Wood Science and Technology (2021) 55:757–776
https://doi.org/10.1007/s00226-021-01272-y

ORIGINAL

Utilization of genetic algorithms to optimize *Eucalyptus globulus* pulp yield models based on NIR spectra

Tu X. Ho[1] · Laurence R. Schimleck[1] · Arijit Sinha[1]

Received: 5 August 2020 / Accepted: 10 February 2021 / Published online: 8 March 2021
© The Author(s), under exclusive licence to Springer-Verlag GmbH, DE part of Springer Nature 2021

Abstract

An optimization problem was developed by using a genetic algorithm to select wavelengths for establishing multivariate calibration models based on partial least squares (PLS) regression. Two near infrared (NIR) data sets represented by untreated and second derivative spectra were used to predict *Eucalyptus globulus* pulp yield. The optimization process was run with the number of variables (i.e., wavelengths) varied from 10 to 100 to determine the optimum wavelengths and number of latent variables for PLS regression model. A linear function of R-squares for calibration and prediction sets was utilized as the objective function of the optimization problem. The optimum wavelengths selected by genetic algorithm helped to considerably improve the performance of the PLS regression model, not only for the calibration sets but also for the prediction sets. The optimum number of latent variables varied over a wide range, from the maximum allowed (20) to a lower limit of six. Representative wavelengths for each data set were also statistically determined and assigned to corresponding wood components through a band assignment process, which showed strong agreement.

Introduction

Efficient utilization of forest resources requires information on wood property variation at multiple scales. Information is often limited owing to the cost and time associated with measuring many wood properties and various methodologies have been developed for estimating these properties rapidly (Schimleck et al. 2019). Near infrared (NIR) spectroscopy is one such technique that has been widely applied to wood (Tsuchikawa and Kobori 2015; Schimleck and Tsuchikawa 2020), and it is the only nondestructive technique that can provide an estimate of pulp yield (the yield of chemically derived pulp from a given volume of wood). Pulp yield is critical to

✉ Tu X. Ho
tu.ho@oregonstate.edu

[1] Department of Wood Science and Engineering, College of Forestry, Oregon State University, Corvallis, OR 97331, USA

the economics of the pulp and paper industry (Greaves and Borralho 1996) and is very expensive to measure (Meder et al. 2011). Hence, there is increased interest in utilizing a rapid, inexpensive approach for its determination (Michell 1995).

Owing to its importance and direct relationship with wood chemistry, the estimation of pulp yield by NIR spectroscopy began with the earliest wood—NIR papers (Birkett and Gambino 1988; Wright et al. 1990) and pulp yield has remained a consistent focus of NIR-wood related research (Downes et al. 2009, 2010, 2011; Meder et al. 2011; White et al. 2009). However, efforts to improve calibration performance through the utilization of advanced selection techniques are rare. For example, Mora and Schimleck (2008) utilized three different sample selection techniques (CADEX, DUPLEX and SELECT algorithms) to identify samples most representative of their data set for the development of pulp yield calibrations. They showed calibration performance was improved by utilizing only selected samples and recommended that these methods be employed to identify unique samples prior to doing any wood property determination utilizing models based on NIR spectra. More recently, Li et al. (2019) utilized a particle swarm optimization (PSO)—support vector machine (SVM) approach and observed improved density prediction for four commercially important Chinese species.

The selection of the most representative wavelengths in the spectra data of all samples might improve both calibration and prediction performances of partial least squares (PLS) regression and reduce computational workload. This selection problem can be defined as an optimization problem (Bangalore et al. 1996).

Many complex real-world problems involve optimizing goals, which means searching for the maximum and / or minimum values of these goals (objective values). For example, in manufacturing, maximizing profit and minimizing cost are common aims; whereas, in logistics, goods or services management, distribution and transportation are of interest, such that goods or services can be delivered in the shortest time and in a cost effective manner. In these examples, profit, cost and time are objective values. Objective values are affected by many factors, and these are called design variables (or decision variables). If there are restrictions, which are typically expressed mathematically as inequalities or equations, they are called constraints. A function, which expresses relationships between objective values and design variables, is called the objective function.

The common optimization problems in the field of NIR spectroscopy/chemometrics include wavelength selection (i.e., variable selection) and selection of the appropriate number of components (i.e., latent variables) in partial least-square (PLS) regression. Others include preprocessing techniques, such as feature selection and optimization of the parameters in calibration models with Support Vector Machines (Ramirez-Morales et al. 2016); and instrumentation optimization (signal precision and wavelength resolution) (Greensill and Walsh 2000). These problems have been solved by many optimization methods: the binary dragonfly algorithm (Chen and Wang 2019), genetic algorithms (GA) (Bangalore et al. 1996; Villar et al. 2014; De et al. 2017), artificial bee colony (Sun et al. 2019), particle swarm optimization (De et al. 2017; Lou et al. 2014), ant colony optimization (Xiaowei et al. 2014) and simulated annealing (Swierenga et al. 1998; Balabin and Smirnov 2011) are some examples. These metaheuristic algorithms help to save time and computational resources, especially in the case of wavelength selection problems in which the solution space is too large. Among the mentioned algorithms, simulated annealing is a single solution approach to improve a local search heuristic to find a better solution; while the others are population-based approaches which maintain and improve multiple potential solutions by generating a new population based on principles of natural systems. Evolutionary algorithm (e.g. genetic algorithm) and swarm-intelligence-based algorithm (e.g. binary dragonfly algorithm, artificial bee colony, particle swarm optimization and ant colony optimization) are two common categories of population-based methods. It is worthy to note that in spite of the popularity of these optimization methods in the field of NIR spectroscopy/chemometrics, their applications in the field of wood-NIR are very limited.

Xiaobo et al. (2010) and Balabin and Smirnov (2011), reviewed variable selection methods for NIR spectroscopy, including GA. Xiaobo et al. (2010) concluded that GA combined with PLS regression showed superiority over other applied multivariate methods because wavelengths selected by GA did not lose prediction capacity and provided useful information about the chemical system.

Villar et al. (2014) applied three variable selection methods, including Martens Uncertainty Test, interval Partial Least Squares (iPLS) and GA to Visible-NIR spectra. The application of iPLS and GA resulted in considerable improvement of the calibration model with the number of latent variables being reduced while also decreasing the root mean square error of the cross-validation (RMSECV) and the standard error of cross-validation (SECV) and increasing the ratio of prediction to deviation (RPD) compared to a full spectrum model.

Evolutionary genetic algorithms are a branch of evolutionary computation, which are inspired by natural evolutionary and adaption processes. Evolutionary algorithms include three major algorithms, i.e., evolution strategies, evolutionary programming and genetic algorithms. Rechenberg (1973) introduced evolutionary strategies as a numerical optimization technique, while the current framework of genetic algorithms was first proposed by Holland (1975) and his students (Jong 1975). An important addition was the development and introduction of the population concept into evolution strategies by Schwefel (1981, 1995). Evolutionary algorithms have been adapted to various optimization problems, with examples including numerical optimization, for example, both constrained (Michalewicz and Schoenauer 1996; Kim and Myung 1997) and unconstrained (Yao and Liu 1996, 1997) and multi-objective optimization (Fonseca and Fleming 1995, 1998).

All evolutionary algorithms have two prominent features, which distinguish themselves from other search algorithms. First, they are all population-based and second, there is communication and information exchange among individuals in a population. They are the result of selection and/or recombination in evolutionary algorithms. Most recombination (crossover) operators use two parents and produce two offspring which inherit the information (genes) from their parents.

