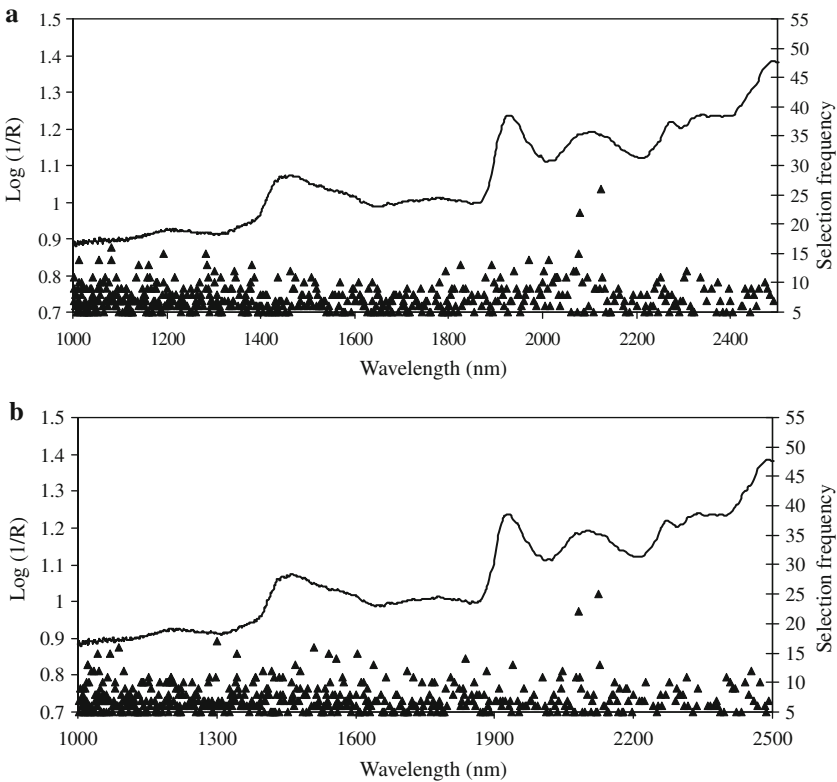


**Fig. 5** Frequency distribution of GILS selected wavelengths for both **a** lignin and **b** extractive content of Turkish pine (*Pinus brutia* Ten.) samples in the third set

As can be seen from the figure, there are a number of regions where selection frequencies are very high compared to the rest of the spectrum. The wavelength region around 2,000 nm for lignin content indicates a strong tendency for GILS method to select while for extractive content, the most frequently selected region is around 1,600 nm.

## Conclusion

The determination of lignin and extractive content of Turkish pine (*Pinus brutia* Ten.) trees was carried out with standard wet chemical methods. The results obtained from these standard methods were used as reference values for the near-infrared spectroscopic determination of these properties. Spectroscopic measurements were performed on milled wood samples in diffuse reflectance mode of NIR spectrometer and the spectra were used to generate multivariate calibration models with GILS. NIR spectroscopic determination of these properties offers much faster analysis for screening purpose. Results demonstrated that NIR spectroscopy coupled with multivariate calibration could be used for routine and fast analysis of these quality parameters of wood samples. In addition, GILS helps to improve prediction ability of the calibration models through variable selection.

## References

Arnold SA, Crowley J, Vaidyanathan S, Matheson L, Mohan P, Hall JW, Harvey LM, McNeil B (2000) At-line monitoring of a submerged filamentous bacterial cultivation using near infrared spectroscopy. Enzyme Microb Tech 27:691–697

Cogdill RP, Schimleck LR, Jones PD (2004) Estimation of the physical wood properties of *Pinus taeda* L. radial strips using least squares support vector machines. J Near Infrared Spec 12(4):263–269

Delwiche SR (1998) Protein content of single kernels of wheat by near-infrared reflectance spectroscopy. J Cereal Sci 27(3):241–254

DeThomas FA, Hall JW, Monfre SL (1994) Real-time monitoring of polyurethane production using near infrared spectroscopy. Talanta 41:425–431

Ferré J, Rius FX (1996) Selection of the best calibration sample subset for multivariate regression. Anal Chem 68:1565–1571

Ferrioa JP, Villegasb D, Zarcob J, Aparicíob N, Arausc JL, Royob C (2005) Assessment of drum wheat yield using visible and near-infrared reflectance spectra of canopies. Field Crops Res 94(2–3):126–148

Hareland GA (1994) Evaluation of flour particle size distribution by laser diffraction, sieve analysis and near-infrared reflectance spectroscopy. J Cereal Sci 20(2):183–190

Hedrick SE, Bennett RM, Rials TG (2007) Correlation of near-infrared spectroscopy measurements with the properties of treated wood. J Mater Civil Eng 19(4):279–285

