Wood Science and Technology (2022) 56:1835–1850
https://doi.org/10.1007/s00226-022-01431-9

ORIGINAL

On the selection of the weighting parameter value in optimizing *Eucalyptus globulus* pulp yield models based on NIR spectra

Yi Zhen[1] · Tu X. Ho[2] · Lindon Roberts[3] · Laurence R. Schimleck[2] · Arijit Sinha[2]

Received: 22 July 2022 / Accepted: 3 November 2022 / Published online: 15 November 2022

Abstract

Prediction of pulp yield of *Eucalyptus globulus* wood samples based on partial least squares (PLS) regression can be optimized by utilizing specific near infrared (NIR) wavelengths. A critical feature of this approach is the weighting of constraint conditions. Equal weighting balances optimization in terms of calibration and prediction; however, there is a lack of knowledge regarding prediction performance of wood property models when different weight factors are used. In this study, pulp yield models were developed using two *E. globulus* data sets characterized by narrow (5%) and extreme (22.6%) yield ranges and represented by untreated and second derivative NIR spectra. The global optimization solver pySOT was used to optimize the performance of a PLS regression model in terms of wavelengths selected and number of latent variables. A linear function of R-squares for calibration (R_c^2) and prediction (R_p^2) sets was utilized as the objective function with the aim of maximizing $\alpha R_c^2 + (1-\alpha)R_p^2$ for all values of α between 0 (maximizing R_p^2 without concern for R_c^2) and 1 (only maximizing R_c^2). Values of $\alpha \leq 0.8$ provided good predictive performance, whereas $\alpha \geq 0.9$ tended to overfit the calibration data indicating that models are robust for values of α from 0 to 0.8. Representative wavelengths for each data set were identified and assigned to corresponding wood components through a band assignment process. Strong agreement was observed for $\alpha \leq 0.8$; however, for $\alpha \geq 0.9$, identified wavelengths generally occurred in regions unrelated to vibrations arising from specific wood components.

✉ Tu X. Ho
tu.ho@oregonstate.edu

[1] School of Mathematics and Statistics, University of Melbourne, Parkville, VIC 3010, Australia

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

[3] Mathematical Sciences Institute, Australian National University, Canberra, ACT 2601, Australia

Introduction

Near infrared (NIR) spectroscopy has been widely utilized for analyzing wood (Tsuchikawa and Kobori 2015; Schimleck and Tsuchikawa 2021). Among existing nondestructive methods of wood analysis (e.g., SilviScan, acoustics, resistograph), NIR spectroscopy is unique in that it can be used to estimate properties related to wood chemistry (Schimleck et al. 2019). From an industrial perspective, the most important of these properties is pulp yield (Greaves and Borralho 1996), which is defined (in percent) as the yield of pulp from a given volume of wood (Raymond et al. 2001). Traditional assessment of pulp yield, which involves cooking wood chips to a fixed Kappa number (lignin content) in a laboratory digestor, is time-consuming, costly, and limits the number of samples that can be examined (Raymond et al. 2001). NIR diffuse reflectance spectroscopy provides a viable alternative for estimation of pulp yield (Downes et al. 2011; Trung et al. 2015), as the variation at specific wavelengths in the range 1100–2500 nm is directly related to variation in pulp yield (Michell and Schimleck 1996). Indeed, it has been successfully applied in tree breeding programs to improve pulp yield of short rotation plantations that are established on a large-scale to provide wood for pulping facilities worldwide (Schimleck 2008).

NIR spectroscopic estimation of pulp yield involves two main components, namely data collection and model development. Spectra of milled wood from a large number of samples with known pulp yield are first collected. These samples are expected to typically represent individual trees or composites of several trees (Downes et al. 2006) and should ideally be from a variety of sites (Downes et al. 2011). Then, a partial least squares (PLS) regression model is built based on the collected spectral and pulp yield data. The model relates the spectral information of each sample to its pulp yield and is used to predict the pulp yield of unexamined samples.

Increasingly sophisticated approaches are available for model development (Cogdill et al. 2004; Mora and Schimleck 2010; Fernandes et al. 2013; Li et al. 2019; Nasir et al. 2019; Ayanleye et al. 2021); however, exploration of their use for pulp yield estimation, or for estimating other wood properties, is rare in wood-related research. Recent papers (Ho et al. 2021, 2022) have investigated the potential to utilize genetic algorithms (GA) in model development (Bangalore et al. 1996; Villar et al. 2014; De et al. 2017). Ho et al. (2021) focus on optimization of pulp yield models for Tasmanian blue gum (*Eucalyptus globulus* Labill.) samples, and Ho et al. (2022) expand the investigation to several loblolly pine (*Pinus taeda* L.) wood properties (density, microfibril angle, modulus of elasticity and tracheid coarseness, radial diameter, tangential diameter, and wall thickness), which were measured by SilviScan (Evans 1994, 1999, 2006). It is shown that the GA can improve PLS regression models in calibration and prediction by identifying critical wavelengths for model building. Furthermore, the most representative wavelengths selected by the GA consistently have band assignments arising from the wood components that directly impact pulp yield (e.g., cellulose, hemicellulose, or lignin) and SilviScan measured properties.

In both studies (Ho et al. 2021, 2022), the coefficients of determination of the PLS regression model for calibration and prediction sets (R_c^2 and R_p^2, respectively) and corresponding standard errors (SEC and SEP, respectively) are specified as constraint conditions. In effect, the performance indicators of optimized models have to be better than those in the model using all wavelengths with a small number of latent variables. The GA is employed to optimize an objective function consisting of R_c^2 and R_p^2 with equal weight to balance the optimization process equally in terms of calibration and prediction and to avoid overfitting. However, there are knowledge gaps in utilization of optimization methods to the wood properties prediction models based on NIR data. For example, how is the prediction performance of optimized models affected when different weight factors are used? Or how will the most representative wavelengths, which are selected from optimization process, change with different weight factors?

