

# Multi-species NIR calibration for estimating holocellulose in plantation timber

Vimal Kothiyal<sup>1</sup> · Jaideep<sup>2</sup> · Seema Bhandari<sup>2</sup> ·  
H. S. Ginwal<sup>2</sup> · Sachin Gupta<sup>2</sup>

Received: 18 November 2014 / Published online: 3 April 2015  
© Springer-Verlag Berlin Heidelberg 2015

**Abstract** Feasibility of near-infrared (NIR) spectroscopy for developing multi-species model for plantation timber was explored for estimation of holocellulose in un-extracted milled wood samples. Six commonly planted species of *Eucalyptus tereticornis*, *E. camaldulensis*, *E. grandis*, *Leucaena leucocephala*, *Dalbergia sissoo* and *Populus deltoides* from a wide range of locations and varying age groups were taken for the present study. Few samples of *E. hybrid* between *E. tereticornis* and *E. camaldulensis* were also included in the study to make the model useful for practical application. NIR models were evaluated using partial least squares regression (PLSR-1—full cross-validation, PLSR-2—cross-validation which leaves more than one out) and by dividing the samples into calibration and prediction (test) sets and interchanging them from calibration to prediction sets. The predictive ability of the model was assessed by calculating four ratios of multivariate statistics for individual species model and combined species models. A final combined model for all the species having component range of 76.14–63.03 % and standard deviation of 2.586 % was developed in the spectral range of 7502–4246  $\text{cm}^{-1}$  wave number using 1st derivative plus multiplicative scatter correction using factor of nine by removing samples with outliers found in all the PLSR-2 evaluation steps and in most of the models. The model remained stable even when 30 % of the samples were left out with no outlier detected.

---

✉ Vimal Kothiyal  
vimal\_kothiyal@yahoo.com; kothiyalv@icfre.org

<sup>1</sup> Research Planning Division, Directorate of Research, Indian Council of Forestry Research and Education, Dehradun, India

<sup>2</sup> Forest Research Institute, Indian Council of Forestry Research and Education, Dehradun, India

## Introduction

Assessment of the quality of the wood produced is extremely important to optimize utilization of raw material resource for economic gains and improve wood properties of plantation crop. Cost, time and simplicity of method adopted, and accuracy of results are important in addition to multiplicity of simultaneous analysis that can be done by noninvasive means. Near-infrared (NIR) spectroscopy has given such option fitting to all the above criteria for estimating many wood properties. Reviews by various researchers have highlighted the emergence of NIR spectroscopy along with its utility in the field of wood science, paper and forestry and development phases in which it is at present (Workman 1999; Pasikatan et al. 2001; Pasquini 2003; Barton 2004; So et al. 2004; Gong and Zhang 2008; Schimleck 2008; Tsuchikawa 2007; He et al. 2010; Yao and Pu 2009; Aenugu et al. 2011; Tsuchikawa and Schwanninger 2013). Sample preparation, recording of NIR spectrum and its resolution (Schimleck et al. 2004), pretreatment of NIR data (Rinnan et al. 2009) for removal of noise and unwanted signals, optimization of suitable wavelength range and the latent variables (factors or rank) and evaluation of NIR models are all important steps in the NIR calibration development. Being an indirect method, NIR spectroscopy needs calibration to be established between spectral dataset and its known data (property of interest) which is normally referred to as calibration dataset. The accuracy of such NIR measurements is influenced by the precision of the reference method used for calibration. The influences of spectral noise and reference data noise have been studied by various workers (DiFoggio 1995; Lu and McClure 1998; Geladi 2002; Rodrigues et al. 2006; Yao et al. 2010b, 2011). Some researchers have reviewed the band assignments in NIR spectra of wood and wood components (Schwanninger et al. 2011a). Bokobza (1998) in his review attempted to explain the basic concepts of near-infrared spectroscopy, and Tsuchikawa (1998) modeled the concept of applying NIR spectroscopy to biological samples having cellular structure. Researchers are also attempting to overcome the problem of anisotropy of woody material (Kothiyal et al. 2014) and changes in NIR spectra due to moisture content (Kothiyal and Raturi 2011; Fujimoto et al. 2012). Others have investigated the feasibility of multi-site and multi-species calibrations (Downes et al. 2010; Schimleck et al. 2010; Yao et al. 2010a) along with reducing efforts in sample preparation (Kothiyal et al. 2012). Improving the robustness and predictive ability of the model is a continuous effort, and some good results have been obtained by Inagaki et al. (2012) and Alves et al. (2012).

Lignocellulosic biomasses are physically and chemically heterogeneous. Optimal processing requires understanding of the material quality and properties. Characteristics of biomass differ as a function of many factors, including plant genetics, growth environment, downstream harvesting and processing methods adopted. Improving the wood industry productivity by supplying quality wood material in terms of wood density, lignin, cellulose content and other chemical, physical and mechanical properties is important for pulp and solid wood industry. Indian plantations are dominated by eucalyptus, poplar, sissoo, casuarina, acacia and melia to meet the demand of pulp and other wood-based industries. Tree improvement

programs have attempted to increase the yield through selections and by developing superior planting material. Stable quality in terms of chemical, physical, mechanical, anatomical and pulp properties is important for any wood-based industry to produce products of the desired uniform quality. Wood quality assessment by faster and cost-effective methods is being attempted in the recent past (Kothiyal and Raturi 2011; Kothiyal et al. 2012, 2014; Raturi et al. 2012). NIR spectroscopy is increasingly being used to replace traditional methods of wood property assessment and, as a result, multi-site, multi-species calibrations are of interest to research organizations that assess wood properties on a large scale. Such calibrations are common in agricultural applications (Berzaghi et al. 2002; Dardenne 2004), and some attempts have been made on wood-related aspects (Garbutt et al. 1992; Schimleck et al. 2001, 2003, 2010; Hodge and Woodbridge 2004, 2010; Yao and Pu 2009; Downes et al. 2010). Multi-site and multi-species calibrations are beneficial in the sense that they reduce the multiplicity of NIR models for individual species and sites to a single model and overall reduction in cost and time. Efforts in upgrading NIR calibration in future will also be reduced to a single model for a component (property of interest) and will make handling of such NIR calibration much easier.

This study focuses on developing multi-species NIR model (combined model) for estimation of holocellulose. Six plantation species, namely *Eucalyptus tereticornis*, *Eucalyptus grandis*, *Eucalyptus camaldulensis*, *Leucaena leucocephala* (subabul), *Dalbergia sissoo* (shisham) and *Populus deltoides* (poplar), were taken for the study. In India, *E. tereticornis* is often referred to as *E. hybrid* or Mysore Gum and gets mixed up with the hybrid of *E. tereticornis* and *E. camaldulensis*. Natural hybrids between *E. tereticornis* and *E. camaldulensis* are also quite common. Most of the plantations in India are dominated by Mysore Gum. Few known samples of hybrid of *E. tereticornis* and *E. camaldulensis* have therefore been included in the study to make the combined model more robust for practical applications. Estimation of holocellulose by NIR attempted by few researchers (Huang et al. 2007; Ding et al. 2009; Hou and Li 2010; Ishizuka et al. 2012) in the past is mostly for single species. The present study will be applicable to six species in addition to *E. hybrid*.