Hibbert DB (1993) Genetic algorithms in chemistry. Chem Intell Lab Syst 19:277–293

Jones PD, Schimleck LR, Peter GF, Daniels RF, Clark A III (2006) Nondestructive estimation of wood chemical composition of sections of radial wood strips by diffuse reflectance near infrared spectroscopy. Wood Sci Technol 40:709–720

Jonsson P, Sjostrom M, Wallbacks L (2004) Strategies for implementation and validation of on-line models for multivariate monitoring and control of wood chip properties. J Chemometr 18(3–4):203–207

Kalivas JH (1997) Two data sets of near infrared spectra. Chem Intell Lab Syst 37(2):255–259

Kelley SS, Rials TG, Snell R, Groom LH, Sluiter A (2004) Use of near infrared spectroscopy to measure the chemical and mechanical properties of solid wood. Wood Sci Technol 38:257–276

McCaig TN (2002) Extending the use of visible/near-infrared reflectance spectrophotometers to measure colour of food and agricultural products. Food Res Int 35(8):731–736

McClure WF (1994) Near infrared spectroscopy-the giant is running. Anal Chem 66:43A–53A

Miralbés C (2004) Quality control in the milling industry using near infrared transmittance spectroscopy. Food Chem 88:622–628

Mosley RM, Williams RR (1998) Determination of the accuracy and efficiency of genetic regression. Appl Spectrosc 52:1197–1202

Özdemir D (2005) Determination of octane number of gasoline using near infrared spectroscopy and genetic multivariate calibration methods. Petroleum Sci Technol 23:1139–1152

Özdemir D (2006) Genetic multivariate calibration for near infrared spectroscopic determination of protein, moisture, dry mass, hardness and other residues of wheat. Inter J Food Sci Tech 41(Suppl 2):12–20

Özdemir D, Dinç E (2004) Determination of thiamine HCl and pyridoxine HCl in pharmaceutical preparations using uv–visible spectrophotometry and genetic algorithm based multivariate calibration methods. Chem Pharm Bull 52(7):810–817

Özdemir D, Öztürk B (2004) Genetic multivariate calibration methods for near Infrared (NIR) spectroscopic determination of complex mixtures. Turk J Chem 28:497–514

Özdemir D, Williams RR (1999) Multi-instrument calibration in uv-visible spectroscopy using genetic regression. Appl Spectrosc 53:210–217

Paradkar RP, Williams RR (1997) Genetic regression as a calibration technique for solid phase extraction of dithizone-metal chelates. Appl Spectrosc 51:92–100

Pizarro MC, Forina M, Casolino MC, Leardi R (1998) Extraction of representative subsets by potential functions methods and genetic algorithms. Chem Intell Lab Syst 40:33–51

Poke F, Raymond CA (2006) Predicting extractives, lignin, and cellulose contents using near infrared spectroscopy on solid wood in eucalyptus globulus. J Wood Chem Tech 26:187–199

Punchwein G, Eibelhuber A (1989) Outlier detection in routine analysis of agricultural grain products by near-infrared spectrometry. Anal Chim Acta 223:95–103

Schimleck LR, Kube PD, Raymond CA (2006) Extending near infrared reflectance (NIR) pulp yield calibrations to new sites and species. J Wood Chem Technol 26(4):299–311

So CL, Eberhardt TL (2006) Rapid analysis of inner and outer bark composition of Southern Yellow Pine bark from industrial sources. Holz Roh-Werkst 64(6):463–467

Sorvaniemi J, Kinnunen A, Tsados A, Mälkki Y (1993) Using partial least squares regression and multiplicative scatter correction for FT-NIR data evaluation of wheat flours. Food Sci Tech 26(3):251–258

Sykes R, Li B, Hodge G, Goldfarb B, Kadla JF, Chang H-M (2005) Prediction of loblolly pine wood properties using transmittance near-infrared spectroscopy. Can J For Res 35:2423–2431

Tran CD, Oliveira D, Grishko VI (2004) Determination of enantiomeric compositions of pharmaceutical products by near-infrared spectrometry. Anal Biochem 325:206–214

Tsuchikawa S, Hirashima Y, Sasaki Y (2005) Near-infrared spectroscopic study of the physical and mechanical properties of wood with meso- and micro-scale anatomical observation. Appl Spect 59(1):86–93

Yeh T, Yamada T, Capanema E, Chang HM, Chiang V, Kadla JF (2005) Rapid screening of wood chemical component variations using transmittance near-infrared spectroscopy. J Agric Food Chem 53:3328–3332

Zobel B, Talbert J (1984) Applied forest tree improvement. Wiley Interscience, New York

Zobel BJ, van Buijtenen JP (1989) Wood variation: its causes and control. Springer, Berlin