Analysis of Mixed Pulping Raw Materials of *Eucalyptus globulus* and *Acacia mangium* by Near Infrared Spectroscopy Technique Combined with LASSO Algorithm

Ting Wu,[a] Guigan Fang,[a,b,*] Long Liang,[a] Yongjun Deng,[a] Yan Lin,[a] and Zhixin Xiong [a,b]

To meet the current demand in China for *Eucalyptus globulus* and *Acacia mangium* mixed pulping, a study was conducted to collect the near infrared (NIR) spectra of 150 mixed samples of *E. globulus* and *A. mangium* in which the content of *E. globulus* was manually controlled. After the original spectra were pretreated by first derivative and standard normal variate (SNV), the least absolute shrinkage and selection operator (LASSO) algorithm and cross-validation were used to calculate the optimal adjustment parameters of 14.30, 19.16, 12.10, and 9.74, respectively. The optimal calibration models for the content of *E. globulus*, holocellulose, pentosan, and acid insoluble lignin were generated. An independent verification of the calibration models showed that the root mean square error of prediction (RMSEP) for these models was 1.59%, 0.54%, 0.66%, and 0.40%, respectively. The absolute deviation (AD) was -2.58% to 2.73%, -0.91% to 0.84%, -1.19% to 1.06%, and -0.61% to 0.64%, respectively. The prediction performance of the four models was sufficient for real-time analysis in the pulping production line. The LASSO algorithm was judged to be efficient for the prediction and analysis of mixed raw materials in pulping industry.

*Contact information: a: Institute of Chemical Industry of Forest Products, Chinese Academy of Forestry, Jiangsu Province, Nanjing 210042, China; b: Collaborative Innovation Center for High Efficient Processing and Utilization of Forestry Resources, Nanjing Forestry University, Nanjing 210037, China; c: College of Light Industry Science and Engineering, Nanjing Forestry University, Nanjing 210037, China; *Corresponding author: ppfangguigan@163.com*

INTRODUCTION

Mixed raw materials of *Eucalyptus globulus* and *Acacia mangium* are sometimes used in the Chinese pulping industry to complement advantages of each other and obtain higher pulp properties, with moderate single raw material requirements (Chen *et al.* 2009; Deng *et al.* 2015). During production, it is difficult to ensure that mixed raw materials have a consistent mixing uniformity because raw materials obtained from different locations or stored in different conditions have varied chemical compositions (Muhammad *et al.* 2017). If the original pulping parameters are used in the production process, the pulp properties are substandard (Qu and Fang 2009; Tsuchikawa and Schwanninger 2013). To ensure normal operation of the production, the real-time analysis of mixed raw materials of *E. globulus* and *A. mangium* is necessary to adjust the pulping parameters online. However, the conventional chemical analysis methods are often time-consuming and unable to determine the mixing degree of raw materials (Downes *et al.* 2012).

Near infrared (NIR) spectra are generated by anharmonic levels of vibration of covalent bonds and reflect the information of hydrogen groups X-H (X= O, C, N, S). The chemical composition of wood raw materials contains a large number of hydrogen-containing groups, so near infrared spectroscopy can be used to analyze the main chemical constituents of wood raw materials (Xu *et al*. 2000).

In addition, near infrared light in the sample will undergo reflection, refraction, diffraction, absorption, and interaction with the internal molecules of the sample. As a result of this process, near infrared spectra contain a great deal of compositional and structural information, which can be used for the analysis of physical properties (Wang *et al*. 2008). There are many overlapping bands in near infrared spectra, and it is difficult to distinguish the spectral bands. Therefore, it is not possible to read the useful information directly, and it is necessary to optimize the spectral data with the help of computer and chemometric methods. The analytical process of near infrared spectroscopy is as follows. First of all, typical samples are selected as calibration set samples and near infrared spectra are collected. Thereafter, standard methods or widely accepted methods are used to determine the composition or structure of calibration set samples. The calibration model is established between near infrared spectra and sample information by the chemometric method.

Finally, the near infrared spectra of unknown samples are collected and the spectral data are loaded in calibration set to predict unknown sample information. Such a procedure can improve the efficiency of regular quantitative analysis (Alves *et al*. 2012). As a rapid, non-destructive analysis method, near infrared spectroscopy combined with chemometrics has been widely used in agriculture (Labbé *et al*. 2008; Talens *et al*. 2013), forestry (Stirling 2013), and petrochemicals (Balabin *et al*. 2011). It also has played an important role in pulping industry. The identification of common pulping raw materials was realized with high recognition rate 100% (Cui and Fang 2015). The air-dry density of *Eucalyptus pellita* was analyzed and optimal analysis conditions were determined (Zhao *et al*. 2012). The determination of specific chemical constituents in specific pulpwood species were also studied and showed good results (Hodge and Woodbridge 2010; Mora and Schimleck 2010; Yao *et al*. 2010). However, these researches did not attempt to analyze the typical mixed raw materials in the pulping industry, and the algorithm used was traditional partial least squares.