The objective of this paper is to address these questions by investigating the performance of optimized models and the most representative wavelengths with different weight factors in the objective function. The state-of-the-art global optimization solver pySOT (Eriksson et al. 2015), which implements a surrogate optimization technique as developed in Regis and Shoemaker (2007, 2013), is utilized in this paper to investigate these questions. The optimization process is illustrated by a flow chart shown in Fig. 1. In essence, pySOT evaluates the objective function at several points, internally constructs an approximation to the objective function based on this data and attempts to find points that either maximize the approximation (and hopefully maximize the objective) or explore areas where the objective function has not been evaluated. As the points evaluated are generated randomly at each iteration, pySOT also is a non-deterministic algorithm. Its advantage over the GA is that a smaller number of function evaluations are required and hence more practical for computationally expensive objective functions (Eriksson et al. 2019). Wavelength selection and number of PLS components result from the optimization process using pySOT, and, in addition, the performance of optimized model based on selected wavelengths and different weight factors are compared. Furthermore, a robustness analysis is conducted to suggest sensible choices of the weights used in the objective function.

Materials and methods

Data sets

Two data sets (pulp yield-min and pulp yield-max) containing Tasmanian blue gum samples with different pulp yield variation are used as in Ho et al. (2021). Pulp yield-min contains 67 clonal blue gum samples of the same age and similar pulp yields (ranging from 50.8% to 55.8%), while pulp yield-max consists of 30 blue gum samples from different forests of various ages and with more diverse pulp yields (ranging from 37.6% to 60.2%) (Ho et al. 2021). Both untreated and second derivative datasets for pulp yield-min and pulp yield-max are utilized with the same

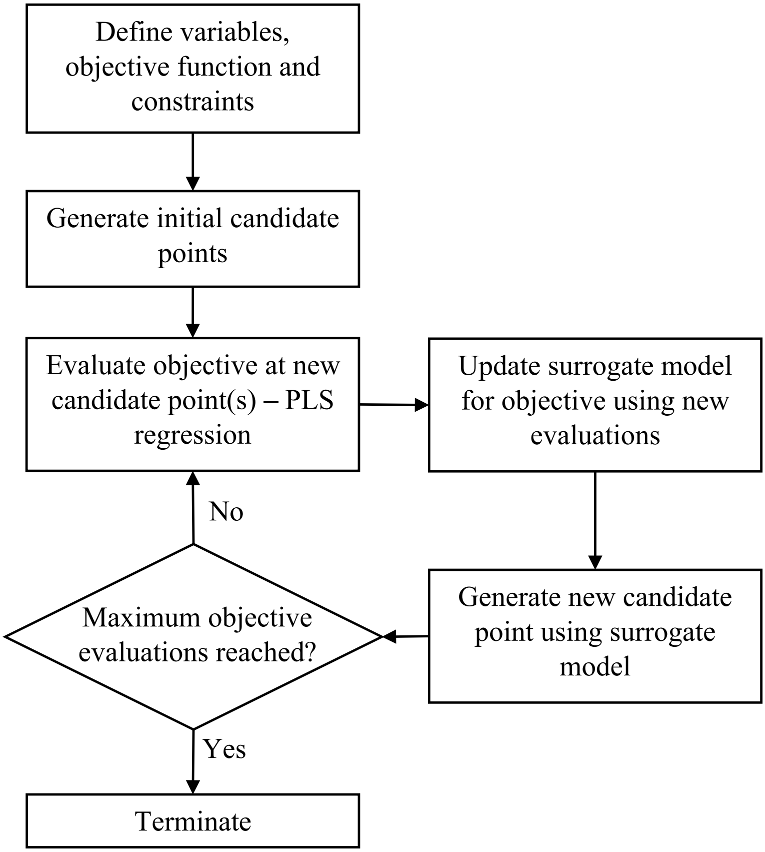
Define variables, objective function and constraints
Generate initial candidate points
Evaluate objective at new candidate point(s) – PLS regression
Update surrogate model for objective using new evaluations
Generate new candidate point using surrogate model
Maximum objective evaluations reached?
No
Yes
Terminate

Fig. 1 Flowchart of pySOT optimization process for NIR wavelength selection

partitioning as Ho et al. (2021) into calibration and prediction datasets based on the DUPLEX selection method (Snee 1977).

Optimization problem

Wavelength selection and number of latent variables (N_{comp}) for PLS regression are investigated at a specific number of wavelengths (N_{Wvl}). Specifically, N_{Wvl} wavelengths are selected from the 700 wavelengths ranging from 1100 to 2500 nm in 2-nm increments (i.e., {1100, 1102, …, 2496, 2498}). These wavelengths, together with N_{comp}, result in (N_{Wvl}+ 1) variables for each PLS model.

The performance of the PLS model is assessed using associated R^2 scores, denoted as R_c^2 and R_p^2 for the calibration dataset and prediction datasets, respectively, with the aim of maximizing both. Noting that these two quantities may be in conflict (i.e., in order to increase one quantity, the other must be decreased), therefore, the problem presented here is a *multi-objective* optimization problem (Ehrgott 2005) with R_c^2 and R_p^2 being the two *objective values* to optimize simultaneously.

The most common goal of an optimization method is to find inputs, which are *Pareto optimal*—where one objective value cannot be unilaterally improved without causing another to worsen. Selecting an optimal input from a collection of Pareto optimal points is a difficult problem, where the relative trade-offs in each objective value must be assessed using expert judgment. In this case, all Pareto optimal points may be found by maximizing an objective function $\alpha R_c^2 + (1 - \alpha)R_p^2$ for all values of α between 0 and 1 inclusive (Ehrgott 2005). Here, α represents a relative weighting of the two objective values: $\alpha = 0$ corresponds to only maximizing R_p^2 without concern for R_c^2, and $\alpha = 1$ corresponds to only maximizing R_c^2.