## Materials and methods

### Wood samples for NIR calibration development

For the development of NIR method for holocellulose estimation, wood samples of six species, namely *Eucalyptus tereticornis* (Mysore Gum), *E. grandis*, *E. camaldulensis* and *Leucaena leucocephala* (subabul), *Dalbergia sissoo* (shisham) and *Populus deltoides* (poplar) were taken for the present study. The samples varied largely in age and location. Most of the plantation of *E. tereticornis* in India is dominated by Mysore Gum and is often referred to as *E. hybrid* and thus gets mixed up with the hybrid of *E. tereticornis* and *E. camaldulensis*. Few known samples of *E. hybrid* (*E. tereticornis* × *E. camaldulensis*) were therefore also included in the

study to make the model suitable for use with *E. hybrid* samples. While developing the NIR models, the *E. hybrid* samples were grouped with *E. camaldulensis*. Details of the samples are given in Table 1. For samples of *L. leucocephala*, details are given in Kothiyal et al. (2012).

Discs of about 5–7 cm at breast height were cut from bottom log, chipped and then milled into dust in a Wiley mill fitted with mesh (mesh 18; 1000  $\mu\text{m}$ ). Milled sample fraction of 40–60 mesh screen (250–400  $\mu\text{m}$ ) was retained and collected for NIR study. Particle size of samples does have an effect on NIR absorption, and therefore, uniformity was maintained in the sample preparation. The chemical analysis was performed by three different operators independently in the same laboratory at different point of time.

Operator 1: *E. tereticornis*, *E. grandis*, *E. camaldulensis* and *E. hybrid* and *D. sissoo*

Operator 2: *Leucaena leucocephala*

Operator 3: *Populus deltoides*

## Chemical analysis

### Holocellulose estimation

Holocellulose of extractive-free milled wood samples (40–60 mesh; 250–400  $\mu\text{m}$ ) was determined in duplicate according to the method of Wise et al. (1946). Ash content and alcohol benzene extractives were determined for each milled sample as per procedure T 211M-58 and T 16m-OS-59, respectively. Test specimen (5.0 g O.D) was put in an Erlenmeyer flask containing distilled water (160 ml). Sodium chlorite (1.50 g) and glacial acetic acid (0.5 ml) were added and refluxed at 70–80  $^{\circ}\text{C}$  for 1 h. After 1 h, sodium chlorite (1.50 g) and glacial acetic acid (0.5 ml) were again added. The process was repeated four times till the material became white. White dust was then filtered in a  $G_2$  crucible washed with distilled water

**Table 1** Details of sampled trees for NIR model development for holocellulose estimation

Name of species	Number of trees/ sample	Diameter range (over bark) at breast height (cm)	Age range (years)	No. of locations
<i>Leucaena leucocephala</i>	78	11–107	4–20	7
<i>E. tereticornis</i>	66	39–100	5–15	3
<i>E. grandis</i>	20	39–57	8	1
<i>E. camaldulensis</i>	36	37–60	8	1
<i>E. hybrid</i>	03	40–60	8	1
<i>Dalbergia sissoo</i>	26	35–120	4–25	2
<i>Populus deltoides</i>	35	35–70	4–8	2
Total	264		4–25	–

(4 × 25 ml) followed by acetone (2 × 25 ml), dried to constant weight in an oven at 105 ± 2 °C. Holocellulose was calculated as follows:

$$\text{Holocellulose, \%} = \frac{(A - B) \times 100}{W} \quad (1)$$

*A*: weight of crucible after extraction. *B*: weight of crucible before extraction. *W*: O.D. weight of test specimen.

The estimate obtained by Eq. (1) was converted to un-extracted wood meal, and the same is reported in Table 2. The samples are also divided into test sets (TS) and calibration sets (CV) and referred to as CS1/TS1 and CS2/TS2 in Table 2. Detailed discussion of the two sets follows in the subsequent sections and in “[Result and discussion](#)”.

### NIR spectra measurement

FT-NIR spectrophotometer Bruker Optics, MPA with wavenumber between 4000 and 12,820 cm<sup>-1</sup> (780–2500 nm), working in the diffuse reflectance mode, resolution: 8 cm<sup>-1</sup> using zero filling of factor 2 fitted with integrating sphere with RT-Pbs (external, NEP 5 × 10<sup>-10</sup> WHZ<sup>-1/2</sup>) detector setting and sample area of 15-mm spot size was used for collecting the spectra of the present study.

Un-extracted wood milled samples were transferred into quartz vials. Vials were placed above the window of the integrating sphere. Spectra of un-extracted wood milled samples were recorded three times with thirty-two scans each and averaged into one spectrum. The average spectrum therefore consisted of a total of ninety-six scans.

### Calibration development

Research in science often involves easy to measure variables (factors) to predict the behavior of other variables (responses or properties of interest). When factors are few in number, not significantly collinear, and have a well-understood relationship to the responses, then multiple linear regressions (MLR) are a good way to turn data into information. In case of NIR spectroscopy, the researcher is faced with many variables (say 2000 frequencies—more than the number of observation samples) with ill-understood relationships. Partial least squares is a method for constructing predictive models when the factors are many and highly collinear to extract only a few underlying factors or latent factors that account for most of the variation in the responses (Yeniav and Goktas 2002; Haenlein and Kaplan 2004). Partial least square regression (PLSR) is applied to describe the relationship (Geladi and Kowalski 1986; Hauksson et al. 2001) between dependent variables (wood holocellulose values in the present study) and near-infrared spectra (independent variable). The analysis was performed with the OPUS 6.5 software (Bruker “Quant 2”). Using full cross-validation techniques with the random subset method provided by this software package, the optimum rank or factors were found and validation model for holocellulose was developed. The model selection criteria were based on

**Table 2** Summary of the holocellulose content (on un-extracted milled wood basis) for calibration and prediction sets

Species name	Validation procedure	Total samples collected	No. of samples used for NIR	Min. (%)	Max. (%)	Avg. (%)	SD (%)
<i>Eucalyptus tereticornis</i>	Cross	66	63	66.03	72.92	68.52	2.195
Calibration/test set	CV1/TS1	–	33	63.03	71.57	68.50	2.189
Calibration/test set	CV2/TS2	–	30	63.44	72.92	68.55	2.238
<i>Eucalyptus hybrid</i>		3	3	70.26	73.96	72.10	0.933
<i>Eucalyptus grandis</i>	Cross	20	20	72.84	76.14	74.54	0.933
<i>Eucalyptus camaldulensis</i>	Cross	39	32	67.48	75.63	71.11	1.733
<i>Leucaena leucocephala</i>	Cross	78	67	65.23	74.76	70.59	2.067
Calibration/test set	CV1/TS1	–	34	65.23	74.76	70.91	2.144
Calibration/test set	CV2/TS2	–	33	66.62	73.12	70.27	1.963
<i>Dalbergia sissoo</i>	Cross	26	22	67.29	74.97	70.99	2.102
<i>Populus deltoides</i>	Cross	35	32	71.03	74.96	73.08	1.070
All species	Total	264	236	63.03	76.14	70.82	2.608
Calibration/test set	CV1/TS1	–	118	63.03	76.14	70.93	2.624
Calibration/test set	CV2/TS2	–	118	63.44	76.02	70.70	2.598
All species—(poplar)	Cross	–	204	63.03	76.14	70.46	2.601
Calibration/test set	CV1/TS1	–	101	63.03	76.14	70.62	2.685
Calibration/test set	CV2/TS2	–	103	63.44	76.02	70.31	2.519
All species—(poplar, shisham)	Cross	–	182	63.03	76.14	70.40	2.653
Calibration/test set	CV1/TS1	–	91	63.03	76.14	70.64	2.773
Calibration/test set	CS2/TS2	–	91	63.44	76.02	70.16	2.520
All Eucalyptus species	Cross	–	115	63.03	76.14	70.29	2.944
Calibration/test set	CV1/TS1	–	57	63.03	76.14	70.48	3.095
Calibration/test set	CV2/TS2	–	58	63.44	76.02	70.10	2.802
All—(poplar, subabul)	Cross	–	137	63.03	76.14	70.40	2.831
All—(poplar, subabul, <i>E. grandis</i> )	Cross	–	117	63.03	75.63	69.69	2.406

Min.: minimum; Max.: maximum; SD: standard deviation; Avg.: Average; CV1/CV2: samples divided into two halves, CV2 used as TS2 for test validation with CV1, and CV1 used as TS1 for test validation with CV2

maximizing the  $r_p^2$  (coefficient of determination of the validation set) and minimizing the root-mean-square error of the prediction (RMSEP) set as well as the minimum of latent variables (factors) used for developing the model. In order to improve signal quality, different spectral pretreatment methods were applied.