The least absolute shrinkage and selection operator (LASSO) algorithm has a strong out-sample prediction ability. It is widely used in economics (Yu and Zhang 2014) and statistics (Shi *et al*. 2012). The aim of this study was to combine near infrared spectroscopy with the LASSO algorithm to perform rapid analyses of mixed *E. globulus* and *A. mangium* raw materials. In the pulping industry, the content of the holocellulose is directly related to the pulp yield and the hemicellulose, which in broad-leaved wood is usually expressed as pentosans.

The chemical content has a great influence on the pulp yellowing and paper-forming properties. The lignin content is related to the degree of chemical dosage and bleaching in pulping process. The acid soluble lignin molecule is small and easy to dissolve, and it has little effect on the pulping process (Li 2001). Based on the above reasons, the contents of *E. globulus*, holocellulose, pentosans, and acid insoluble lignin in mixed raw materials were considered as the most important indices.

EXPERIMENTAL

Materials

Eucalyptus globulus Labill. aged 4 to 6 years was obtained from Guangdong province in China. *Acacia mangium* Willd. aged 6 to 7 years was obtained from Fujian Province in China. The wood was peeled and ground into powder. All powder that passed through a 40-mesh sieve and was retained on the 60-mesh sieve was collected. Two kinds of wood powder were put under the environmental condition of constant temperature, constant humidity until the equilibrium moisture content was reached; then they were mixed into 150 samples (Li *et al*. 2017) in different given proportions. The content of *E. globules* was artificially controlled around 0/149, 1/149, 2/149 ... 149/149 proportion, ensure uniform coverage of 0 to 100% intervals. Of the 150 samples, 30 samples were chosen according to the content gradient as a validation set for the independent verification of models. The remaining 120 samples were used to establish the models as a calibration set. Additionally, 10 *E. globulus* and 10 *A. mangium* single wood samples were collected to determine the chemical composition.

Spectroscopic Method

A Hadamard transform near infrared spectrometer (Zhejiang China Invent NIR1000, Jiaxing, China) was used to collect the near infrared spectra of all 150 mixed samples. Spectra were acquired by 50 scans over the wavelength range of 1600 to 2400 nm, with the spectral points of 100. For comprehensive sample analysis, the original spectrum of each sample took the average of five spectra.

Chemical Method

The holocellulose content of 150 mixed samples and 20 single samples were determined by GB/T 2677.10 (1995). Accordingly, sodium chlorite was used in the pH range 4 to 5 to treat the sample after extraction of benzene/ethanol to remove the lignin, then the holocellulose was determined gravimetrically.

The pentosan content of 150 mixed samples and 20 single samples were determined by GB/T 2677.9 (1994). Accordingly, 0.5 g of sample was boiled with 12% HCl in order to hydrolyse pentosan to pentose. Then pentose dehydrated into furfural. The distilled furfural was collected by condensation, and the furfural content was measured quantitatively by bromination method, and converted to the content of pentosans.

The acid insoluble lignin content of 150 mixed samples and 20 single samples were determined by GB/T 2677.8 (1994). In this method, 1 g of sample extracted with benzene/ethanol was first treated with 72% H_2SO_4 for 2 h at 20 °C. Subsequently, the sample was diluted with water to reduce the H_2SO_4 concentration to 3% and boiled for 4 h in order to hydrolyse the polysaccharides to soluble monosaccharides. The solid residue (acid insoluble lignin) is then collected and weighed.

Lasso Algorithm and Analysis Procedure

The LASSO algorithm conducts a biased estimation on data with multicollinearity. Specifically, it minimizes the residual sum of squares (RSS) given the constraint that the L^1 norm of the coefficient vector must be smaller than a constant. It generates a model with stronger explanatory power because some of the coefficients are strictly equal to zero (Tibshirani 1997).

The first assumption is that there are independent variables x_1, x_2, ..., x_p and a dependent variable y that follow the linear regression relationship shown in Eq. 1,

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \epsilon \tag{1}$$

where α is a constant term, β_1, β_2, ..., β_p are regression coefficients, and ϵ is the random disturbance term.

Next, assume $(x_{i1}, x_{i2}, ..., x_{ip};\ y_i)$, i=1, 2, ..., n are observations of variables. Furthermore, assume the data has been centrally standardized, meaning that $\sum_{i=1}^{n} y_i = 0$, $\sum_{i=1}^{n} x_{ij} = 0 \sum_{i=1}^{n} x_{ij}^2 = 1$, j =1, 2, ..., p. Meanwhile, $\beta = (\beta_1,\ \beta_2,\ ...,\ \beta_p)^T$.