Genetic algorithms have been applied in the area of NIR spectroscopy since the 1980s. Koljonen et al. (2008) reviewed applications of GAs, including wavelength selection, wavelength interval selection, feature selection, co-optimization for wavelength selection and the number of PLS components, pre-processing optimization and wavelet transformation. The authors also proposed some potential research directions and applications of GAs in chemometrics.

In this paper, the GA approach was applied to a variable selection problem, which can be considered as an optimization problem, for NIR spectroscopy data sets. Two data sets represented by untreated, and second derivative spectra were used to predict pulp yield. The goals of the optimization problem were reducing the number of variables (i.e., wavelengths) for PLS regression and identifying the most frequent optimum wavelengths (i.e., representative wavelengths) for each data set. NIR band assignment was utilized to provide useful information about the wood components related to the optimum wavelengths.

Materials and methods

Optimization problem

Wavelength selection, number of wavelengths (N_{WvL}) and number of latent variables (N_{comp}) for PLS regression are often optimized in the same procedure using GA. Using an approach first implemented by Bangalore et al. (1996), a chromosome includes a series of ($N+1$) genes, in which N is the total number of wavelengths in the wavelength domain. Therefore, each gene in the first N genes corresponds to a specific wavelength. The value of a gene is binary, which indicates whether the wavelength is included in the model or not (i.e., 1 = yes and 0 = no). The number of wavelengths for the regression model (N_{WvL}) is counted as the number of genes among the first N genes assigned the value of 1. However, as a result, the number of selected wavelengths could not be controlled. The last gene represents the number of latent variables, which is an integer. By developing the problem in this way, the wavelengths, N_{WvL} and N_{comp} are co-optimized.

In the study presented here, the optimum wavelengths and number of latent variables for PLS regression are investigated at a specific number of wavelengths, which increased from 10 to 100. This approach allows the observation of how these variables and PLS model metrics change versus the number of wavelengths. Therefore, the implementation of GA to the optimization problem will be different from the aforementioned studies (Bangalore et al. 1996; Koljonen et al. 2008) and summarized as follows.

Each calibration model for PLS regression includes ($N_{WvL}+1$) variables, which are a combination of wavelengths selected from the wavelength domain and the number of latent variables for PLS regression. They are combined into a vector, called a chromosome (or an individual) $x^* = \left[x_1 x_2 \dots x_{NWvL+1}\right]^T$. Each value in a chromosome is called a gene. The first gene x_1 represents the number of latent variables while the others (from x_2 to x_{NWvL+1}) are assigned integer values which belong to a N_{WvL}-combination without repetition of all wavelength values in their domain (N elements). This combination is sorted in ascending order before being assigned to genes.

Data sets

The optimization problem was developed and applied to two NIR data sets selected as they represented two extremes in terms of pulp yield variation. The first (pulp yield-min) was comprised of 67 clonal blue gum (*Eucalyptus globulus*) samples (Schimleck and French 2002) all the same age and with Kraft pulp yields that ranged from 50.8 to 55.8%. The second (pulp yield-max) included 30 blue gum samples (Michell 1995) from several different native forests in Tasmania, Australia. The forests were of various ages and pulped samples had a much wider yield range (soda pulp yields = 37.6 to 60.2%). Details regarding sample preparation and collection of NIR spectra are described in Michell (1995) and Schimleck and French (2002). Briefly, wood chip samples (representative of individual trees or clones) were milled in a model 4 Wiley mill (Thomas Scientific, Swedesboro, NJ, USA). For both data sets, milled wood was placed in a large NIR systems sample cup (NR-7070) and duplicate spectra (the cell was repacked between scans) collected using a NIR Systems Inc. Model 5000 scanning spectrophotometer (Silver Spring, Maryland, USA). Duplicate spectra (wavelength range 1100–2500 nm in 2 nm increments, total $N = 700$) were averaged prior to analysis. For the pulp yield-max samples, a static sample holder was used, whereas a spinning sample holder was utilized for the collection of spectra from the pulp yield-min samples. Schimleck and French (2002) and Turner et al. (1983) provide information regarding the determination of pulp yield for samples included in the two datasets.

Each data set was separated into two subsets (i.e., calibration set and prediction set) based on the DUPLEX selection method (Snee 1977), which use Euclidean distance to determine the proximity of samples to others in a factor space (Mora and Schimleck 2008). The basic information of data sets is shown in Table 1.

The maximum number of latent variables was selected to be 20 for pulp yield-min data and 15 for pulp yield-max data. The value of 20 for the maximum N_{comp} was considered more than necessary for a PLS model using this set but we wanted to allow for instances where the optimization required more latent variables as suggested by preliminary models using 10 latent variables. Based on an analysis of the percentage of variance explained of Y for full data set, $N_{\mathrm{comp}} = 20$ explained 99.54% variance of Y. Therefore, a number of latent variables greater than 20 does little in terms of improving the PLS model and might actually make the model more complicated and therefore increase the computing time. In case of the pulp yield-max data set, the maximum number of latent variables was limited by the size of the calibration set (20 samples). When the cross-validation sets of 4 were used, the number of latent variables should not be larger than the size of the training set (i.e., 15). Again,

Table 1 Data set information

Data set	Pulp yield-min			Pulp yield-max		
	Total	Calibration set	Prediction set	Total	Calibration set	Prediction set
Number of samples	67	51	16	30	20	10

the analysis of the percentage of variance explained of Y showed that $N_{\text{comp}}=15$ explained 99.94% variance of Y.

Therefore, the domains of optimum variables were defined as:

For number of latent variables:

$D[x_1] = [1\ 2\ 3\ \dots\ 18\ 19\ 20]$ for pulp yield-min data
$D[x_1] = [1\ 2\ 3\ \dots\ 13\ 14\ 15]$ for pulp yield-max data

For wavelength variables:

$D[x_i] = [1100\ 1102\ 1104\ \dots\ 2494\ 2496\ 2498]$ (nm)
$(i = 2 \dots N_{WvL}+1)$

The performance of a calibration model for PLS regression (i.e., a chromosome or an individual) in this study was evaluated by four inequality constraints ($m=4$) and two objective values. The constraint conditions are R-squares for the calibration and prediction sets (R_c^2 and R_p^2, respectively) and the standard errors for the calibration and prediction sets (SEC and SEP, respectively). These constraint conditions can be expressed as follows:

$$\begin{cases} R_c^2 - R_{c,\min}^2 \geq 0 \\ R_p^2 - R_{p,\min}^2 \geq 0 \\ \text{SEC}_{max} - SEC \geq 0 \\ \text{SEP}_{max} - SEP \geq 0 \end{cases}$$

in which $R_{c,\min}^2$, $R_{p,\min}^2$, $\text{SEC}_{\max}$ and $\text{SEP}_{\max}$ are limited values for R_c^2, R_p^2, SEC and SEP, respectively. These limited values were selected so that the performance of an optimum calibration model is equal to, or better than that, of a calibration model using all wavelengths in the data set and $N_{\text{comp}}=6$. The details of the constraint values are shown in Table 2.