## Wavelength selection and spectral pretreatment (preprocessing)

Opus Quant 2 software was used for data preprocessing and for the calculation and validation of the PLSR models. The wavenumber ranges, and the preprocessing of NIR spectra were determined from several combinations to give high coefficients of determination. Several preprocessing methods such as no preprocessing, offset correction, multiplicative scatter correction (MSC), vector normalization, derivatives, straight-line correction and combinations of them were applied along with smoothing. The spectral pretreatments giving the best models are discussed here. Spectral range  $7502\text{--}4246\text{ cm}^{-1}$  was found to be most appropriate as almost all treatments and wavenumber ranges fell within this range. Two spectral ranges ( $7502\text{--}6098 + 5450\text{--}4246$  and  $7502\text{--}4246\text{ cm}^{-1}$ ) with 1st derivative (9-point smoothing) plus MSC were found to be the most suitable for the present study, although few good results were also obtained with 1st derivative plus vector normalization (SNV) and reduced spectral range within  $7502\text{--}4246\text{ cm}^{-1}$ . Results of the models developed with both the spectral ranges are reported in Table 3 (spectral range  $7502\text{--}6098 + 5450\text{--}4246\text{ cm}^{-1}$ ) and Table 4 ( $7502\text{--}4246\text{ cm}^{-1}$ ). Some good results of SNV and reduced spectral range are also discussed in the appropriate section of “[Result and discussion](#)”.

## Optimization of rank (latent variables)

The optimum number of PLS components (rank or latent variables or factors) are determined by full inner cross-validation method (leave one out). The Quant 2 software package calculates optimum factors (rank) separately during cross-validation and test set validation (prediction) process, and differences were observed between the two. Tables 3 and 4 therefore reports the latent variables (factors) obtained through Quant 2 software optimization during both the processes. This feature is unique to Quant 2 software, and therefore, the models developed with software automatic optimization selection process of factors are described in detail for integrating sphere in “[Results and discussion](#)” section. Researchers have discussed differences in the latent variable in detail and have concluded that a difference of two or less is acceptable.

## Identification of outliers

The method of identification of outliers by Quant 2 is based on Mahalanobis distance considered as a measure of the similarity between the spectra. Schwanninger et al. (2011b, c) have discussed it in detail, and any spectra lying outside the set limit (limit =  $X \times R/M$ , where  $X$  = factor normally taken as 2,  $R$  = rank-latent variables, and  $M$  = number of samples) will be identified as outlier. A factor of two is normally used for calculation of the Mahalanobis distance, and the same was applied in the present study. Schwanninger et al. (2011b, c), however, described that the factor of two is too restrictive during the prediction of unknown samples. Some other workers have also discussed the detection of outliers in detail (Jouan-Rimbaud et al. 1999).

**Table 3** Summary of the NIR models developed for individual species and combination of species in spectral range 7502–6098 + 5450–4246 cm<sup>-1</sup> using 1st derivative (9-point smoothing) plus multiplicative scatter correction (MSC)

Samples	CV/test validation	No. of samples	Component range (cross/test), %	RMSECV (%)	$r^2_{cv}$	RMSEP (%)	$r^2_p$	No. of factors	Outliers	RPD (cross/test)	RER	RMSEP/RMSECV	SD/RMSECV	RMSEECV/RMSEC
All species	CV	236	63.03–76.14	0.848	0.89	–	–	9	0	3.07	–	–	3.08	1.14
All species—(poplar)	CV	204	63.03–76.14	0.864	0.89	–	–	8	0	3	–	–	3.01	1.18
All—(poplar, subabul, <i>E. grandis</i> )	CV	117	63.03–75.63	0.815	0.88	–	–	8	0	2.94	–	–	2.95	1.27
All species—(shisham, poplar)	CV	182	63.03–76.14	0.851	0.89	–	–	9	1	3.11	–	–	3.12	1.21
All species—(poplar, subabul)	CV	137	63.03–76.14	0.838	0.91	–	–	8	2	3.36	–	–	3.38	1.27
All eucalyptus species	CV	115	63.03–76.14	0.841	0.92	–	–	8	1	3.49	–	–	3.50	1.38
<i>Eucalyptus tereticornis</i>	CV	63	63.03–72.92	0.798	0.86	–	–	6	1	2.73	–	–	2.75	1.37
<i>E. camaldulensis</i>	CV	32	67.48–75.63	0.669	0.84	–	–	6	0	2.55	–	–	2.59	1.77
<i>E. grandis</i>	CV	20	72.84–76.14	0.635	0.51	–	–	4	0	1.43	–	–	1.47	2.08
<i>Leucaena leucocephala</i>	CV	67	65.23–74.76	0.91	0.80	–	–	4	0	2.25	–	–	2.27	1.18
<i>Dalbergia sissoo</i>	CV	22	67.29–74.97	0.641	0.90	–	–	4	0	3.2	–	–	3.28	1.71
<i>Populus deltoides</i>	CV	32	71.03–74.96	0.83	0.38	–	–	4	0	1.27	–	–	1.29	1.37



Table 3 continued

Samples	CV/test validation	No. of samples	Component range (cross/test), %	RMSECV (%)	$r_{cv}^2$	RMSEP (%)	$r_p^2$	No. of factors	Outliers	RPD (cross/test)	RER	RMSEP/RMSECV	SD/RMSECV	RMSECV/RMSEC
All species	CV1/TS1	118/118	63.03–76.14/ 63.44–76.02	0.938	0.87	0.876	0.88	6/7	0	2.79/2.96	14.36	0.93	2.80	1.14
	CV2/TS2	118/118	63.44–76.02/ 63.03–76.14	0.897	0.88	0.866	0.89	9/7	0	2.88/3.02	15.14	0.97	2.90	1.38
All species— (poplar)	CV1/TS1	101/103	63.03–76.14/ 63.44–76.02	0.964	0.87	0.899	0.87	7/8	0	2.77/2.79	13.99	0.93	2.79	1.23
	CV2/TS2	103/101	63.44–76.02/ 63.03–76.14	0.899	0.87	0.873	0.89	9/9	2/0	2.79/3.08	15.02	0.97	2.80	1.12
All species— (shisham, poplar)	CV1/TS1	91/91	63.03–76.14/ 63.44–76.02	0.991	0.87	0.893	0.87	7/8	0/	2.78/2.81	14.09	0.90	2.80	1.12
	CV2/TS2	91/91	63.44–76.02/ 63.03–76.14	0.962	0.85	0.876	0.90	6/9	1/0	2.61/3.16	14.97	0.91	2.62	1.13
All eucalyptus species	CV1/TS1	57/58	63.03–76.14/ 63.44–76.02	0.962	0.90	0.901	0.89	7/7	1/1	3.19/3.09	13.96	0.94	3.22	1.53
	CV2/TS2	58/57	63.44–76.02/ 63.03–76.14	0.871	0.90	0.822	0.93	11/8	1/1	3.19/3.74	15.95	0.94	3.22	3.69
<i>Eucalyptus tereticornis</i>	CV1/TS1	33/30	63.03–71.57/ 63.44–72.92	1.11	0.73	0.703	0.90	4/6	1/1	1.95/3.13	13.49	0.63	1.97	1.32
	CV2/TS2	30/33	63.44–72.92/ 63.03–71.57	0.886	0.84	0.724	0.89	6/6	1/0	2.48/3.02	11.80	0.82	2.53	2.19
<i>Leucaena leucocephala</i>	CV1/TS1	34/33	65.23–74.76/ 66.62–73.12	1.14	0.71	1.02	0.72	3/4	1/1	1.85/1.92	6.37	0.89	1.88	1.26
	CV2/TS2	33/34	66.62–73.12/ 65.23–74.76	0.843	0.81	0.948	0.80	4/5	0	2.3/2.26	10.05	1.12	2.33	1.27