Using an algorithm that penalized the sum of the absolute values of the coefficients solves a problem that minimizes the sum of RSS and a penalty function that increases with the absolute values of the coefficients, as shown in Eq. 2,

$$(\hat{\alpha}, \hat{\beta}) = \arg\min \sum_{i=1}^{n}(y_i - \alpha - \sum_{j=1}^{p}\beta_j x_{ij})^2, subject\ to\ \sum_{j=1}^{p}|\beta_j| \leq \lambda \tag{2}$$

where $\lambda \geq 0$ is a constraint constant. For arbitrary $\lambda \geq 0$, the solution of α in Eq. 2 is $(\hat{\alpha}) = 0$ because the data have already been centrally standardized. The constrained optimization problem is equivalent to an unconstrained optimization problem with a penalty function. Equation 2 is the same as Eq. 3,

$$\hat{\beta}(Lasso) = \arg\min\{\sum_{i=1}^{n}\ (y_i - \sum_{j=1}^{p}\beta_j x_{ij})^{\ 2} + \mu\sum_{j=1}^{p}|\beta_i|\} \tag{3}$$

where μ is a penalty constant. As μ increases, the $\sum_{j=1}^{p}|\beta_i|$ term in the optimal solution decreases. The dimensionality of the dataset is therefore reduced, since some coefficients shrink to 0 during the process. With the penalty on L^1 norm, LASSO sets the coefficients of some uninformative variables to 0 and keeps an information-intensive subset of the variables. The model will have the largest coefficient of determination (R^2) possible given the constraint on L^1 norm of the coefficients (Tibshirani 2011). Each μ value corresponds to a LASSO solution, so the crux of the LASSO algorithm is the determination of the optimal adjustment parameter μ. In general, Matlab software (MathWorks, Natick, MA, United States of America) and cross validation method are used to calculate the predicted residual sum of squares (PRESS), and when the PRESS is minimum, the μ value is optimal. At the same time, to make $\hat{\alpha} = 0$, the optimal model parameters are computed.

Evaluation Standard

Model evaluation methods consist of cross validation and independent verification. The R^2 of cross validation (R^2_{cv}) and root mean square error of cross validation (RMSECV) are reference data. The R^2 of independent verification (R^2_{val}) and the ratio of performance to deviation (RPD) of independent verification are secondary evaluation standards. The root mean square error of prediction (RMSEP) and absolute deviation (AD) in independent verification are the main evaluation standards (Feng *et al.* 2016). Usually, a value of R^2 closer to 1 indicates better regression precision and prediction precision of the model. However, the R^2 value is closely related to the distribution range of the sample properties, so it is not the main basis for judging the model performance. The RPD value standardizes the predictive result of the model such that a higher value represents better prediction precision. However, the value size is also related to the sample properties distribution range and is thus not independent as the model evaluation standard. The RMSEP is the basic index of evaluating model accuracy and reliability. The AD directly reflects the difference

between predicted value and measured value, and its range determines the upper and lower limit of the predictive error, which should conform to the permissible error range of the model application. Finally, the system error in the independent verification is reflected by the bias value, which is the mean value of the algebraic sum of AD (Krasznai *et al.* 2012).

RESULTS AND DISCUSSION

Distribution of Measured Values

The content distribution of *E. globulus* and chemical composition of all 150 mixed samples is shown in Table 1. The content distribution of *E. globulus* was uniformly covered by the entire range of 0% to 100%. The content of holocellulose in 10 *E. globulus* samples was 79.04% to 81.57%, the mean value was 80.50%, and the standard deviation was 0.88%. The content of holocellulose in 10 *A. mangium* samples was 74.00% to 76.12%, the mean value was 75.27%, and the standard deviation was 0.76%. The content of holocellulose in 150 mixed samples was 74.51% to 81.46%. The content of pentosans in 10 *E. globulus* samples was 26.25% to 30.50%, the mean value was 28.26%, and the standard deviation was 1.36%. The content of pentosans in 10 *A. mangium* samples was 20.93% to 26.58%, the mean value was 23.47% and the standard deviation was 1.97%. The content of pentosan in 150 mixed samples was 21.72% to 30.47%. The content of acid insoluble lignin in 10 *E. globulus* samples was 23.34% to 26.60%, the mean value was 24.85%, and the standard deviation was 1.08%. The content of acid insoluble lignin in 10 *A. mangium* samples was 22.63% to 27.19%, the mean value was 25.11%, and the standard deviation was 1.54%. The content of acid insoluble lignin in 150 mixed samples was 22.67% to 26.93%. The chemical compositions of mixed samples were widely distributed, which covered the probable range.