The objective values in this study are the aforementioned R-squares for the calibration and prediction sets. It demonstrates that the overall goal of the optimization

Table 2 Constraint values for data sets and the result of optimized values

Data set		Pulp yield-min		Pulp yield-max	
Treatment		Untreated	2nd Derivative	Untreated	2nd Derivative
$R_{c,\min}^2$	Full set	0.75	0.86	0.97	0.98
$R_{c,\text{opt}}^2$	Optimized set	0.840–0.995	0.917–0.999	0.978–0.998	0.992–1.000
$R_{p,\min}^2$	Full set	0.87	0.63	0.94	0.95
$R_{p,\text{opt}}^2$	Optimized set	0.902–0.999	0.971–0.999	0.962–0.990	0.990–1.000
$\text{SEC}_{\max}$	Full set	0.64	0.48	1.28	0.93
SEC_{opt}	Optimized set	0.105–0.513	0.045–0.395	0.504–1.226	0.016–0.859
$\text{SEP}_{\max}$	Full set	0.42	0.68	1.39	1.48
SEP_{opt}	Optimized set	0.034–0.367	0.031–0.190	0.575–1.106	0.156–0.727

Note: $R_{c,\min}^2$, $R_{p,\min}^2$, $\text{SEC}_{\max}$ and $\text{SEP}_{\max}$ are constraint values for the optimization problem. They are determined from a calibration model using all wavelengths in the data set and $N_{\text{comp}}=6$

$R_{c,\text{opt}}^2$, $R_{p,\text{opt}}^2$, SEC_{opt} and SEP_{opt} are the results from optimized wavelength sets problem presented here is to obtain a set of wavelengths which can produce good R-square values for both sets. If the objective value is only the R-square for the calibration set, it might result in an overfitting problem for PLS regression and as a result, the fitness of PLS regression in terms of prediction would be reduced. The objective function is defined as: $f_{obj} = \alpha \times R_c^2 + \beta \times R_p^2$, in which α and β are weighted factors for R_c^2 and R_p^2, respectively ($\alpha + \beta = 1$). In this study, α and β were selected to be 0.5.

Optimization process

In the optimization problem for PLS regression, the first generation is created randomly. The first generation of parents, P, is represented by the following matrix, in which k is the number of parents (or the population size), and each row represents an individual's chromosome. In this study, the number of parents (k) is 100.

$$P = \begin{bmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,n-1} & x_{1,NWvL+1} \\ x_{2,1} & x_{2,2} & \cdots & x_{2,n-1} & x_{2,NWvL+1} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{k,1} & x_{k,2} & \cdots & x_{k,n-1} & x_{k,NWvL+1} \end{bmatrix}$$

In the initial step, a hundred individuals were created by randomly selecting gene values from pre-defined variable domains by the uniform distribution. The strength (or fitness) of each individual is evaluated by the objective function f_{obj}. Good individuals are selected to be parents based on their fitness to create the next generation (offspring) during the search process. After that, the objective function of each individual offspring is evaluated and compared to their parents using a penalty function (Van de Lindt and Dao 2007) in a process named tournament selection. The best individuals are identified and become new parents of the next generation. The process is repeated until pre-determined convergence criteria are satisfied.

The searching process is performed through the crossover (recombination) and mutation operators. In the crossover operator, two or more offspring are often produced by randomly exchanging genes from two or more parents. In most cases, two parents will be selected randomly, thus, only two offspring will be created and inherit genes from parents. The number of individuals selected to perform the crossover operator depends on a crossover rate. A crossover point, where genes exchange occurs, is chosen randomly between 1 and (n-1). There are possibly more than one crossover points. However, only one crossover point will be used in this study. For example, individuals X_i and X_j are selected to take the crossover operator at the crossover point k, the two offspring are expressed as:

$$X_i' = \begin{bmatrix} x_{i,1} & \cdots & x_{i,k-1} & x_{j,k} & \cdots & x_{j,NWvL+1} \end{bmatrix}$$

$$\text{and} \quad X_j' = \begin{bmatrix} x_{j,1} & \cdots & x_{j,k-1} & x_{j,k} & \cdots & x_{i,NWvL+1} \end{bmatrix}$$

The mutation operator changes some genes in some individuals in every generation. Similar to the crossover operator, the number of chromosomes selected to be mutated depends on the pre-defined mutation rate while the mutation points are chosen randomly between 1 and n. At a mutation point q on a selected chromosome p, a gene's value is changed to a random value which is within the gene's domain. A new offspring is expressed as:

$$X_p' = \left[x_{p,1} \; \cdots \; x_{p,q-1} \; x_{p,q}' \; x_{p,q+1} \; \cdots \; x_{p,NWvL+1} \right]$$

Since the genes from x_2 to x_{NWvL+1} are required to create a set of unique wavelengths, the new offspring produced by crossover and mutation operators are checked. If the values of wavelength genes are not unique, the operator is repeated until a set of unique wavelength values is obtained. In addition, the large domains defined for variables result in numerous possible individuals. Therefore, in this study, the crossover and mutation rates were chosen to be 0.5 to introduce various new genes to the population.

The offspring matrix, O, obtained from crossover and mutation operators, is expressed as:

$$O = \begin{bmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,n-1} & x_{1,n} \\ x_{2,1} & x_{2,2} & \cdots & x_{2,n-1} & x_{2,n} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{r,1} & x_{r,2} & \cdots & x_{r,n-1} & x_{r,n} \end{bmatrix}$$

The selection process is conducted for parents and offspring using the tournament selection method. Each individual is a PLS regression model for the respective data set. The regression produces constraint values, and the fitness of each individual is evaluated based on the objective function. The fitness vector **Y** and the constraint value matrix **C** for $(k+r)$ individuals can be expressed as:

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ \cdots \\ y_{(k+r)} \end{bmatrix} \text{and}\, C = \begin{bmatrix} c_{1,1} & c_{1,2} & \cdots & c_{1,m} \\ c_{2,1} & c_{2,2} & \cdots & c_{2,m} \\ \cdots & \cdots & \cdots & \cdots \\ c_{(k+r),1} & c_{(k+r),2} & \cdots & c_{(k+r),m} \end{bmatrix}$$

where m is the number of constraint values being considered (as described earlier, $m=4$).

As proposed by Van de Lindt and Dao (2007), one should concentrate on searching for individuals around those individuals having the best fitness, so that the approach to global optimization is as stable as possible. In that scenario, some individuals, which do not satisfy the constraint conditions but have very good fitness, might be considered for retention. A penalty function was proposed by Van de Lindt and Dao (2007) so the individuals having fitness values around the best fitness value will have a higher probability of survival. The mathematical form of this penalty function for a minimum optimization problem can be expressed as:

$$f_p(x) = \begin{cases} f(x) & \textit{if } \left[g_k(x) \geq 0 \text{ and } h_m(x) = 0\right] \text{ or } \left[f_b(x) < f(x)\right] \\ f_b(x) + \left[f_b(x) - f(x)\right] & \textit{if } \left[g_k(x) < 0 \text{ or } h_m(x) \neq 0\right] \text{ and } \left[f_b(x) \geq f(x)\right] \end{cases}$$