Preprocessing: 1st derivative (9-point smoothing) plus MSC; spectral range: 7502–6098+ 5450–4246  $\text{cm}^{-1}$ ; mean centering, zero filling of factor 2

CV: cross-validation; CV1/CV2; samples divided into two halves, CV2 used as TS2 for test validation with CV1 and CV1 used as TS1 for test validation with CV2

**Table 4** Summary of the NIR models developed for individual species and combination of species in spectral range 7502–4246 cm<sup>-1</sup> using 1st derivative (9-point smoothing) plus multiplicative scatter correction (MSC)

Samples	CV/Test validation	No. of samples	Component range (cross/test), %	RMSECV (%)	$r^2_{cv}$	RMSEP (%)	$r^2_p$	No. of factors	Outliers	RPD (cross/test)	RER	RMSEP/RMSECV	SD/RMSECV	RMSEECV/RMSEC
All species	CV	264/236	63.03–76.14	0.846	0.89	–	–	9	0	3.07	–	–	3.08	1.15
All species—(poplar)	CV	229/204	63.03–76.14	0.826	0.90	–	–	9	1	3.14	–	–	3.14	1.19
All—(poplar, subabul, <i>E. grandis</i> )	CV	131/117	63.03–75.63	0.815	0.88	–	–	8	0	2.94	–	–	2.95	1.37
All species—(shisham, poplar)	CV	203/182	63.03–76.14	0.826	0.90	–	–	9	1	3.2	–	–	3.21	1.20
All species—(poplar, subabul)	CV	151/137	63.03–76.14	0.773	0.92	–	–	9	2	3.65	–	–	3.66	1.41
All eucalyptus species	CV	125/115	63.03–76.14	0.774	0.93	–	–	8	1	3.79	–	–	3.80	1.38
<i>Eucalyptus tereticornis</i>	CV	66/63	63.03–72.92	0.798	0.86	–	–	6	1	2.73	–	–	2.75	1.33
<i>E. camaldulensis</i>	CV	39/32	67.48–75.63	0.774	0.79	–	–	5	0	2.21	–	–	2.24	1.55
<i>E. grandis</i>	CV	20/20	72.84–76.14	0.59	0.58	–	–	4	0	1.56	–	–	1.58	2.13
<i>Leucaena leucocephala</i>	CV	78/67	65.23–74.76	0.932	0.79	–	–	4	0	2.2	–	–	2.22	1.19
<i>Dalbergia sissoo</i>	CV	26/22	67.29–74.97	0.627	0.90	–	–	4	0	3.27	–	–	3.35	1.81
<i>Populus deltoides</i>	CV	35/32	71.03–74.96	0.82	0.39	–	–	4	0	1.28	–	–	1.30	1.38

**Table 4** continued

Samples	CV/Test validation	No. of samples	Component range (cross/test), %	RMSECV (%)	$r^2_{cv}$	RMSEP (%)	$r^2_p$	No. of factors	Outliers	RPD (cross/test)	RER	RMSEP/RMSECV	SD/RMSECV	RMSEECV/RMSEC
All species	CV1/TS1	118/118	63.03–76.14/ 63.44–76.02	0.957	0.87	0.861	0.89	7/8	0/0	2.73/3.01	14.6	0.89	2.74	1.16
	CV2/TS2	118/118	63.44–76.02/ 63.03–76.14	0.889	0.88	0.817	0.90	9/9	2/0	2.91/3.2	14.74	0.92	2.92	1.42
All species— (poplar)	CV1/TS1	101/103	63.03–76.14/ 63.44–76.02	0.937	0.88	0.887	0.87	7	1/0	2.85/2.83	14.18	0.94	2.86	1.20
	CV2/TS2	103/101	63.44–76.02/ 63.03–76.14	0.911	0.87	0.832	0.90	8/9	1	2.75/3.21	15.75	0.91	2.76	1.37
All species— (shisham, poplar)	CV1/TS1	91/91	63.03–76.14/ 63.44–76.02	0.968	0.88	0.864	0.88	7/8	1/0	2.85/2.91	14.56	0.89	2.86	1.20
	CV2/TS2	91/91	63.44–76.02/ 63.03–76.14	0.94	0.86	0.872	0.90	7/9	0/0	2.67/3.17	15.03	0.93	2.68	1.19
All eucalyptus species	CV1/TS1	57/58	63.03–76.14/ 63.44–76.02	0.864	0.92	0.881	0.90	7/7	1/1	3.55/3.16	14.28	1.02	3.58	1.52
	CV2/TS2	58/57	63.44–76.02/ 63.03–76.14	0.889	0.90	0.735	0.94	9/8	2/0	3.13/4.17	17.83	0.83	3.15	2.07
<i>Eucalyptus terreicornis</i>	CV1/TS1	33/30	63.03–71.57/ 63.44–72.92	1.08	0.75	0.644	0.91	4/6	1/1	1.99/3.43	14.72	0.60	2.03	1.33
	CV2/TS2	30/33	63.44–72.92/ 63.03–71.57	0.853	0.85	0.778	0.87	6/5	1/0	2.58/2.78	10.97	0.91	2.62	2.00
<i>Leucaena leucocephala</i>	CV1/TS1	34/33	65.23–74.76/ 66.62–73.1	1.03	0.76	1.05	0.71	4/3	1/0	2.06/1.85	6.19	1.02	2.08	1.34
	CV2/TS2	33/34	66.62–73.12/ 65.23–74.76	0.856	0.80	0.963	0.79	4/5	0/0	2.26/2.21	9.89	1.13	2.29	1.28

Preprocessing: 1st derivative (9-point smoothing) plus MSC; spectral range: 7502–4246  $\text{cm}^{-1}$ ; mean centering, zero filling of factor 2

*Leucaena leucocephala*: subabul; *Dalbergia sissoo* (shisham); *Populus deltoides*; poplar

CV: cross-validation; CV1/CV2; samples divided into two halves, CV2 used as TS2 for test validation with CV1 and CV1 used as TS1 for test validation with CV2