Table 1. Content Distribution of the Mixed Samples

Content	Number	Min. (%)	Max. (%)	Mean (%)	Median (%)	Standard Deviation (%)
Eucalyptus globulus	150	0	100	50.52	50.06	29.16
Holocellulose	150	74.51	81.46	77.92	77.85	2.14
Pentosan	150	21.72	30.47	25.86	25.54	2.48
Acid insoluble lignin	150	22.67	26.90	24.95	25.17	1.26

NIR Spectra of Mixed Samples and Pretreatment

Figure 1 shows the original near infrared spectra of mixed samples, in which the horizontal axis is the spectral wavelength, and the ordinate reflects the absorption. The sample spectra were similar and difficult to distinguish. The reason is that wood raw materials has a complex composition, not only including cellulose, pentosans, and other polysaccharides, lignin and other aromatic compounds, but also contains resin, tannin, pigments, and minerals. In addition, the intensity of the near infrared spectroscopy signal was weak, and the overlapping interference of spectrum was distinct. Thus, it was necessary to preprocess the spectral data.

Usually the pretreatment methods of near infrared spectra include derivative method, multiscatter calibration (MSC) method, and standard normal variate (SNV) method. The first derivative method and the second derivative method are used to eliminate the shifting and interference of the baseline in the spectra, which can effectively eliminate the interference of other backgrounds, distinguish overlapping peaks, and improve resolution and sensitivity. The MSC method is used to eliminate the influence of the scattering of particles and particle size on the spectra. The effect of SNV method is similar to that of MSC method, but the calibration ability is stronger (Sun 1997). In order to eliminate the shift of the baseline in the spectra and to eliminate the nonspecific scattering caused by the particle heterogeneity of the sample, the first derivative and SNV method was adopted (Huang *et al*. 2012). Figure 2 contains the pretreated spectra.
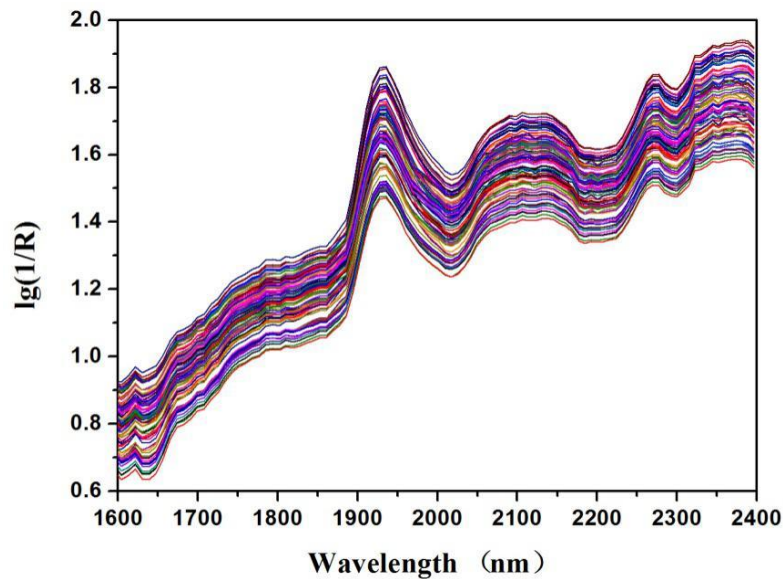

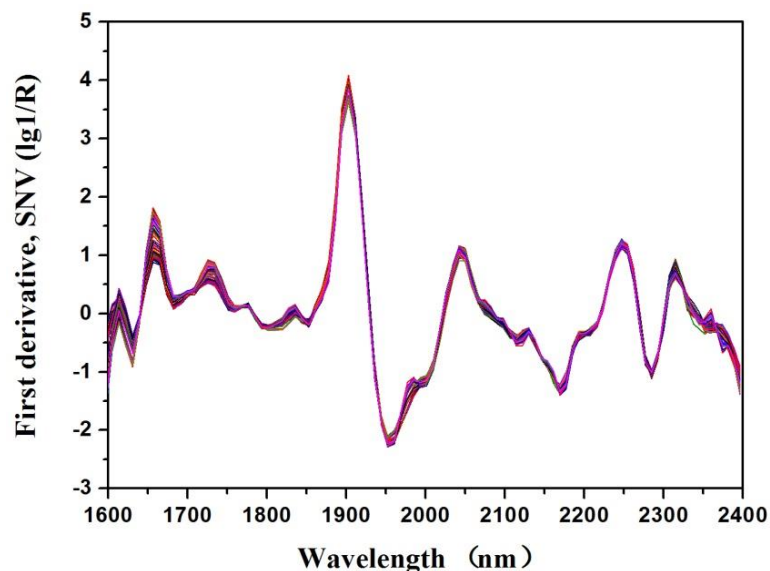**Fig. 1.** The near infrared spectra of 150 mixed samples

Fig. 2. The near infrared spectra pretreated by first derivative and SNV

Model Establishment

Due to the overlapping of output in the near infrared spectra, some shift around the original peak position was inevitable. The important spectral regions for the prediction were identified previously. The biggest spectral peak of the samples was observed at the wavelength of about 1900 nm, and it can be attributed mainly to the absorption of water. The spectral regions of 1600 nm to 2355 nm were mainly associated with holocellulose. The spectral regions of 1600 nm to 1836 nm and 2173 nm to 2355 nm were mainly associated with pentosans. The spectral regions of 1695 nm to 2394 nm were mainly associated with the acid insoluble lignin (Schwanninger *et al*. 2011; He and Hu 2013). However, the spectral points of Hadamard transform spectrometer used in this experiment were less than for the Fourier transform spectrometer that was used in the above studies. If bands selection is used, more useful information may be lost, so the whole-band modeling was adopted.