The optimization in Ho et al. (2021) corresponds to considering $\alpha = 0.5$, and this study extends the analysis by considering the robustness of the wavelength selection process relative to the choice of α. It is expected that choosing α too large would cause overfitting to the calibration data; however, there is no way of knowing what numerical values are "too large" in any single instance. The extent of any possible overfitting and whether the negative impact would be substantial is also unclear. Hence, multi-objective optimization techniques are applied in this study to extend the work of Ho et al. (2021) in order to assess both the overall quality of calibrated NIR models based on automatic wavelength selection and also the impact of the choice of α on the results.

Optimization method

Optimization for $\alpha = 0, 0.1, \ldots, 0.9, 1$ is performed to estimate the full set of Pareto optimal points. There is a standard (single-objective) optimization problem for each value of α with (N_{Wvl}+1) variables that take discrete values within lower and upper bounds. PySOT is employed to solve the optimization problems, and the budget of each optimization is set to be 500 function evolutions. Moreover, each optimization is run 5 times as pySOT is a non-deterministic algorithm.

In order to investigate the impact exerted by the number of latent variables on the performance of the PLS model, the ranges for N_{comp} is selected to be 1–7 or 8–14 or 15–20 for pulp yield-min datasets, and 1–7 or 8–15 for pulp yield-max datasets. As mentioned in Ho et al. (2021), the upper bounds for the number of latent variables (20 for pulp yield-min datasets and 15 for pulp yield-max datasets) are large enough to explain the variance of the full datasets. All the results for N_{Wvl} from 20 to 30 with different ranges of N_{comp} are saved to compare performance of the PLS model among different ranges of N_{comp}.

To investigate the robustness of model performance with respect to α and suggest sensible choices of α, the regions in R_c^2/R_p^2 space, which are covered by the optimized values obtained by different weight vectors with all ranges of N_{comp} and all values of N_{Wvl}, are analyzed.

Finally, representative wavelengths for PLS regression in different ranges of α are inspected. The frequency of each wavelength present in the optimal sets of all optimization cases is determined, as some optimized wavelengths can be just a local maximizer for the objective values and lack generalizability (Ho et al. 2021).

Results and discussion

There are 1210 optimization cases for pulp yield-max datasets and 1815 optimization cases for pulp yield-min datasets. These include the combination of 11 choices of N_{Wvl} (varying from 20 to 30), 11 choices of α (varying from 0 to 1 by 0.1 increments), and two/three different sets of constraints on N_{comp}, with each combination being run five times. The optimization results presented in the following sections consist of the combined results of all five iterations.

Performance of PLS model with different numbers of latent variables

To investigate the impact of N_{comp} on the performance of PLS regression, R_c^2 is plotted against R_p^2 for all optimization cases with the ranges of the number of latent variables being specified in Fig. 2, and the extreme values are listed in Table 1. It is observed that the R_c^2 scores of PLS models with the ranges of larger N_{comp} are significantly greater than that of PLS models with the ranges of smaller N_{comp}, suggesting that larger N_{comp} generally results in better performance in calibration. However,

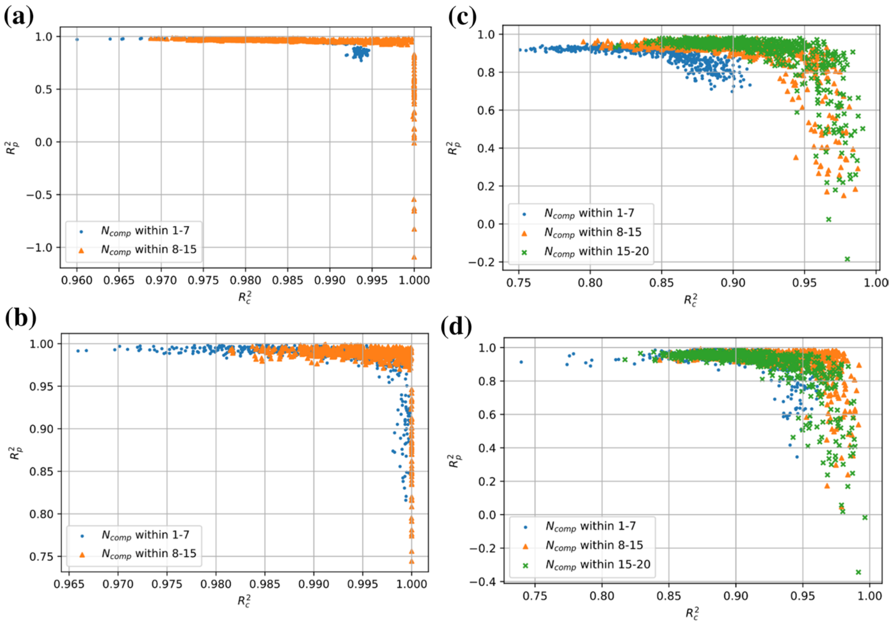

Fig. 2 PLS model performance for different ranges of N_{comp} and different sets of NIR spectra **a** pulp yield-max untreated, **b** pulp yield-max second derivative, **c** pulp yield-min untreated, and **d** pulp yield-min second derivative