## Evaluation of models

Cross-validation with one sample omitted (leave out one) also referred to as PLSR-1 was performed for the validation of models. It is stated (Alves et al. 2012) that when the number of samples used for calibration is high or very high, the results obtained by full cross-validation are too “optimistic”. A large number of cancellation groups correspond to validation with a small perturbation of the statistical sample, whereas a small number of cancellation groups correspond to a heavy perturbation. Using a large number of samples combined with the leave out one method for cross-validation, the perturbation of the model at each step will be small. This procedure tends to “over fit” the model and often leads to too optimistic modeling results. The higher the number of samples, the more optimistic is the result obtained by full cross-validation, especially when the number of samples is very high, the results are too optimistic. To confirm the results, normally two procedures are applied: (a) to use a test set and (b) to increase the number of samples left out (referred to as PLSR-2) during cross-validation. The latter additionally proves the stability of a model. The progression of  $r_{cv}^2$ ,  $r_p^2$ , RMSECV (root-mean-square error of cross-validation) and RPD (ratio of performance to deviation) is evaluated. The present study reports the models for the two spectral ranges (7502–6098 + 5450–4246 and 7502–4246  $\text{cm}^{-1}$  with 1st derivative and 9-point smoothing using MSC preprocessing) in Tables 3 and 4 as discussed in the section “Wavelength selection and spectral pretreatment”. The models are developed for individual species and for combination of species (as given in Tables 2, 3 and 4). The procedures suggested for evaluation of models in the present section have been used selectively depending on the number of samples used in each model and its overall importance in the present study.

## The predictive ability of the model

The predictive ability of the model is also assessed and compared by calculating the ratio of the standard deviation of the reference data to standard error of performance (SEP) and is referred to as RPD which was introduced by Williams and Norris (2004). From the analytical point of view, in accordance with AACC method 39-00 (AACC 1999), the RPD should be in the following range:  $\geq 2.5$  screening in the breeding program;  $\geq 5$  acceptable for quality control;  $\geq 8$  good for process control, development and applied research (AACC 1999). Schimleck et al. (2005) reported in their study that RPD of  $> 1.5$  is good enough for preliminary screening. However, RPD is only correct and comparable when the data are normally distributed and the range in the validation set is not artificially increased. This method and the defined limits were developed for the analysis of grains; they are being used as rough rule of thumb until limits of wood are available. While analyzing the RPD results, one needs to specifically mention the nature of SEP calculated (with or without bias and skew correction). The RPDs in the present study have not been corrected for bias (Tables 3 and 4).

Range error ratio (RER = the ratio of the range in validation reference data to the SEP—standard error of performance) was introduced by Starr et al. (1981). The RER should be  $\geq 4$  for screening calibration,  $\geq 10$  to be acceptable for calibration for quality control and  $\geq 15$  for calibration for quantification. However, RER is also not immune to artificially extending the range (Fearn 2002) of the validation and thus manipulating RER. Tables 3 and 4 also report RER obtained for different models in the present study.

The ratio of SEP/SEC or RMSEP/RMSECV is also used for evaluation of the models. The ratio should be smaller than 1.2. Other criteria generally used to assess the quality of a model (Derkyi et al. 2011) include the root mean square of calibration (RMSEC), root-mean-square error of cross-validation (RMSECV) and the correlation coefficient ( $r^2$ ). A good model should have a low RMSEC, a low RMSECV and a high correlation coefficient ( $r^2$ ) between the predicted and measured values, but a small difference between RMSEC and RMSECV. The correlation coefficient  $r^2$ , the most commonly used, should be higher than 0.8 for quantitative predictions. For excellent models, the RMSEC-to-SD ratio should be  $\leq 0.2$ , where SD is the standard deviation of the reference values. If  $0.2 < \text{RMSEC-to-SD ratio} \leq 0.5$ , quantitative prediction is possible. The SD-to-RMSECV ratio should be  $\geq 2$ , RMSECV-to-RMSEC  $\leq 1.2$  and the SD-to-RMSECV ratio should be  $\geq 2.5$ . Tables 3 and 4 also report on the results of these parameters.

## Results and discussions

### Samples information

Samples from three species of Eucalyptus (*E. tereticornis*, *E. grandis* and *E. camaldulensis*), few samples of *E. hybrid* (*E. tereticornis*  $\times$  *E. camaldulensis*), along with that of *L. leucocephala* (subabul), *D. sissoo* (shisham) and *P. deltoides* (poplar) were taken for the present study. One sample per tree was taken for model development. All 264 samples, as detailed in Table 1, were collected from different locations. Particle size of all the samples was kept constant at 250–400  $\mu\text{m}$  as it is reported that NIR diffuse reflectance measurement is sensitive to light scattering and powder absorption by particle size (Pasikatan et al. 2001; Yeh et al. 2004; Hein et al. 2010). NIR spectra were recorded from un-extracted wood meal. After careful observation of the data from chemical analysis (holocellulose content), twenty-eight samples having inconsistent results in duplicate analysis were rejected. The results of the remaining 236 samples are given in Table 2 per species along with minimum, maximum, standard deviation and total sample number. The overall range of holocellulose for all the species is 63.03–76.14 % (lowest belongs to *E. tereticornis* and highest to *E. grandis*). Samples selected are from a wide range of sites and have wide variation in age as detailed in Table 1. Two-pronged approach was applied to evaluate the models. In the first approach, models were developed for individual species and combination of species using full cross-validation (PLSR-1: leave one out). Leave more out (PLSR-2) was also used for selected species and combinations for evaluating the models.

For selected models where number of samples was sufficient, the models were also evaluated by dividing the samples randomly (in the ratio of 50:50) into calibration and prediction sets. Care was taken that the same samples are selected in calibration and prediction sets for all the models. The statistics of samples randomly divided for calibrations (CV) and test validations (TS) is also reported in Table 2. The sets (groups) were interchanged from cross-validation to test validation and vice versa and are referred to as CV1/TS2 (calibration set/test set) and CV2/TS1 (groups interchanged).

## Spectral information

NIR region of the vibration spectroscopy measures the interaction between light and the relevant material. It is based on selective absorption of light by chemical compounds and determined by vibration of chemical bond specific to sample constituents. NIR region  $4000\text{ cm}^{-1}$ – $128,020\text{ cm}^{-1}$  is characterized by highly overlapping absorbance ( $4000$ – $5000\text{ cm}^{-1}$ ) with low noise (combination band that is difficult to chemically analyze), and first and second harmonic region ( $5000$ – $9000\text{ cm}^{-1}$ ) is informative region with low noise that can be chemically analyzed and the third harmonic region of  $9000$ – $125,000\text{ cm}^{-1}$  has high noise/low absorption and results in poor information quality. Most of the studies use first and second harmonic regions for qualitative and quantitative information (Schwaninger et al. 2011a), although some studies have effectively used third harmonic region for quantitative information (Kothiyal et al. 2014). NIR spectroscopy is useful because all properties are somehow influenced by chemical constitution of wood which is reflected in NIR absorption. The spectral range selection is the better fit of data to the calibration model. The present study therefore used the region of  $7502$ – $4246\text{ cm}^{-1}$  for the development of the models.