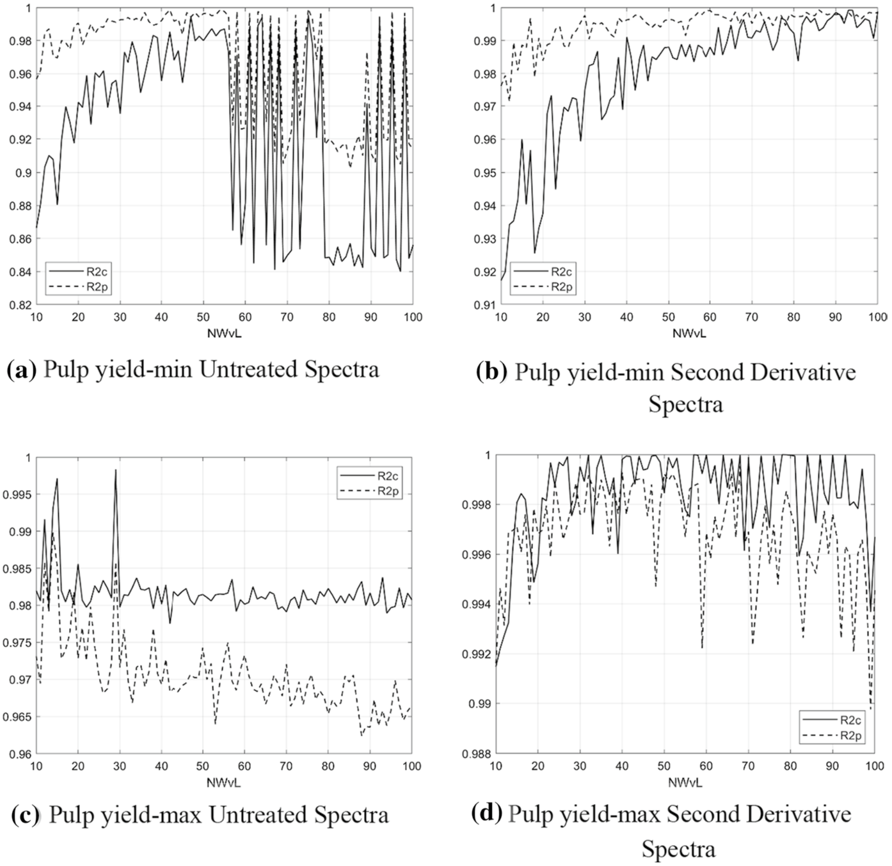
$$for\, k = 1,2, \ldots , N; h = 1,2, .., M$$

where $f_p(x)$ is fitness after penalizing; $f(x)$ is fitness before penalizing, and $f_b(x)$ is fitness of the best individual in the constraint domain $\left[g_k(x) \geq 0\,\text{and}\,h_m(x) = 0\right]$, in which $g_k(x)$ and $h_m(x)$ are constraint functions for N inequality constraints and M equality constraints, respectively.

Results and discussion

Optimization results

As mentioned, optimization was implemented for a specific number of wavelengths. There were 91 optimization cases corresponding to the change in number of wavelengths from 10 to 100. Figure 1 shows objective values (R_c^2 andR_p^2) from



(a) Pulp yield-min Untreated Spectra

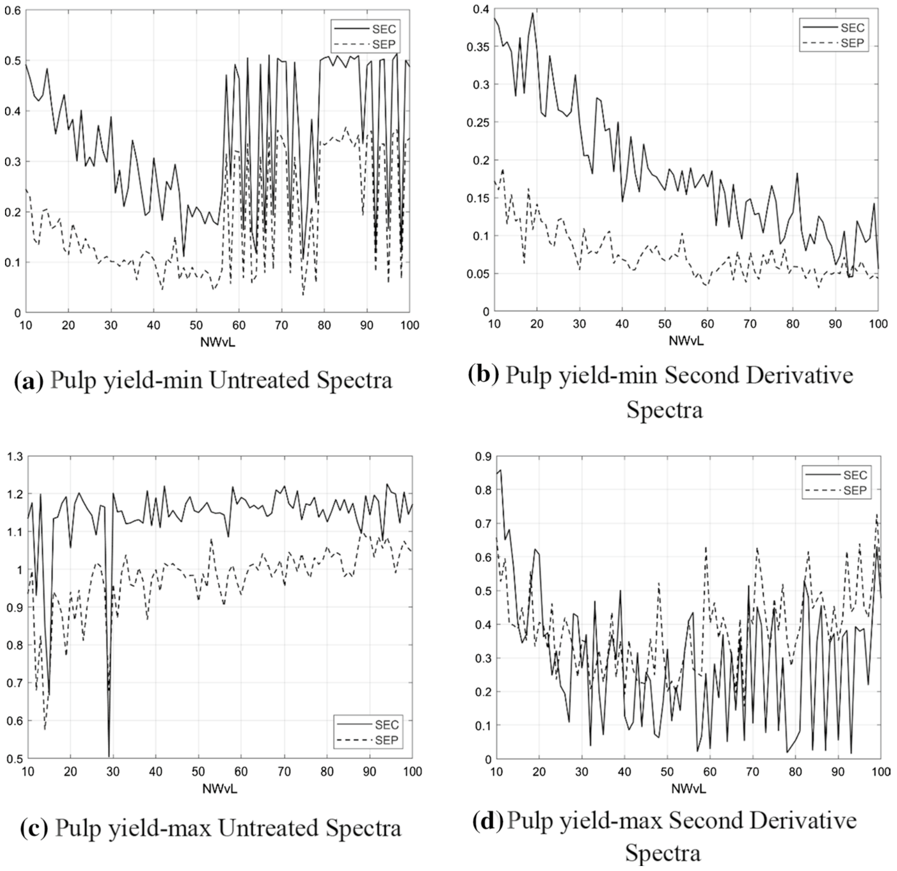
(b) Pulp yield-min Second Derivative Spectra

(c) Pulp yield-max Untreated Spectra

(d) Pulp yield-max Second Derivative Spectra

Fig. 1 Optimum objective values result for different spectra data sets

each optimum set of wavelengths resulting from the optimization process for the four data sets. Overall, the objective values were greatly improved compared to the corresponding values obtained from using all wavelengths (see Table 2). The best number of wavelengths for optimizing the prediction result (R_p^2) differed among data sets. For pulp yield-min untreated spectra, R_p^2 increases from 0.96 to 0.98 when N_{WvL} increases from 10 to 22. R_p^2 stays above 0.98 before it fluctuates drastically in the range of 0.9-1when N_{WvL} is larger than 56. R_p^2 of pulp yield-max untreated spectra reaches a peak value of 0.99 at $N_{WvL} = 14$ which is then followed by a downward trend to around 0.965 as N_{WvL} increased. Excellent values of R_p^2 were observed for both the second derivative sets. R_p^2 of pulp yield-min second derivative spectra increases from 0.97 and remains above 0.99 with $N_{WvL} \geq 27$, while R_p^2 of pulp yield-max second derivative spectra is higher than 0.99 for all investigated cases of N_{WvL}.The optimization also reduces *SEC* and *SEP* values indicating an improvement in model fitting and predictive performance (Fig. 2).