Table 1 Extreme values comparison for different ranges of N_{comp}

Dataset	N_{comp}	Min R_c^2	Min R_p^2	Max R_c^2	Max R_p^2
Pulp yield-max raw	1–7	0.960	0.771	0.995	0.982
Pulp yield-max raw	8–15	0.969	– 1.092	1.000	0.986
Pulp yield-max 2nd	1–7	0.966	0.816	1.000	0.999
Pulp yield-max 2nd	8–15	0.981	0.744	1.000	0.999
Pulp yield-min	1–7	0.751	0.698	0.912	0.953
Pulp yield-min	8–14	0.795	0.150	0.987	0.989
Pulp yield-min	15–20	0.818	– 0.185	0.991	0.985
Pulp yield-min 2nd	1–7	0.740	0.345	0.967	0.991
Pulp yield-min 2nd	8–14	0.843	0.048	0.992	0.992
Pulp yield-min 2nd	15 – 20	0.817	– 0.344	0.996	0.988

improvement in prediction performance is small with an increase in N_{comp}, especially for second derivatives datasets. A slight improvement of 0.005 in the highest R_p^2 scores of PLS models is observed in pulp yield-max raw dataset when increasing N_{comp} from 1–7 to 8–15, and no notable difference is observed in the second derivative pulp yield-max dataset. The highest R_p^2 score of PLS models increases from 0.953 to 0.989 and drops to 0.985 when N_{comp} varies from 1–7 to 8–14 to 15–20 for pulp yield-min dataset. A similar trend is observed for second derivative pulp yield-min dataset with much smaller improvement when increasing N_{comp} from 1–7 to 8–14. Moreover, the lowest R_p^2 scores with the ranges of larger N_{comp} are significantly lower than that of smaller N_{comp}. As shown in Table 1, negative R_p^2 scores for PLS models are observed with large N_{comp}(-1.092 when N_{comp} is 8–15 for pulp yield-max dataset; -0.185 and -0.344 when N_{comp} is 15–20 for pulp yield-min datasets). These indicate that if N_{comp} is chosen to be too large, it may exacerbate the overfitting of PLS regression (with improper choice of α), leading to poor performance in prediction. Hence, lower N_{comp} may be selected to decrease computing time while still providing equivalent model performance.

Furthermore, the variation of R_p^2 in untreated pulp yield-max is larger than that in untreated pulp yield-min as PLS models with extremely poor performance in prediction ($R_p^2 \leq -0.5$) are observed for untreated pulp yield-max. This observation may result from the large variation (both in terms of origin and pulp yield) in the samples contained in pulp yield-max.

Choice of Alpha

There are a number of PLS models with nearly perfect calibration performance but poor prediction performance from Fig. 2, suggesting that focusing only on calibration performance (by setting large alpha) results in overfitting. To investigate the range of alpha generating reasonable fit to the prediction data, optimal objective values (R_c^2 and R_p^2) achieved by each value of alpha are analyzed. The regions (convex hulls) of optimal objective values (in R_c^2/R_p^2 space) for each value of alpha are plotted in Fig. 3, and the mean and variance of R_p^2 for each value of alpha are also

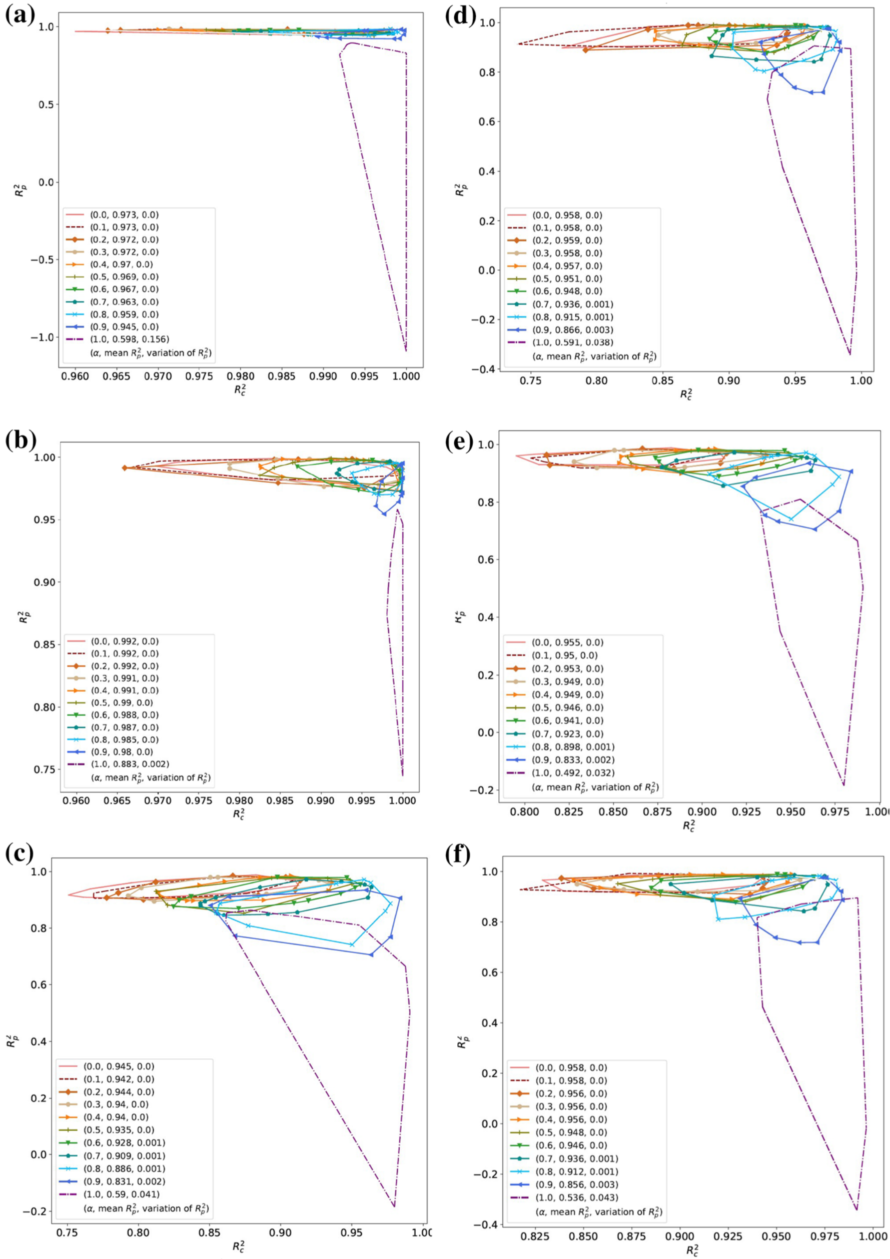

Fig. 3 R_c^2/R_p^2 of different values of α for **a** pulp yield-max untreated, **b** pulp yield-max second derivative, **c** pulp yield-min untreated, **d** pulp yield-min second derivative, **e** pulp yield-min untreated omitting small N_{comp}, and **f** pulp yield-min second derivative omitting small N_{comp}