## PLS calibrations (full cross-validation)

In the first step, near-infrared spectral dataset of samples of all species (236 samples) selected for the present study was regressed against holocellulose content using inner full cross-validation (leaving one sample out) using 1st derivative plus MSC preprocessing. Two combinations of spectral ranges, namely  $7502$ – $6098 + 5450$ – $4246\text{ cm}^{-1}$  (Table 3) and  $7502$ – $4246\text{ cm}^{-1}$  (Table 4) gave the best regression results with maximum regression coefficient with minimum RMSECV. The procedure was repeated for all species taken independently and in combinations (all eucalyptus species; all species barring poplar; all species barring sissou and poplar; all species barring poplar and subabul; all species barring poplar, subabul and *E. grandis*). Tables 3 and 4 report the model statistics of all the models. Models of individual species were constructed with four to six factors. Except for *E. grandis* and *P. deltoides*, all models have RPD of more than 2.25 with *D. sissou* (component range 67.29–74.97 %) giving the best results of 3.2/3.27. *E. tereticornis* (component range 63.03–72.92 %) gave RPD of 2.73/2.73 with six factors and one outlier which when removed gave the RPD of 2.84/3.01 with 1/0 outliers, respectively. Models of subabul (component range 65.23–74.76 %) with the two spectral ranges gave an RPD of 2.2

and 2.25. Individual models of *E. grandis* and *P. deltooides* were not found suitable even for preliminary screening. Component range for both of these species is very narrow (*E. grandis*: 72.84–76.14 %; *P. deltooides*: 71.03–74.96 %) which could be one of the reasons. Sample size for *E. grandis* was also very small (20). However, RMSECV for all the models with full cross-validation was <1 %. Slight improvement in the models of *E. grandis*, *P. deltooides* and subabul was observed with reduced spectral range within 7502–4246  $\text{cm}^{-1}$  but was of not much significance. From the above results, it is also concluded that different operators used in the study did not affect the overall results. The models constructed with 7502–4246  $\text{cm}^{-1}$  spectral range in general had better overall statistics for all the models.

Following the trend of models for individual species, a model was constructed with all eucalyptus species (component range 63.03–76.14 %) taken together for both the spectral ranges (Tables 3 and 4). Remarkable improvement was obtained for both the spectral ranges as evident from Tables 3 and 4 with  $r_{\text{cv}}^2$  of determination for cross-validation of 0.92/0.93, RMSECV of 0.841/0.774 %, RPD of 3.49/3.79, respectively, was achieved using eight factors with one outlier common in both the models belonging to *E. tereticornis*. This is interesting as no suitable model could be developed with *E. grandis* alone. Removal of the outlier although improved the models statistics, resulting in more outliers. The models were extended by adding the samples of *D. sissoo* (component range 63.03–76.14 %), and model statistics was with  $r_{\text{cv}}^2$ —0.91/0.92, RMSECV—0.838/0.773 %, RPD 3.36/3.65 using 8/9 factors obtained with two outliers from *E. tereticornis* (one is common in both models). Removal of the outliers from both the models improved the models statistics ( $r_{\text{cv}}^2$ —0.92/0.93, RMSECV—0.783/0.738 %, RPD—3.6/3.81, factors—9/9, outlier—0/2). The removal of *E. grandis* from this combination did not improve the model (not reported in Tables 3, 4). The model was therefore further extended by adding samples from subabul, and the model statistics slightly decreased ( $r_{\text{cv}}^2$ —0.89/0.90, RMSECV—0.864/0.826 %, factors—8/9, RPD—3/3.14, outliers—0/1 belonging to *E. tereticornis* and same as before). Removal of outlier marginally improved RPD from 3.14 to 3.18. Another model was constructed using all the eucalyptus species and subabul with almost similar model statistics ( $r_{\text{cv}}^2$ —0.89/0.90, RMSECV—0.851/0.826 %, factors—9/9, RPD—3.11/3.2, outlier—1/1 same in both the models). Removing the outliers detected during the process improved the model statistics ( $r_{\text{cv}}^2$ —0.90/0.91, RMSECV—0.838/0.808 %, factors—9/9, RPD—3.15/3.27, outlier—nil). *P. deltooides* samples were added to this combination giving model statistics of  $r_{\text{cv}}^2$ —0.89/0.89, RMSECV—0.848/0.846 %, factors—9/9, RPD—3.07/3.07 with no outliers. The six samples detected as outliers while constructing individual species and combination of species models were no outliers in this model. However, removing the three common (*E. tereticornis*—2; *E. grandis*—1) outliers detected in the earlier process only marginally improved the model ( $r_{\text{cv}}^2$ —0.90/0.90, RMSECV—0.806/0.783 %, factors—9/10, RPD—3.18/3.27). Individual models of *E. grandis* (RPD—1.43/1.56), *P. deltooides* (RPD—1.27/1.28) and subabul (RPD—2.25/2.2) gave the lowest RPD. When a combined model was developed with these three species using spectral range of 7502–4246  $\text{cm}^{-1}$  with 1st derivative plus MSC, the model statistics have improved ( $r_{\text{cv}}^2$ —0.84, RMSECV—0.933 %, RPD—2.47, factors—4). The best

model with these three species was achieved with spectral range 7502–5446  $\text{cm}^{-1}$  using 1st derivative plus MSC ( $r_{\text{cv}}^2$ —0.86, RMSECV—0.848 %, RPD—2.72, factors—6, outlier—1). Removal of outlier, however, did not improve model statistics except for decreasing the rank by a factor of one. A combined model of *E. grandis* and *P. deltoides* in spectral range 7502–4597  $\text{cm}^{-1}$  with straight-line subtraction gave better model statistics ( $r_{\text{cv}}^2$ —0.67, RMSECV—0.702 %, RPD—1.75, factors—6, outlier—nil) compared to their individual models.

## Outliers

In all the models discussed above, six samples each were found to be outliers in both the spectral range of 7502–6098  $\text{cm}^{-1}$  + 5450–4246  $\text{cm}^{-1}$  and 7502–4246  $\text{cm}^{-1}$ . Five samples (*E. tereticornis*—2; *E. grandis*—2, subabul—1) were common in both the spectral ranges. One additional sample belonging to subabul was found to be outlier in spectral range 7502–6098 + 5450–4246  $\text{cm}^{-1}$ , and one sample of poplar was found to be outlier in spectral range 7502–4246  $\text{cm}^{-1}$ . Outlier samples belonging to *E. tereticornis* were of low holocellulose content (between 63 and 67 %), that of *E. grandis* was of high cellulose content (between 72 and 74 %), subabul (between 72 and 75 %) and poplar (between 70 and 71 %).

## Analysis of $r_{\text{cv}}^2$ , RMSECV and rank (factors)

Coefficient of determination and RMSECV are plotted against rank (factors) obtained for *E. tereticornis*, subabul, shisham and *E. camaldulensis* by the process of full cross-validation for the two spectral ranges and are given in Figs. 1a–d and 2a–d. It is observed that first four factors (rank) explain 80 % of the variation and have <1 % of RMSECV. It can be concluded that first four PLS factors are important for construction of good NIR models. Similar trends are observed for *E. grandis* and *P. deltoides* (graphs not plotted here) although coefficient of determination achieved was much lower as evident from Tables 3 and 4.

Progression of  $r_{\text{cv}}^2$ , RMSECV and rank are also plotted for combination models and given in Figs. 3a–d and 4a–d. Similar trends are observed for  $r_{\text{cv}}^2$ , whereas RMSECV <1 % is achieved only with six factors.