(a) Pulp yield-min Untreated Spectra

(b) Pulp yield-min Second Derivative Spectra

(c) Pulp yield-max Untreated Spectra

(d) Pulp yield-max Second Derivative Spectra

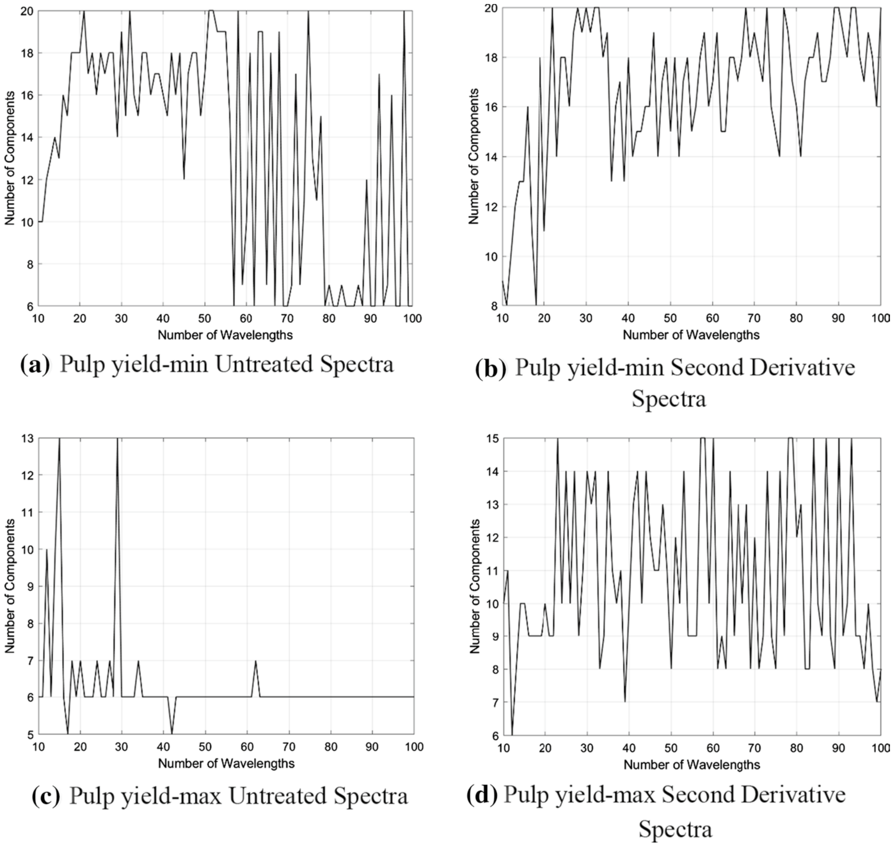
Fig. 2 SEC and SEP from optimum results for different spectra data sets

Figure 3 shows the optimization results for the number of latent variables (N_{comp}). These are values, which combined with the corresponding optimum wavelength sets, resulted in the highest objective values. Only pulp yield-max untreated spectral data shows a convergence of $N_{\text{comp}}=6$ versus number of wavelength (N_{WvL}). For the other data sets, the optimum N_{comp} fluctuates over a wide range. However, $N_{\text{comp}}=6$ tends to be the lower limit while the upper limit reaches the preselected maximum number of latent variables in some cases. This suggests that the true upper limit might go higher if the maximum number of latent variables were increased.

For optimization based on different numbers of wavelengths, the sets of identified wavelengths share few common components. For example, the result from optimization for pulp yield-min untreated spectra shows that the optimum wavelength sets for $N_{WvL}=10$ and $N_{WvL}=11$ are as follows:

$$\text{WvL}_{NWvL=10} = [1470\ 1636\ 1706\ 1790\ 1852\ 1854\ 2286\ 2364\ 2474\ 2476]\,(\text{nm})$$

$$\text{WvL}_{NWvL=11} = [1152\ 1154\ 1198\ 1200\ 1472\ 1918\ 2032\ 2322\ 2328\ 2364\ 2372](\text{nm})$$



(a) Pulp yield-min Untreated Spectra

(b) Pulp yield-min Second Derivative Spectra

(c) Pulp yield-max Untreated Spectra

(d) Pulp yield-max Second Derivative Spectra

Fig. 3 Optimization results for the number of latent variables

These two sets only share common wavelengths in the range 1470–1472 nm and 2364 nm. It suggests that the optimization result for a specific number of wavelengths might be just a local optimized point for that specific case. Therefore, the local optimized point contains not only the common wavelengths but its own distinguishing wavelengths. This means not all the optimized wavelengths in that case contribute to global optimization and help to explain, or understand, the relationship between wavelengths and wood components or wood properties.

Most frequently identified wavelengths

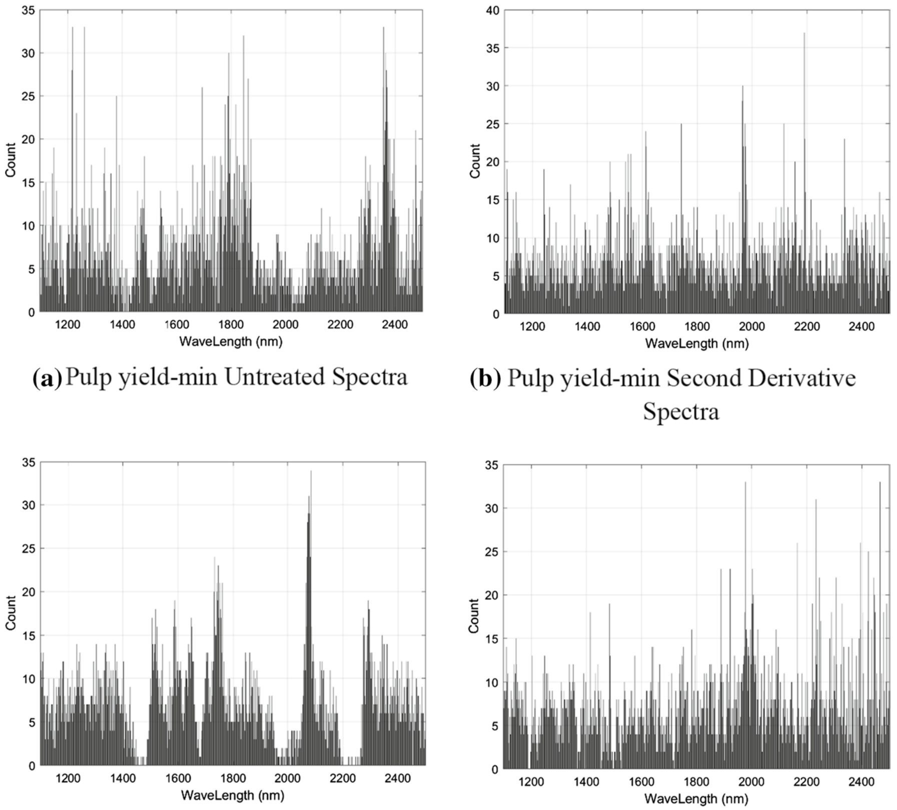
A statistical approach was applied to analyse the optimization results. Each wavelength in the domain (i.e., from 1100 to 2498 nm) was counted for its presence in the different optimum wavelength sets resulting from 91 optimization cases. The most frequent wavelengths of a data set were considered representative for that data set. The frequency of wavelengths across all optimization cases for a given data set is plotted in Fig. 4. Frequency distribution for the untreated spectral data sets is more concentrated than that for second derivative data sets (Fig. 4). Moreover, although the distributions are concentrated for untreated spectra, the wavelengths with highest frequency of pulp yield-min and -max untreated spectra are not the same indicating that representative wavelengths are different for the untreated spectra.

Different sets of the most frequent wavelengths were determined for each data set based on different minimum frequency values. For an example of pulp yield-min untreated spectra, there are 304 wavelengths presented at least seven (7) times and 12 wavelengths with a minimum frequency of 26. The objective values result of the models using representative wavelengths sets as their input are plotted in Fig. 5. In general, the representative wavelength sets also greatly improved the PLS model, although the performance was not as high as that provided by the optimized wavelength sets. Moreover, Fig. 5 shows that R_c^2 increases with the number of representative wavelengths (NRWvL). Moreover, R_p^2 peaks when NRWvL is in the range of 100–200 wavelengths and tends to decrease when more wavelengths are added to the model input.