calculated. The smaller variance and smaller region indicate more robust model performance relative to the choice of value of alpha. As shown in Fig. 3, the variance of R_p^2 increases dramatically when α varies from 0.9 to 1, which is consistent with the rightmost large convex hulls in the figures. The means of R_p^2 are remarkably small at $\alpha = 1$, and significant decreases in R_p^2 start to be observed at $\alpha = 0.9$ as in Fig. 3.

Interestingly, even if a small α is chosen, relatively good calibration results are still obtained, especially for pulp yield-max datasets ($R_c^2 \geq 0.96$ for all values of α). The variation in R_c^2 for pulp yield-min datasets is larger than that for pulp yield-max datasets. It is probably because the number of samples in the pulp yield-min dataset are around two times of the number of samples in the pulp yield-max dataset, and the number of latent variables varying from one to seven is not enough for PLS regression. Omitting the results for model performance with $1 \leq N_{comp} \leq 7$, the variation in R_c^2 is much smaller for untreated data as shown in Fig. 3(e), and R_c^2 is larger than 0.8 for all values of α for both untreated and second derivate treated data as shown in Fig. 3(e–f). Though these results are to some extent dataset specific, the similarities in the performance across the datasets are sufficient to conclude that $\alpha \leq 0.8$ should be chosen in order to obtain a reasonably good predictive performance, and $\alpha \geq 0.9$ tends to overfit the calibration data.

Representative wavelengths

For each wavelength in the domain (i.e., {1100, 1102, …, 2496, 2498}), the number of optimization cases for which it was present in the optimum wavelength set selected by pySOT was counted. Since $\alpha \geq 0.9$ tends to overfit the calibration data, the spectra for $\alpha \leq 0.8$ and $\alpha \geq 0.9$ are plotted separately to observe if the most representative wavelengths selected from the optimization process are impacted by the choice of weight vector. The frequency plots for the four datasets are presented in Figs. 4 and 5, with the ten most frequently identified wavelengths being marked.

The selected wavelengths for $\alpha \leq 0.8$ are more concentrated than wavelengths selected when $\alpha \geq 0.9$. Moreover, untreated spectra and second derivative spectra for different datasets also show different frequency distributions. Yet similarities exist across the datasets in the wood components corresponding to the representative wavelengths, and these wood components are observed to directly impact pulp yield (cellulose, lignin, and hemicellulose).

For $\alpha \leq 0.8$ and untreated NIR spectra, wavelengths utilized most frequently are listed in Table 2. Origins of the wavelengths are based on band assignments presented by Schwanninger et al. (2011) for the primary components of wood. As observed in Ho et al. (2022), regions arising from CH and OH bond vibrations are consistently utilized and include 2100–2400 nm (predominately CH and CH_2 combination bands), 1350–1800 nm (first overtone CH stretch bond vibrations), 1400–1600 nm (first overtone OH stretch bond vibrations) and 1140–1225 nm (second overtone CH stretch bond vibrations), regions where bands assigned to cellulose and lignin are common.