The LASSO algorithm was used in Matlab software. After the pretreated spectral data and content data of the calibration set was loaded, the "leave one out" method was used to carry out cross validation. That is, the calibration set left a sample to be predicted, and the other samples were used for the model establishment. This process was repeated until each sample in the calibration set had been predicted 1 time and participated in the model establishment 119 times (Zhong *et al*. 2016). When the PRESS values of models of *E. globulus*, holocellulose, pentosan, and acid insoluble lignin content were minimum, μ was the optimal adjustment parameter, and the values were 14.30, 19.16, 12.10, and 9.74, respectively. At the same time, to make $\hat{\alpha} = 0$, the corresponding model parameters were solved as optimal model parameters, and the optimal calibration models were obtained. The cross validation of calibration models is shown in Table 2. The R^2_{cv} value of the *E. globulus* content model was 0.9982 because of the wide coverage of the content of *E. globulus*. All four models of *E. globulus*, holocellulose, pentosans, and acid insoluble lignin content had a good correlation and accuracy in cross validation.

Table 2. Parameters and Evaluation of the Models

Content	The Optimal Adjustment Parameter	Cross Validation	
	μ	R^2_{cv}	RMSECV
Eucalyptus globulus	14.30	0.9982	1.36%
Holocellulose	19.16	0.9450	0.48%
Pentosan	12.10	0.9497	0.52%
Acid insoluble lignin	9.74	0.9326	0.31%

Independent Verification

Four calibration models were used to predict the samples in the validation set. In the Matlab software, the calibration models were filed, and the spectral data of the validation set after pretreatment were loaded. The predicted values were obtained by calculation and analysis, and the relative parameters of the models were obtained by comparing the predicted values with the measured values, as shown in Table 3. The R^2_{val} and RPD values of the *E. globulus* content model were not significant because of the wide coverage of the content. The value of RMSEP was 1.59%, and AD was -2.58% to 2.73%. The error was relatively large but suitable for rough prediction. This was also consistent with the objective of mixing degree evaluation in the pulping industry. The RMSEP values of the holocellulose, pentosans, and acid insoluble lignin models were 0.54%, 0.66%, and 0.40%, respectively. The AD were -0.91% to 0.84%, -1.19% to 1.6%, and -0.61% to