Comparison of band assignments

Schwanninger et al. (2011) reviewed and provided a summary of band assignments for wood and its components. Results from that study were utilized here and matched to the most frequent wavelengths of each data set. Table 3 shows the representative wavelengths of each data set, their frequency, bands in the NIR spectrum identified as arising from wood, the related bond vibration and the corresponding wood components.

Strong agreement was observed between the most frequently observed representative wavelengths and bands corresponding to wood components. The strong agreement is very encouraging as it indicates that wavelengths identified as important for optimization originate from bond vibrations in wood components that directly influence pulp yield (Poke and Raymond 2006). For the pulp



(a) Pulp yield-min Untreated Spectra

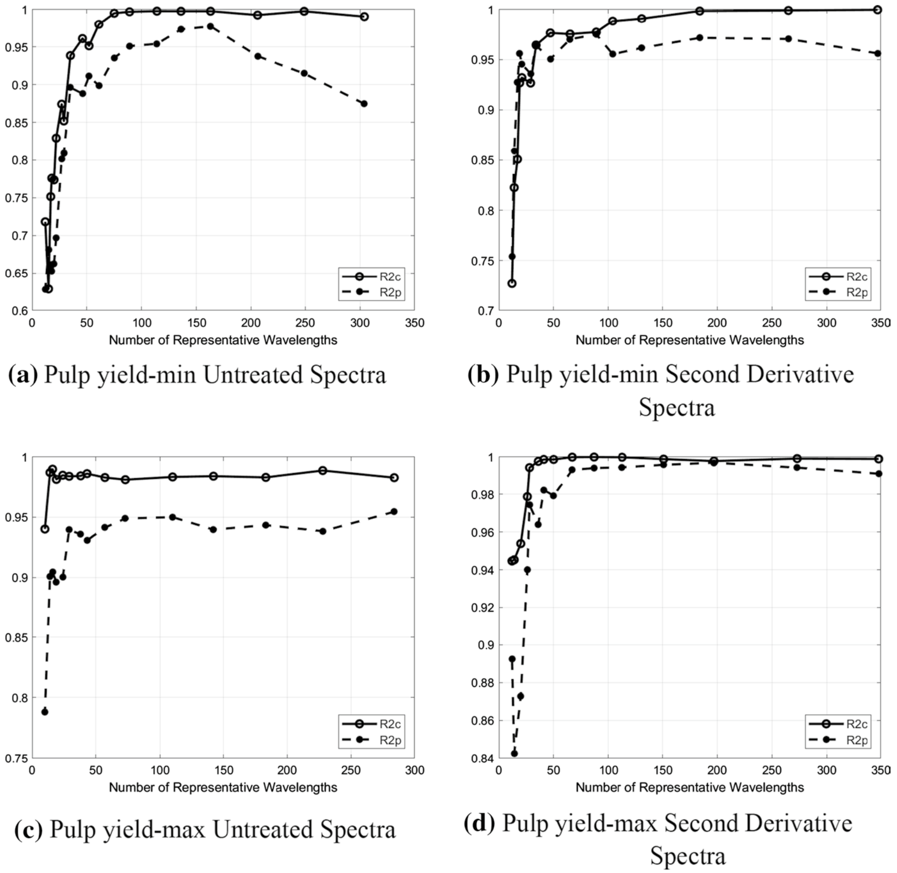
(b) Pulp yield-min Second Derivative Spectra

(c) Pulp yield-max Untreated Spectra

(d) Pulp yield-max Second Derivative Spectra

Fig. 4 Presence frequency of wavelengths in the optimum results

yield-max data, nearly all identified wavelengths that had a wood related analog arose from cellulose while for the pulp yield-min data the frequency of bands related to lignin, while still relatively small, was greater. The contrasting range in yields for the data sets influenced the selection of wavelengths. It is likely that the wide range of yields for the pulp yield-max data set has permitted clear identification of specific wavelengths related to cellulose utilizing untreated spectra (Fig. 4c), whereas for the pulp yield-min untreated spectra (Fig. 4a) the narrow yield range resulted in more wavelengths being identified as important and also allowed lignin-related wavelengths to have a greater influence. This suggests variation in lignin content is more important for pulp yield models based on data that has a narrow range. For the second derivative data, more wavelengths had influence which can be expected as this treatment baselines the data and highlights differences amongst wavelengths (Barton 1989).



(a) Pulp yield-min Untreated Spectra

(b) Pulp yield-min Second Derivative Spectra

(c) Pulp yield-max Untreated Spectra

(d) Pulp yield-max Second Derivative Spectra

Fig. 5 Objective values results for different representative wavelength sets

Conclusion

This study presents an optimization problem for *Eucalyptus globulus* pulp yield models. Two NIR data sets represented by untreated and second derivative spectra were used in multivariate calibration models based on partial least squares (PLS) regression to predict pulp yield. The genetic algorithm was used to select optimum wavelengths, with an objective function including both *R*-squares for the calibration and prediction sets. The optimization process was run for 91 cases corresponding to the change in number of wavelengths from 10 to 100. Results show that optimum wavelengths considerably improved PLS regression model performance (represented by *R*-square and standard error), not only for the calibration sets but also the prediction sets. However, each spectral data set has its own optimum number of wavelengths. Despite differences, *R*-square values for prediction were still greater than 0.96. The optimum number of latent variables varied over a wide range from the maximum allowed (20) to a lower limit of six. A statistical approach was applied

Table 3 Band assignments for optimization results

Identified wavelength (nm)	Band location (nm)	Bond vibration	Wood component
a. Pulp yield-min untreated spectra (frequency > 20, bold text = frequency > 30)			
1216 (28), **1218 (33)**	1212–1225	2nd OT C–H str	Cellulose
1232 (23)			
1262(33)			
1378 (25)			
1694 (26)	1685 1698	1st OT C–Har str 1st OT C–H str	Lignin Lignin
1778(24), 1788 (25), 1790 (25), **1792 (30)**, 1796 (20)	1780 1788, 1790 1791 1793	1st OT C–H2 str 1st OT C–H str 1st OT C–H str 1st OT C–H str	Cellulose Cellulose Lignin Cellulpose
1818 (24)	1820	O–H str. + 2nd OT C-O str	Cellulose
1846 (32), 1862 (27), 1872 (20)			
2356 (26), **2358 (33),** 2362 (21)**, 2364 (30),** 2366 (22), 2368 (28), 2370 (26), 2372 (22),	2352 2361	C–H str. + C–H def., 2nd OT C- H2 bend O–H def. or C–H def + C–H str. or C–H2 str	Cellulose
2376 (20), 2380 (20)	2384	Not assigned	Lignin
2396 (20)			
2474 (21)	2461	C–H str. + C–C str	Starch (cellulose)
b. Pulp yield-min second derivative spectra (frequency > 15, bold text = frequency > 25)			
1108 (19), 1110 (16)			
1130 (15), 1140 (16)	1143	2nd OT Car–H str., 2nd OT C–H str. of CH3 groups	Lignin
1242 (19)			
1338 (17)	1350	1st OT C–H str. + C–H def	Hemicellulose (tentative)
1482 (20), 1484 (16)	1477–1484	1st OT O–H str	Cellulose
1516 (15)	1510 1515	1st OT O–H str 1st OT O–H str	Cellulose Cellulose?
1538 (20), 1548 (21), 1550 (16), 1558 (21)	1534, 1550 1540 1545 1548	1st OT O–H str 1st OT O–H str 1st OT O–H str 1st OT O–H str	Cellulose Cellulose Cellulose Cellulose
1612 (24), 1614 (22), 1620 (15), 1622 (16)	1616 1632	1st OT C–H str 1st OT O–H str	Not assigned Cellulose
1742 (25)	1731	1st OT C–H str	Cellulose
1954 (16), **1964 (28), 1966 (30),** 1968 (22)			
1974 (25), 1976 (22), 1978 (17), 1982 (15)	1980	O–H str. + O–H def. of $H_2$0	Water
2116 (25)	2110	O–H def. + O–H str	Cellulose
2156 (20)	2170, 2178	Not assigned	Cellulose/hemicellulose