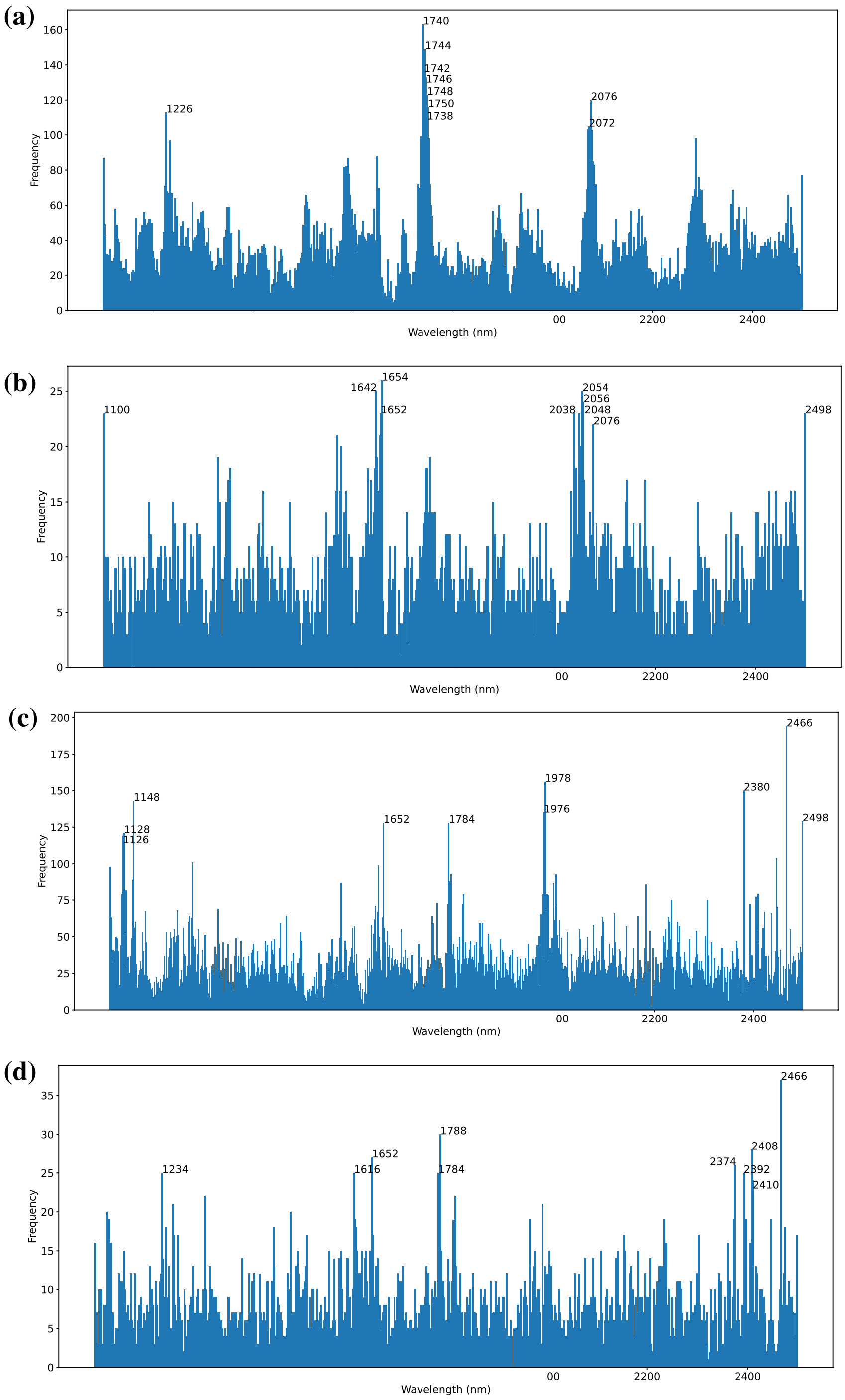

Fig. 4 Pulp yield-max for **a** untreated spectra for $\alpha \leq 0.8$, **b** untreated spectra for $\alpha \geq 0.9$, **c** second derivative spectra for $\alpha \leq 0.8$, and **d** derivative spectra $\alpha \geq 0.9$

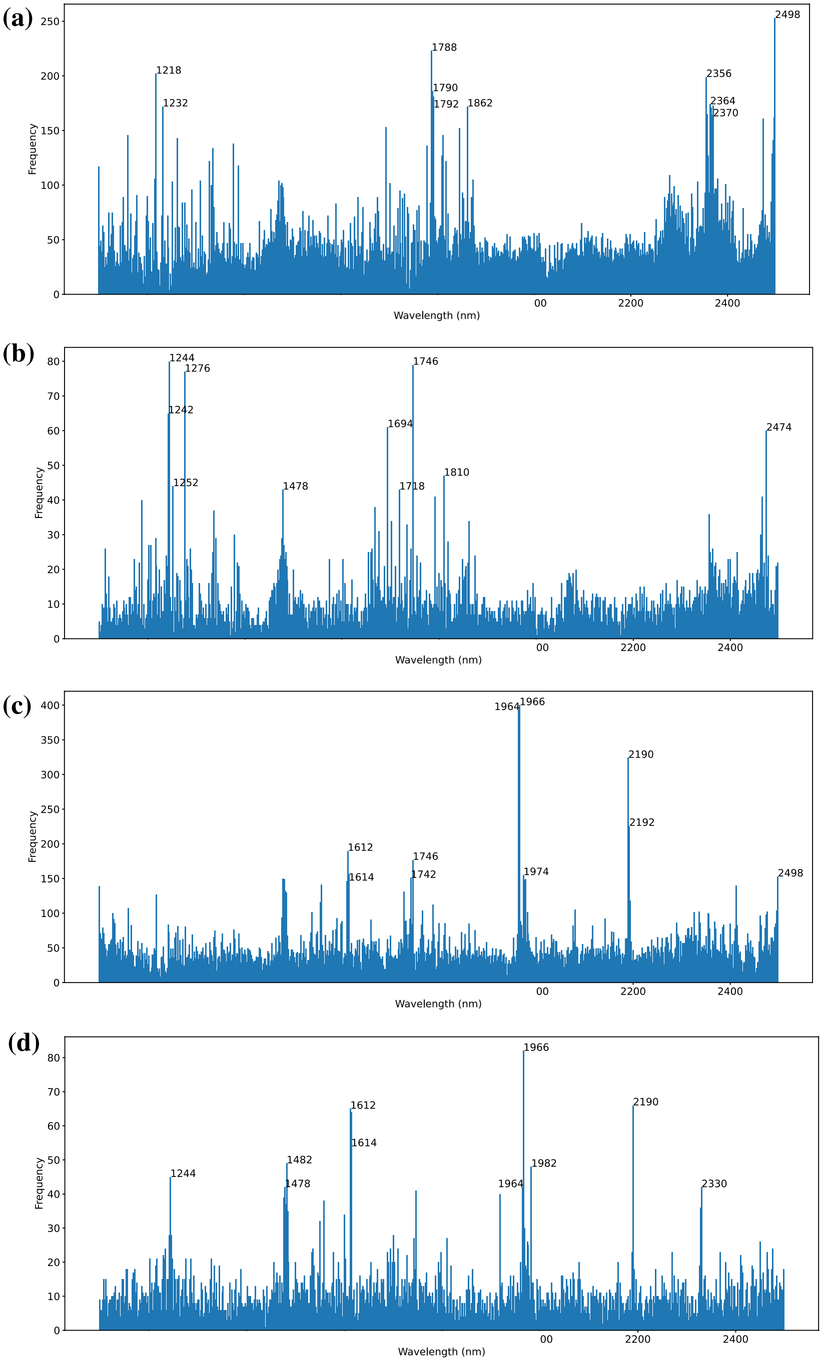

Fig. 5 Pulp yield-min for **a** untreated spectra for $\alpha \leq 0.8$, **b** untreated spectra for $\alpha \geq 0.9$, **c** second derivative spectra for $\alpha \leq 0.8$, and **d** derivative spectra $\alpha \geq 0.9$