0.64%, respectively. The error was small and essentially conformed to the requirement of measuring error in the pulping industry, and it could adapt to the accurate determination of chemical composition content. In order to determine the difference between the LASSO algorithm and the traditional algorithm in the modeling precision, the partial least squares (PLS) operation program was loaded in Matlab, and four PLS calibration models were established and used to predict the samples in validation set. In independent verification of four PLS models, the RMSEP value of *E. globulus* content model was 1.91%, and AD was -3.17% to 3.42%. The RMSEP value of holocellulose content model was 0.59% and AD was -0.90% to 1.03%, the RMSEP value of pentosan content model was 0.67%, and AD was -1.17% to 1.10%. The RMSEP value of acid insoluble lignin content was 0.44% and AD was -0.65% to 0.71%. Obviously, the LASSO algorithm was shown to be superior to the PLS algorithm when analyzing the content of *E. globules*. In the analysis of holocellulose and acid insoluble lignin content LASSO algorithm was slightly better than the PLS algorithm. The accuracy of the two algorithms was similar when analyzing the pentosan content. Generally speaking, the LASSO algorithm was superior to PLS algorithm in the above pulping raw materials analysis.
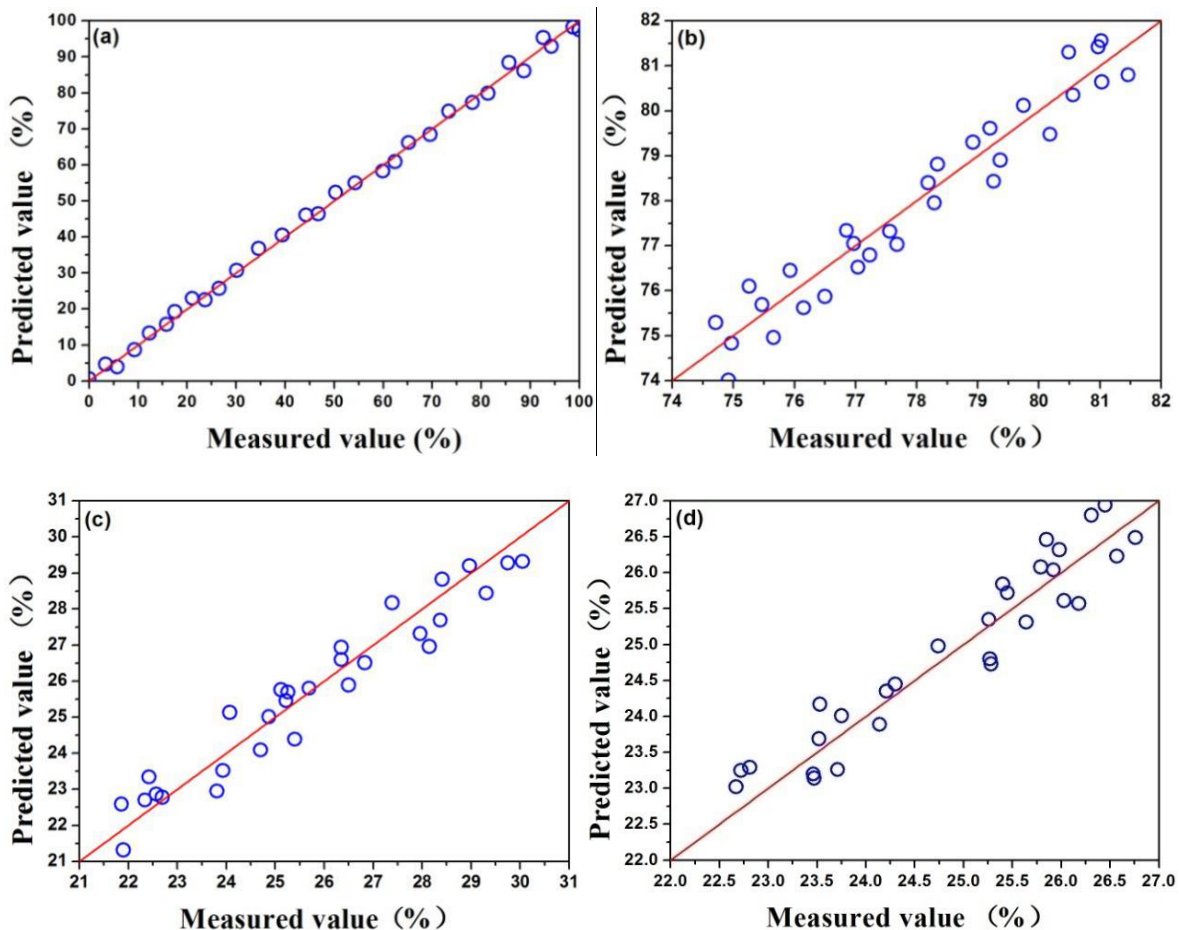
Fig. 3. The predictive performance of four calibration models. (a) *Eucalyptus globulus*, (b) holocellulose, (c) pentosan, (d) acid insoluble lignin

bioresources.com

Figure 3 contains scatter diagrams that display the predictive performance of the four models with measured values as the horizontal axis and the predicted values as the ordinate. The distribution of the points by the holocellulose model and the pentosans model on both sides of the 45° line is relatively uniform. The results showed that there is no significant deviation in the prediction of the holocellulose model and the pentosans model. The points by the *E. globulus* model and the acid insoluble lignin model are more on the left side of the 45° line, and the bias values are 0.16% and 0.08%, which make the predicted value of models on the high side.

Table 3. The Independent Verification of the LASSO Calibration Models

Content	Independent Verification			
	R^2_{val}	RMSEP	RPD	AD
Eucalyptus globulus	0.9975	1.59%	20.04	-2.58% to 2.73%
Holocellulose	0.9333	0.54%	3.87	-0.91% to 0.84%
Pentosan	0.9351	0.66%	3.90	-1.19% to 1.06%
Acid insoluble lignin	0.9037	0.40%	3.23	-0.61% to 0.64%

CONCLUSIONS

1. The least absolute shrinkage and selection operator (LASSO) algorithm and near infrared (NIR) spectroscopy technique were used to build *E. globulus*, holocellulose, pentosan, and acid insoluble lignin content models to be used for the estimation of the contents of mixtures of *E. globulus* and *A. mangium*. The distribution of *E. globulus* content and chemical composition content basically covered the range of possible values. The prediction error of the *E. globulus* content model was suitable for the rough prediction of the mixed condition of pulping raw materials. Holocellulose, pentosan, and acid insoluble lignin models had satisfactory performance and could be used for accurate analysis for small prediction error. There was no obvious deviation in the prediction of holocellulose and pentosan, and the models of *E. globulus* and acid insoluble lignin had bias error, which made the analysis result slightly higher than the measured value. The good adaptability and predictive ability of the models will help to realize the on-line analysis of *E. globulus* and *A. mangium* raw materials, thus improving the efficiency of pulping industry.