Table 3 (continued)

Identified wavelength (nm)	Band location (nm)	Bond vibration	Wood component
2190 (37), 2192 (23), 2194 (16)	2200	C–H str. + C = O str	Lignin
2214 (15)			
2336 (23)	2328–2332	C–H str. + C–H def	Hemicellulose
2464 (16)	2461	C–H str. + C–C str	Starch (cellulose)
c. Pulp yield-max untreated spectra (frequency > 15, bold text = frequency > 20)			
1506 (17)	1510	1st OT O–H str	Cellulose
1518 (18), 1524 (16)	1515	1st OT O–H str	Cellulose (?)
1584 (16), 1586 (18), 1588 (19), 1592 (16)	1580 1588	1st OT O–H str 1st OT O–H str	Cellulose Cellulose
1648 (17), 1650 (16)	1632 1666	1st OT O–H str 1st OT C–H str	Cellulose Hemicellulose
1730 (20), 1732 (24), 1734 (16), 1736 (15), 1738 (15), **1740 (20), 1742 (21),** 1744 (19), **1746 (23),** 1748 (17), **1750 (21)**	1731	1st OT C–H str	Cellulose
1754 (18), 1756 (17), **1760 (21)**	1765	1st OT C–H str	Not assigned to a wood component
2066 (21), 2068 (16), 2070 (24), **2072 (28), 2074 (29), 2076 (31), 2078 (31), 2080 (29), 2082 (24), 2084 (32)**, 2086 (16)	2080 2086 2092	O–H str. + C–H def O–H str. + 0-H and C–H def O–H str. + 0-H and C–H def	Cellulose Cellulose/hemicellulose Cellulose
2274 (17),2276 (15), 2278 (16)	2270 2271 2272 2277	O–H str. + C–O str. C-H2 str. + C-H2 def C-H str. + C-H def O–H str. + C–C str. and/ or C-H str. + C-H def	Cellulose Cellulose/hemicellulose Hemicellulose Cellulose
2288 (15),2290 (16), 2292 (19),2294 (18), 2296 (18)	2291	C–O str. + O–H str. or C–H2 bend + C-H2 str	Cellulose
2342 (15)	2343	C–H str. + C–H def. and/ or 2nd OT C-H def	Cellulose
d. Pulp yield-max Second Derivative Spectra (frequency > 15, bold text = frequency > 25)			
1144 (15)	1143	2nd OT C_{ar}–H str.,2nd OT C–H str. of CH_3 groups	Lignin
1414 (18)	1414 1410 1417	1st OT O–H str 1st OT O–H str 1st OT C–H str. + C-H bend	Water Lignin/extractives Lignin
1482, 1484 (19)	1480	1st OT O–H str	Cellulose

Table 3 (continued)

Identified wavelength (nm)	Band location (nm)	Bond vibration	Wood component
1782 (16)	1780	1st OT C–H$_2$ str	Cellulose
1888 (23)			
1922 (23)	1916–1942	O–H asym. str. + O–H def. of H$_2$0	Water
1974 (18), **1978 (33)**, 1980 (16), 1982 (15)	1980	O–H str. + O–H def. of H$_2$0	Water
1996 (16), 2000 (19), 2002 (23), 2004 (19), 2006 (20)			
2022 (16)			
2088 (16)	2080 2086 2092	O–H str. + C–H def O–H str. + O–H andC–H def O–H str. + O–H and C–H def	Cellulose Cellulose/hemicellulose Cellulose
2166 (26)	2170, 2178	Not assigned	Cellulose/hemicellulose
2220 (19)			
2234 (31), 2238 (16), 2246 (22), 2250 (17)	2343	C–H str. + C–H def. and/or 2nd OT C-H def	Cellulose
2284 (15), 2290 (18)	2291	C–O str. + O–H str. or C– H$_2$ bend + C-H$_2$ str	Cellulose
2306 (22)			
2328 (19)	2328–2332	C–H str. + C–H def	Hemicellulose
2364 (15)	2361	O–H def. or C–H def + C–H str. or C–H$_2$ str	Cellulose
2390 (15), 2394 (16), **2396 (26)**, 2404 (18)	2384	Not assigned	Lignin
2422 (17), **2424 (25)**			
2436 (16), 2444 (22), 2446 (20), 2448 (18)			
2466 (33),	2461	C–H str. + C–C str	Starch (cellulose)
2478 (18), 2490 (19)	2488 2491	C–H str. + C–C str. ? C–H str. + C–C str	Lignin Cellulose (not confirmed)

to determine representative wavelengths for each spectral data set. Representative wavelengths were assigned to corresponding wood components through a band assignment process, which showed strong agreement. The result also suggests variation in lignin content is more important for pulp yield models based on data having a narrow range.

Compliance with ethical standards

Conflicts of interest The authors declare no conflict of interest.

References

Balabin RM, Smirnov SV (2011) Variable selection in near-infrared spectroscopy: benchmarking of feature selection methods on biodiesel data. Anal Chim Acta 692:63–72

Barton FE (1989) Spectra. In: Near infrared reflectance spectroscopy (NIRS): Analysis of forage quality, Marten GC, Shenk JS, Barton FE (ed.). United States Department of Agriculture, Agriculture Handbook No. 643, Govt Pr, Washington DC, pp 30–31

Birkett MD, Gambino MJT (1988) Potential applications for near infrared spectroscopy in the pulping industry. Pap S Afr 5:102

Bangalore AS, Shaffer RE, Small GW, Arnold MA (1996) Genetic algorithm -based method for selecting wavelengths and model size for use with partial least-squares regression: application to near-infrared spectroscopy. Anal Chem 68(23):4200–4212

Chen Y, Wang Z (2019) Wavelength selection for NIR spectroscopy based on the binary dragonfly algorithm. Molecules 24(3):421

De A, Chanda S, Tudu B, Bandyopadhyay RB, Hazarika AK, Sabhapondit S, Baruah BD, Tamuly P, Bhattachryya N (2017) Wavelength Selection for Prediction of Polyphenol Content in Inward Tea Leaves Using NIR. IEEE 7th International Advance Computing Conference (IACC), Hyderabad, 2017 pp. 184-187. doi: https://doi.org/10.1109/IACC.2017.0050

Downes GM, Meder R, Hicks C, Ebdon N (2009) Developing and evaluating a multisite and multispecies NIR calibration for the prediction of Kraft pulp yield in eucalypts. South For 71(2):155–164

Downes GM, Meder R, Ebdon N, Bond H, Evans R, Joyce K, Southerton S (2010) Radial variation in cellulose content and Kraft pulp yield in *Eucalyptus nitens* using near-infrared spectral analysis of air-dry wood surfaces. J Near Infrared Spectrosc 18(2):147–155