Table 2 Band assignments for optimization when $\alpha \leq 0.8$

Identified wavelengths (nm)	Band location (nm)	Bond vibration	Wood component
Pulp yield-max untreated			
1220–1230	1212–1225	2nd OT C-H str	Cellulose
1580	1580	1st OT O–H str	Cellulose
1660	1666	1st OT C-H str	Hemicellulose
1740–1750	1731	1st OT C-H str	Cellulose
2070–2080	2080	O–H str. + C-H def	Cellulose
2280	2277	O–H str. + C-O str. and/or C-H str. + C-H def	Cellulose
Pulp yield-max 2nd deriv			
1120–1130	N/A		
1148	1143	2nd OT C$_{ar}$-H str 2nd OT C-H str. CH$_3$ groups	Lignin
1652	1666	1st OT C-H str	Hemicellulose
1784	1780, 1788	1st OT C-H str	Cellulose
1976, 1978	1980	O–H str. + O–H def	Water
2380	2384	N/A	Lignin
2466	2461	C-H str. + C–C str	Starch (cellulose)
2498	2491	C-H str. + C–C str	Cellulose
Pulp yield-min untreated			
1218, 1232	1212–1225	2nd OT C-H str	Cellulose
1780–1790	1731	1st OT C-H str	Cellulose
1862	N/A		
2356–2370	2352	C-H str. + C-H def 2nd OT C-H$_2$ bend	Cellulose
	2361	O–H def. or C-H def + C-H str. or C-H$_2$ str	Cellulose
2498	2491	C-H str. + C–C str	Cellulose
Pulp yield-min 2nd deriv			
1612, 1614	1616	1st OT C-H str	N/A
1964, 1966, 1974	1980	O–H str. + O–H def	Water
2190, 2192	2200	C-H str. + C = O str	Lignin
2498	2491	C-H str. + C–C str	Cellulose

OT overtone, *str.* stretching vibration, *bend* bending vibration, *def.* deformation vibration, + combination band; *ar* aromatic; *t.a.* tentative assignment; *N/A* no vibration specific to wood component assigned

With the exception of 1862 nm (pulp yield-min), all important wavelengths had recognized bond vibrations occurring at, or within 5 to 10 nm of, a wavelength band assigned to a wood component. For second-derivative-treated NIR spectra and $\alpha \leq 0.8$, similar correspondence is observed with only two wavelengths in the region 1120–1130 nm (pulp yield-max) not related to a bond vibration arising from a specific wood component (Table 2).

For $\alpha \geq 0.9$, the most frequently used wavelengths generally occurred in regions unrelated to vibrations arising from specific wood components. While weaker predictive performance may be expected for models emphasizing calibration statistics, this observation may help explain why they do not perform as well as the models developed using values of α that balance both calibration and prediction performance.

Discussion on wavelength selection problem and optimization algorithms

Optimization can improve PLS calibration performance by identifying a relatively small number of NIR wavelengths for model development (Ho et al. 2021, 2022). Weighting of objective values is a critical decision in this approach, and it was chosen to balance calibration and prediction performance, i.e., $\alpha = 0.5$. To understand the impact of different values of α (0, 0.1, …, 0.9, 1) on the robustness of wavelength selection, the global optimization solver pySOT was utilized to optimize PLS regression model performance. Using pySOT, it was demonstrated that models are robust over a wide range of α values (0 to 0.8).

A key feature of optimized PLS models is the consistent identification of wavelengths occurring in NIR regions associated with C-H and O–H bond vibrations arising from cellulose, hemicellulose, and lignin. The identification of wavelengths whose origin is directly related to the main components of wood explains why both calibration and prediction performance (for samples included in the optimization process) can be improved.

Further work is required to investigate the applicability of optimized models to the estimation of wood properties, for example a comparison of estimated pulp yield genetic parameters (Raymond et al. 2001; Schimleck 2008) based on predicted pulp yield data from PLS models and optimized PLS models. Ho et al. (2022) investigated GA optimized model prediction of microfibril angle (MFA) for loblolly pine; however, the prediction set employed (referred to as “Maps”) differed in age and geographic origin and was unrelated to the calibration / prediction samples (“Agenda 2020”) used for model development. Ho et al. (2022) recognized this issue and stated: “The Agenda 2020 model did not perform as well as the optimized MFA model that incorporated the Maps data for its prediction phase, indicating that characteristic wavelengths for the Maps data set differ from those of the Agenda 2020 samples.” This is an important finding and while consistent with other NIR-based studies of wood that have attempted to apply models based on samples from one location to those from a different location (Schimleck 2008). It indicates that optimized models are also site(s) limited and that important wavelengths likely differ to some extent with age and location. Hence, an assessment of optimized model performance should be based on samples sharing the same characteristics (age and location) as those used for model development but exclusive of the calibration/prediction phases of optimization. This question should be explored in future studies.

Furthermore, the performance of optimized models can be evaluated by comparing among different optimization methods in terms of associated computational times and trade-offs between optimized and unoptimized models. Utilizing an optimization approach not only likely results in better models, but also adds time and complexity (in terms of model development). Further studies will answer whether better predictive models help achieve incremental increases in the value of wood products derived from forests whose wood properties are accurately characterized, and what types of applications will most benefit from the optimization results.

Conclusion

In this study, pySOT is utilized to develop PLS regression pulp yield models based on two *Eucalyptus globulus* data sets (characterized by narrow and extreme yield ranges) and examine predictive performance of models using different weight factors (i.e., factor α in the objective function $\alpha R_c^2 + (1 - \alpha)R_p^2$ to be maximized in the optimization algorithm). Values of $\alpha \leq 0.8$ provided good predictive performance, whereas $\alpha \geq 0.9$ tended to overfit the calibration data indicating that models are robust for values of α from 0 to 0.8. Representative wavelengths for each data set were identified. For $\alpha \leq 0.8$, almost all wavelengths arose from C-H and O–H bond vibrations observed in cellulose, hemicellulose, and lignin. However, identified wavelengths for $\alpha \geq 0.9$ generally occurred in regions, which are unrelated to vibrations arising from specific wood components.