2. This study confirmed the feasibility of the LASSO algorithm for pulping raw materials analysis. At the same time, its modeling accuracy is slightly better than the traditional PLS algorithm. LASSO algorithm will provide more possibility when selecting the algorithm to establish accurate calibration models.

ACKNOWLEDGMENTS

bioresources.com

REFERENCES CITED

Alves, A., Simões, R., Santos, C., Potts, B., Rodrigues, J., and Schwanninger, M. (2012). "Determination of *Eucalyptus globulus* wood extractives content by near infrared-based partial least squares regression models: Comparison between extraction procedures," *Journal of Near Infrared Spectroscopy* 20(2), 275. DOI: 10.1255/jnirs.987

Balabin, R. M., Safieva, R. Z., and Lomakina, E. I. (2011). "Near-infrared (NIR) spectroscopy for motor oil classification: From discriminant analysis to support vector machines," *Microchemical Journal* 98(1), 121-128. DOI: 10.1016/j.microc.2010.12.007

Chen, C., Pu, J. W., Yao, S., and Jiang, Y. F. (2009). "Study on the pulping performances of fast growing wood-*Acacia auriculiformis* × *A. mangium*," *Paper Science & Technology* 28(4), 1-3. DOI: 10.3969/j.issn.1671-4571.2009.04.001

Cui, H., and Fang, G. G. (2015). "A method based on near infrared spectra for rapid identification of wood genuses and species," *Chemistry and Industry of Forest Products* 35(2), 169-171. DOI: 10.3969/j.issn.0253-2417.2015.06.016

Deng, Y. J., Fang, G. G., Han, S. M., Jiao, J., Li, H. B., and Liang, L. (2015). "Variation of chemi-mechanical pulping properties from several *Eucalyptus* spp. woodchips," *Chemistry and Industry of Forest Products* 35(1), 63-68. DOI: 10.3969/j.issn.0253-2417.2015.01.010

Downes, G. M., Harwood, C. E., Wiedemann, J., Ebdon, N., Bond, H., and Meder, R. (2012). "Radial variation in kraft pulp yield and cellulose content in *Eucalyptus globulus* wood across three contrasting sites predicted by near infrared spectroscopy," *Canadian Journal of Forest Research* 42(8), 1577-1586. DOI: 10.1139/x2012-083

Feng, Y. C., Zhang, Q., and Hu, C. Q. (2016). "Study on the selection of parameters for evaluating drug NIR universal," *Spectroscopy & Spectral Analysis* 36(8), 1-8. DOI: 10.3964/j.issn.1000-0593(2016)08-2447-08

GB/T 2677.10 (1995). "Fibrous raw material - Determination of holocellulose," Standardization Administration of China, Beijing, China.

GB/T 2677.9 (1994). "Fibrous raw material - Determination of pentosan," Standardization Administration of China, Beijing, China.

GB/T 2677.8 (1994). "Fibrous raw material - Determination of acid-insoluble lignin," Standardization Administration of China, Beijing, China.

He, W. M., and Hu, H. R. (2013). "Prediction of hot-water-soluble extractive pentosan and cellulose content of various wood species using FT-NIR spectroscopy," *Bioresource Technology* 140(7), 299-305. DOI: 10.1016/j.biortech.2013.04.115

Hodge, G. R., and Woodbridge, W. C. (2010). "Global near infrared models to predict lignin and cellulose content of pine wood," *Journal of Near Infrared Spectroscopy* 18(6), 367-380. DOI: 10.1255/jnirs.902

Huang, J., Xia, T., Li, A., Yu, B., Li, Q., and Tu, Y. (2012). "A rapid and consistent near infrared spectroscopic assay for biomass enzymatic digestibility upon various physical and chemical pretreatments in miscanthus," *Bioresource Technology* 121(10), 274. DOI: 10.1016/j.biortech.2012.06.015

Krasznai, D. J., Champagne, P., and Cunningham, M. F. (2012). "Quantitative characterization of lignocellulosic biomass using surrogate mixtures and multivariate techniques," *Bioresource Technology* 110(5), 652-661. DOI: 10.1016/j.biortech.2012.01.089

Labbé, N., Ye, X. P., Franklin, J. A., and Womac, A. R. (2008). "Analysis of switchgrass characteristics using near infrared spectroscopy," *BioResources* 3(4), 1329-1348. DOI: 10.15376/biores.3.4.1329-1348

Li, M., Wang, J., Du, F., Diallo, B., and Xie, G. H. (2017). "High-throughput analysis of chemical components and theoretical ethanol yield of dedicated bioenergy sorghum using dual-optimized partial least squares calibration models," *Biotechnology for Biofuels* 10(1), 206. DOI: 10.1186/s13068-017-0892-z

Li, Z. Z. (2001). "Forestry-paper integration and the pulp and paper making property of major man-made paper making trees in China," *China Pulp & Paper Industry* 22(7), 6-12.