Downes GM, Meder R, Bond H, Ebdon N, Hicks C, Harwood C (2011) Measurement of cellulose content, Kraft pulp yield and basic density in eucalypt woodmeal using multisite and multispecies near infra-red spectroscopic calibrations. South For 73(3–4):181–186

Fonseca CM, Fleming PJ (1995) An overview of evolutionary algorithms in multi-objective optimization. Evol Comput 3(1):1–16

Fonseca CM, Fleming PJ (1998) Multi-objective optimization and multiple constraint handling with evolutionary algorithms-part i: a unified formulation. IEEE Trans Syst Man Cybern A Syst Humans 28(1):26–37

Greaves BL, Borralho NMG (1996) The influence of basic density and pulp yield on the cost of eucalypt kraft pulping: a theoretical model for tree breeding. Appita J 49(2):90–95

Greensill CV, Walsh KB (2000) Optimization of instrumentation precision and wavelength resolution for the performance of NIR calibrations of sucrose in a water-cellulose matrix. Appl Spectrosc 54(3):426–430

Holland JH (1975) Adaptation in natural and artificial systems. The university of Michigan press, Ann Arbor, MI

Jong KAD (1975) An analysis of the behavior of a class of genetic adaptive systems. PhD thesis, University of Michigan, Ann Arbor, MI

Kim J-H, Myung H (1997) Evolutionary programming techniques for constrained optimization problems. IEEE Trans Evol Comput 1(2):129–140

Koljonen J, Nordling TEM, Alander JT (2008) A review of genetic algorithms in near-infrared spectroscopy and chemometrics: past and future. J Near Infrared Spectrosc 16:189–197

Li Y, Via BK, Young T, Li Y (2019) Visible-near infrared spectroscopy and chemometric methods for wood density prediction and origin/species identification. Forests 10:1078

Lou W, Yang K, Zhu M, Wu Y (2014) Liu X and Jin Y (2014) Application of particle swarm optimization-based least square support vector machine in quantitative analysis of extraction solution of yangxinshi tablet using near infrared spectroscopy. J Innovat Opt Health Sci 7(6):1450011. https://doi.org/10.1142/S1793545814500114

Meder R, Brawner JT, Downes GM, Ebdon N (2011) Towards the in-forest assessment of Kraft pulp yield: comparing the performance of laboratory and hand-held instruments and their value in screening breeding trials. J Near Infrared Spectrosc 19(5):421–429

Michalewicz Z, Schoenauer M (1996) Evolutionary algorithms for constrained parameter optimization problems. Evol Comput 4(1):1–32

Michell AJ (1995) Pulpwood quality estimation by near-infrared spectroscopic measurements on eucalypt woods. Appita J 48(6):425–428

Mora C, Schimleck LR (2008) On the selection of samples for multivariate regression analysis: application to near infrared (NIR) calibration models. Can J For Res 38(10):2626–2634

Poke FS, Raymond CA (2006) Predicting extractives, lignin and cellulose contents using near infrared spectroscopy on solid wood in *Eucalyptus globulus*. J Wood Chem Technol 26(2):187–199

Ramirez-Morales I, Rivero D, Fernandez-Blanco E, Pazos A (2016) Optimization of NIR calibration models for multiple processes in the sugar industry. Chemometr Intell Lab Syst 159:45–57

Rechenberg I (1973) Evolution strategy: Optimization of technical systems according to the principles of biological evolution. Frommann Holzboog, Stuttgart, Germany

Schimleck L, Apiolaza L, Dahlen J, Downes G, Emms G, Evans R, Moore J, Pâques L, Van den Bulcke J, Wang X (2019) Non-destructive evaluation techniques and what they tell us about wood property variation. Forests 10:728

Schimleck LR, Tsuchikawa S (2020) Application of NIR spectroscopy to wood and wood derived products (Chapter 37). In: The Handbook of Near-Infrared Analysis, Fourth Edition, Newly revised and expanded. Ciurczak E, Igne B, Workman J, Burns D (ed). CRC Press, ISBN 9781138576483

Schimleck LR, French J (2002) Application of NIR spectroscopy to clonal *Eucalyptus globulus* samples covering a narrow range of pulp yield. Appita J 55(2):149–154

Schwefel H-P (1981) Numerical optimization of computer models. John Wiley & Sons, Chichester, England

Schwefel H-P (1995) Evolution and Optimum Seeking. John Wiley & Sons, New York

Schwanninger M, Rodrigues JC, Fackler K (2011) A review of ban assignments in near infrared spectra of wood and wood components. J Near Infrared Spectrosc 19:287–308

Snee R (1977) Validation of regression models: methods and examples. Technometrics 19:415–428. https://doi.org/10.2307/1267881

Sun H, Zhang S, Chen C, Li C, Xing S, Liu J, Xue J (2019) Detection of the soluble solid contents from fresh jujubes during different maturation periods using NIR hyperspectral imaging and an artificial bee colony. J Anal Meth Chem 2019:5032950. https://doi.org/10.1155/2019/5032950

Swierenga H, de Groot PJ, de Weijer AP, Buydens LMC DMWJ (1998) Improvement of PLS model transferability by robust wavelength selection. Chemometr Intell Lab Syst 41:237–248

Tsuchikawa S, Kobori H (2015) A review of recent application of near infrared spectroscopy to wood science and technology. J Wood Sci 61(3):213–220

Turner CH, Balodis V, Dean GH (1983) Variability in pulping quality of *E. globulus* from Tasmanian provenances. Appita J 36:371–376

Xiaobo Z, Jiewen Z, Povey MJW, Holmes M, Hanpin M (2010) Variables selection methods in near-infrared spectroscopy. Anal Chim Acta 667:14–32

Xiaowei H, Xiaobo Z, Jiewen Z, Jiyong S, Xiaolei Z, Holmes M (2014) Measurement of total anthocyanins content in flowering tea using near infrared spectroscopy combined with ant colony optimization models. Food Chem 164:536–543

Van de Lindt JW, Dao TN (2007) Evolutionary algorithm for performance-based shear wall placement in buildings subjected to multiple load types. J Struct Eng 133(8):1156–1167

Villar A, Fernandez S, Gorritxategi E, Ciria JI, Fernandez LA (2014) Optimization of the multivariate calibration of a Vis-NIR sensor for the on-line monitoring of marine diesel engine lubricating oil by variable selection methods. Chemometr Intell Lab Syst 130:68–75

White DE, Courchene C, McDonough T, Schimleck L, Jones D, Peter G, Purnell R, Goyal G (2009) Effects of specific gravity and wood chemical content on the pulp yield of loblolly pine. Tappi J 8(4):29–34

Wright JA, Birkett MD, Gambino MJT (1990) Prediction of pulp yield and cellulose content from wood samples using near infrared reflectance spectroscopy. Tappi J 73(8):164–166

Yao X, Liu Y (1996) Fast evolutionary programming. In: Evolutionary programming V: Proc. of the fifth annual conference on evolutionary programming. Fogel L J, Angeline P J, Back T (ed) pp 451–460. The MIT Press, Cambridge, MA

Yao X, Liu Y (1997) Fast evolution strategies. International conference on evolutionary programming, evolutionary programming VI, pp.149–161, 1997, Vol. 1213 (ISBN: 978–3–540–62788–3), https://doi.org/https://doi.org/10.1007/BFb0014808

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.