The reliability with which pySOT identified chemically relevant representative wavelengths indicates its usefulness as a global optimization algorithm in this setting. However, because the underlying algorithm is randomized, repeated runs of the optimization are important to assess the variability in results, so deterministic methods may be overall faster (and give more reproducible representative wavelengths).

There was consistency between the optimal ranges of alpha between the two *Eucalyptus globulus* data sets, but it would be useful in future work to assess whether this generalizes to other species/data sets, or whether alternative metrics of PLS performance (such as RMSE) are suitable. Testing the generalizability of the identified representative wavelengths on a separate partition of data (i.e., not calibration or prediction) would also give further evidence of the utility of this framework but is more suited to larger data sets than those used in this work. In addition, the applicability of optimized models to the estimation of wood properties, the comparison between optimization methods and associated computational times, and trade-offs between optimized and unoptimized models are recommended for future studies.

Declarations

Conflict of interest The authors declare no conflict of interest.

References

Ayanleye S, Nasir V, Avramidis S, Cool J (2021) Effect of wood surface roughness on prediction of structural timber properties by infrared spectroscopy using ANFIS, ANN and PLS regression. Eur J Wood Prod 79(1):101–115

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

Cogdill RP, Schimleck LR, Jones PD, Peter GF, Daniels RF, Clark A (2004) Estimation of the physical wood properties of *Pinus taeda* L. radial strips using least squares support vector machines. J Near Infrared Spectrosc 12(4):263–270

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. In: IEEE 7th international advance computing conference (IACC), Hyderabad, 2017 pp 184–187

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

Downes GM, Worledge D, Schimleck LR, Harwood C, French J, Beadle CL (2006) The effect of growth rate and irrigation on the basic density and kraft pulp yield of *Eucalyptus globulus* and *E. nitens*. N Z J For 51(3):13–22

Ehrgott M (2005) Multicriteria Optimization. Springer, Berlin Heidelberg, Germany

Eldridge KG, Davidson J, Harwood CE, vanWyk G (1993) Eucalypt domestication and breeding. Oxford University Press, Oxford

Eriksson D, Bindel D, Shoemaker CA (2015) Surrogate optimization toolbox (pySOT) (2015) Available from https://github.com/dme65/pySOT

Eriksson D, Bindel D, Shoemaker CA (2019) pySOT and POAP: An event-driven asynchronous framework for surrogate optimization. ArXiv, abs/1908.00420

Evans R (1994) Rapid measurement of the transverse dimensions of tracheids in radial wood sections from *Pinus radiata*. Holzforschung 48:168–172

Evans R (1999) A variance approach to the X-ray diffractometric estimation of microfibril angle in wood. Appita J 52(283–289):294

Evans R (2006) Characterization of the cellulosic cell wall. Stokke DG, Groom L (ed) pp 138–146. Blackwell Publishing, Ames, IA, USA

Fernandes A, Lousada J, Morais J, Xavier J, Pereira J, Melo-Pinto P (2013) Measurement of intra-ring wood density by means of imaging VIS/NIR spectroscopy (hyperspectral imaging). Holzforschung 67(1):59–65

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

Ho TX, Schimleck LR, Sinha A (2021) Utilization of genetic algorithms to optimize Eucalyptus globulus pulp yield models based on NIR spectra. Wood Sci Technol 55(3):757–776

Ho TX, Schimleck LR, Sinha A, Dahlen J (2022) Utilization of genetic algorithms to optimize loblolly pine wood property models based on NIR spectra and SilviScan data. Wood Sci Technol 56:1419–1437. https://doi.org/10.1007/s00226-022-01403-z

Li Y, Via BK, Cheng Q, Zhao J, Li Y (2019) New pretreatment methods for visible–near-infrared calibration modeling of air-dry density of *Ulmus pumila* wood. For Prod J 69(3):188–194

Michell AJ, Schimleck LR (1996) NIR spectroscopy of woods from *Eucalyptus globulus*. Appita J 49(1):23–26

Mora C, Schimleck LR (2010) Kernel regression methods for the prediction of wood properties of *Pinus taeda* using near infrared (NIR) spectroscopy. Wood Sci Technol 44(4):561–578

Nasir V, Nourian S, Zhou Z, Rahimi S, Avramidis S, Cool J (2019) Classification and characterization of thermally modified timber using visible and near-infrared spectroscopy and artificial neural networks: a comparative study on the performance of different NDE methods and ANNs. Wood Sci Technol 53(5):1093–1109

Raymond CA, Schimleck LR, Muneri A, Michell AJ (2001) Genetic parameters and genotype-by-environment interactions for pulp yield and pulp productivity in *Eucalyptus globulus* predicted using near infrared reflectance analysis. For Genet 8(3):213–224

Regis RG, Shoemaker CA (2007) A stochastic radial basis function method for the global optimization of expensive functions. INFORMS J Comput 19(4):497–509

Regis RG, Shoemaker CA (2013) Combining radial basis function surrogates and dynamic coordinate search in high-dimensional expensive black-box optimization. Eng Optim 45(5):529–555

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 (2008) Near infrared spectroscopy: A rapid, non-destructive method for measuring wood properties and its application to tree breeding. N Z J for Sci 38(1):14–35

Schimleck LR, Tsuchikawa S (2021) Application of NIR spectroscopy to wood and wood derived products (Chapter 37). In: Ciurczak E, Igne B, Workman J, Burns D (eds) The handbook of near-infrared analysis, fourth edition, newly revised and expanded. CRC Press, Boca Raton, FL, pp 759–780

Schwanninger M, Rodrigues JC, Fackler K (2011) A review of band 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

Trung T, Downes G, Meder R, Allison B (2015) Pulp mill and chemical recovery control with advanced analysers - from trees to final product. Appita J 68(1):39–46

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

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

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