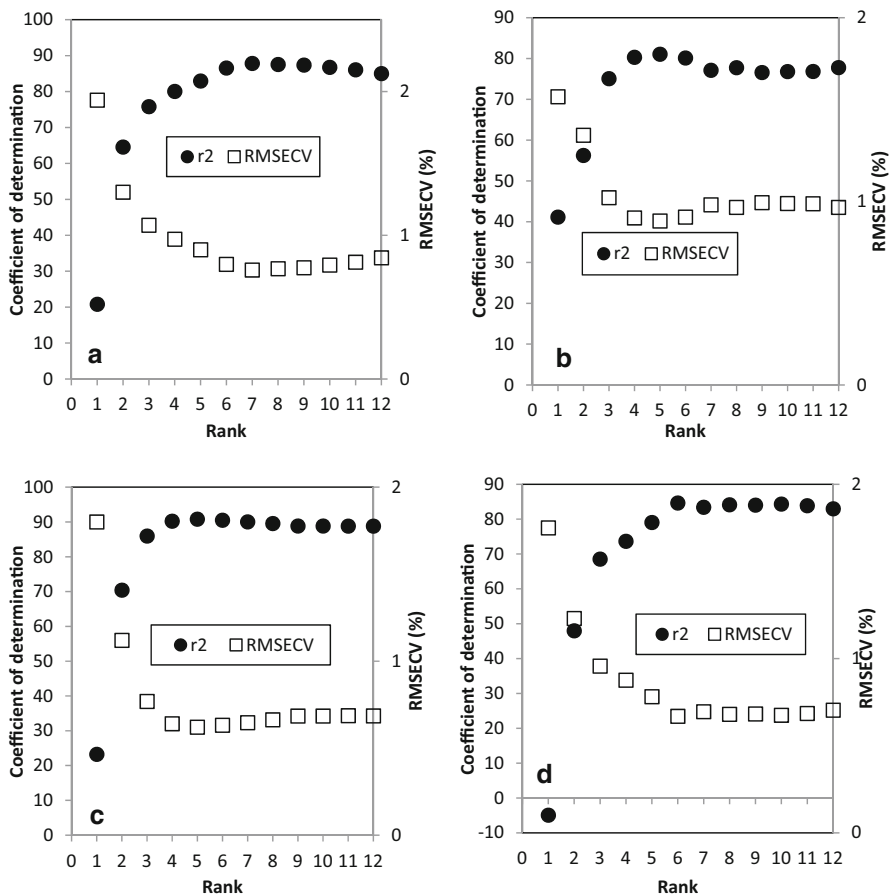
## PLS calibrations (cross and test validation)

The robustness of the selected models developed in the previous section using full cross-validation was evaluated by randomly dividing the samples into two sets (in the ratio of 50:50) for cross-validation (CV) and test validation (TS) and designated as CV1 and TS2. The groups were then interchanged from cross-validation to test validation and vice versa and designated as CV2 (earlier TS2) and TS1 (earlier CV1) by following methodology adopted by Schwanninger et al. (2011b). The procedure was followed in reverse order by starting with the model developed with all species taken together during cross-validation step. This was done to maintain the same samples in the cross and test validation in subsequent step of other



combination models and that for individual species. The component range and data statistics for the two sets were almost identical for all the models as evident from Table 2. The procedure was not attempted for *P. deltooides*, *D. sissoo*, *E. grandis* and *E. camaldulensis* as the number of samples was not sufficient. The statistics of the models is given in Tables 3 and 4 for the two spectral ranges.

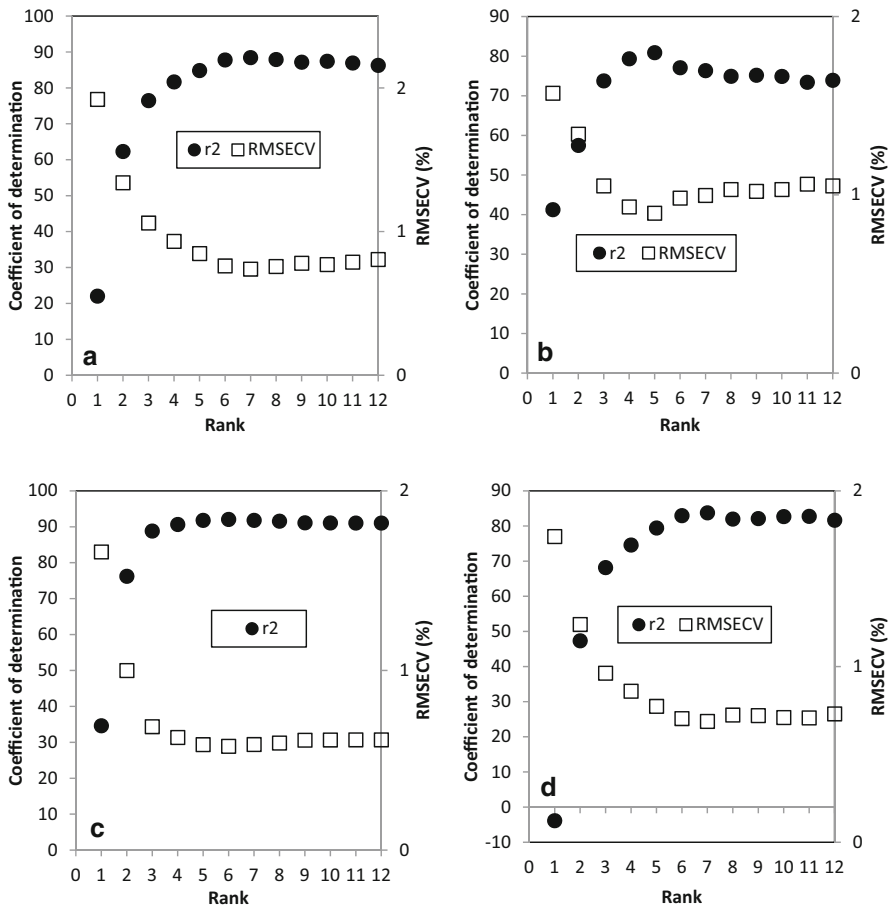
The model statistics decreased marginally in case of all the models selected for cross and test validation in comparison with when all samples were taken together. The difference in ranks was two or  $<2$  for all the models except for two models in Table 3 (all eucalypts plus subabul and all eucalypt species) for the test set. RMSECV was almost equal to 1 % for the subabul models (Tables 3 and 4) with sets CV1 and TS2. Here, outliers also were the same samples as observed during full cross-validation. The models statistics is on the expected lines.



**Fig. 1** Progression of the coefficient of determination ( $r_{cv}^2$ ) and the root-mean-square error of cross-validation (RMSECV) with rank (number of PLS vectors) for spectral range 7502–6098 + 5450–4246  $\text{cm}^{-1}$  in *E. tereticornis* (a), *L. leucocephala* (b), *D. sissoo* (c) and *E. camaldulensis* (d). Dark circles represent  $r_{cv}^2$ , and open squares represent RMSECV

## Over fitting of models

When the number of samples used is high, the full cross-validation process may result in over fitting and the results may be too optimistic. For this reason, the leave more out (PLSR-2) method was also used in the present study for selected models. The number of samples left out during cross-validation was increased up to 40 %, and the results are discussed here. The process was applied to both the spectral ranges used in the present study (Tables 3 and 4). The progression of  $r_{cv}^2$ , RMSECV and RPD was observed. When this method was applied to models of individual species, it was observed that values of RMSECV and  $r_{cv}^2$  remained almost similar when samples left out were increased up to 30–40 % with maximum up to 1 sample