Mora, C. R., and Schimleck, L. R. (2010). "Kernel regression methods for the prediction of wood properties of *Pinus taeda* using near infrared spectroscopy," *Wood Science and Technology* 44(4), 561-578. DOI: 10.1007/s00226-009-0299-5

Muhammad, A. J., Ong, S. S., and Ratnam, W. (2017). "Characterization of mean stem density, fibre length and lignin from two acacia species and their hybrid," *Journal of Forestry Research* (2), 1-7. DOI: 10.1007/s11676-017-0465-9

Qu, A. Y., and Fang, G. G. (2009). "Forecasting methods of pulping properties of fast-growing woods with multiple nonlinear regression," *Scientia Silvae Sinicae* 45(10), 113-119. DOI: 10.3321/j.issn:1001-7488.2009.10.019

Schwanninger, M., Rodrigues, J. C., and Fackler, K. (2011). "A review of band assignments in near infrared spectra of wood and wood components," *Journal of Near Infrared Spectroscopy* 19(5), 287-308. DOI: 10.1255/jnirs.955

Shi, W. F., Hu, X. G., and Yu, K. (2012). "K-part Lasso based on feature selection algorithm for high-dimensional data," *Computer Engineering and Applications* 48(1), 157-161. DOI: 10.3778/j.issn.1002-8331.2012.01.045

Stirling, R. (2013). "Near-infrared spectroscopy as a potential quality assurance tool for the wood preservation industry," *The Forestry Chronicle* 89(5), 654-658. DOI: 10.5558/tfc2013-117

Sun, J. (1997). "Statistical analysis of NIR data: Data pretreatment," *Journal of Chemometrics* 11(6), 525-532. DOI: 10.1002/(SICI)1099-128X(199711/12)11:63.0.CO;2-G

Talens, P., Mora, L., Morsy, N., and Sun, D. W. (2013). "Prediction of water and protein contents and quality classification of Spanish cooked ham using NIR hyperspectral imaging," *Journal of Food Engineering* 117(3), 272-280. DOI: 10.1016/j.jfoodeng.2013.03.014

Tibshirani, R. (1997). "The Lasso method for variable selection in the Cox model," *Statistics in Medicine* 16(4), 385-395. DOI: 10.1002/(SICI)1097-0258(19970228)16:4<385::AID-SIM380> 3.0.CO;2-3

Tibshirani, R. (2011). "Regression shrinkage and selection *via* the Lasso," *Journal of the Royal Statistical Society* 73(3), 273-282. DOI: 10.2307/2346178

Tsuchikawa, S., and Schwanninger, M. (2013). "A review of recent near-infrared research for wood and paper (Part 2)," *Applied Spectroscopy Reviews* 48(7), 560-587. DOI: 10.1080/05704928.2011.621079

Wang, Y. R., Fei, B. H., Fu, F., Jiang, Z. H., Qin, D. C., and Yang, Z. (2008). "A novel method for estimating wood fiber length using near infrared spectroscopy," *China Pulp & Paper* 27(6), 6-9. DOI: 10.3969/j.issn.0254-508X.2008.06.002

Xu, G. T., Yuan, H. F., and Lu, W. Z. (2000). "Development of modern near infrared spectroscopic techniques and its applications," *Spectroscopy and Spectral Analysis* 20(2), 134-142. DOI: 10.3321/j.issn:1000-0593.2000.02.003

Yao, S., Wu, G. F., Xing, M., Zhou, S. K., and Pu, J. W. (2010). "Determination of lignin content in *Acacia* spp. using near-infrared reflectance spectroscopy," *BioResources* 5(2), 556-562.

Yu, S. H., and Zhang, J. (2014). "The study on impact factors of foreign direct investment based on LASSO," *Journal of Hunan University (Social Sciences)* 28(2), 53-56. DOI: 10.3969/j.issn.1008-1763.2014.02.009

Zhao, R. J., Huo, X. M., Xing, X. T., Shangguan, W. W., and Ren, H. Q. (2012). "Comparison of measurement methods of *Eucalyptus pellita* air-dry density," *Journal of Northwest Forestry University* 27(2), 242-244. DOI: 10.3969/j.issn.1001-7461.2012.02.49

Zhong, Y., Kang, L., Zhang, M., Xin, D., and Zhang, J. (2016). "Rapid determination of chemical composition and classification of bamboo fractions using visible–near infrared spectroscopy coupled with multivariate data analysis," *Biotechnology for Biofuels* 9(1), 35. DOI: 10.1186/s13068-016-0443-z