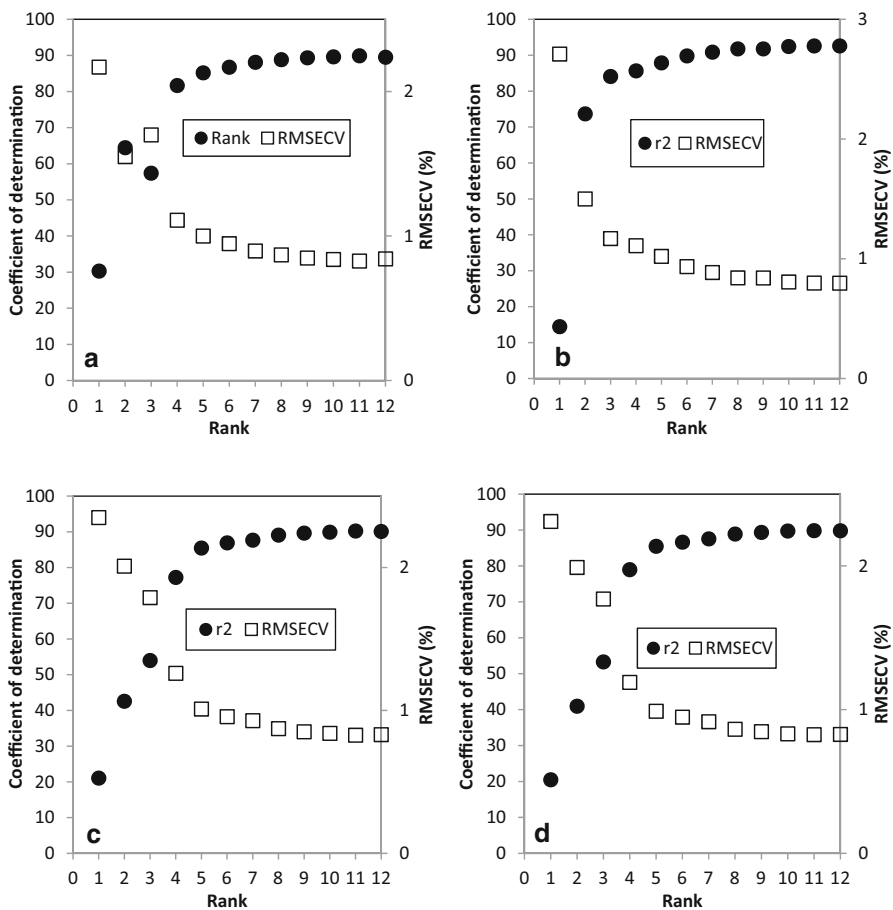


**Fig. 2** Progression of the coefficient of determination ( $r_{cv}^2$ ) and the root-mean-square error of cross-validation (RMSECV) with rank (number of PLS vectors) for spectral range 7502–4246  $\text{cm}^{-1}$  in *E. tereticornis* (a), *L. leucocephala* (b), *D. sissoo* (c) and *E. camaldulensis* (d). Dark circles represent  $r_{cv}^2$ , and open squares represent RMSECV

detected as outlier. RPD also remained above 2.5. For combined models of all species and combination of species, the model statistics (RMSECV and  $r_{cv}^2$ ) remained stable when samples left out were increased up to 20 % (i.e., RMSECV almost equal to 1 % and  $r_{cv}^2$  more than 80 %) and RPD above 2.5. Outliers above 20 % also increased to 3–5. The results of individual species models and combination models are not shown.

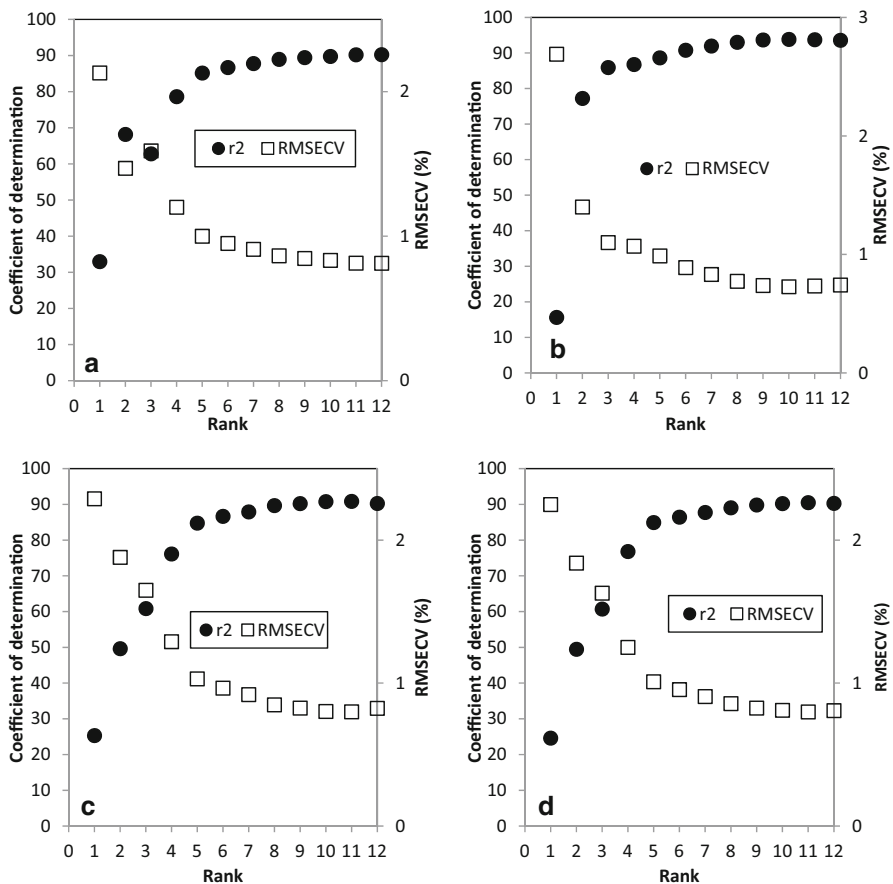
### Construction of final model

For the construction of final model for all the species, samples which were consistently detected outliers during evaluation of all the above models were



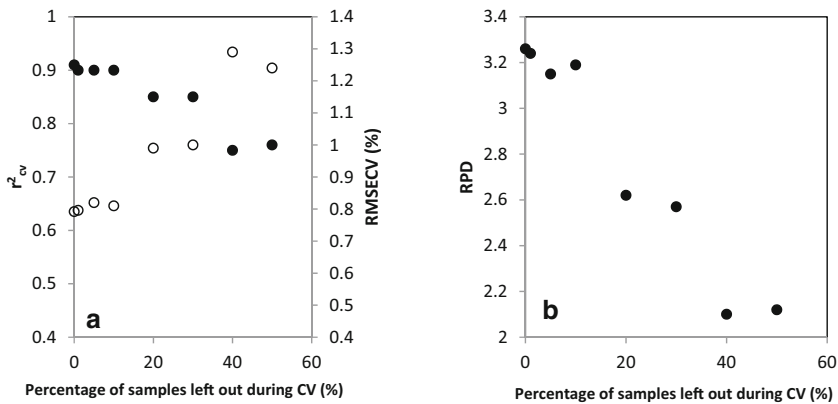
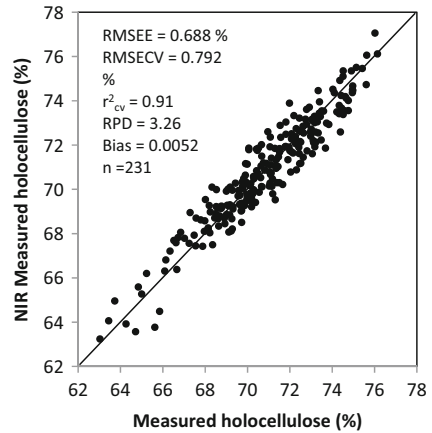
**Fig. 3** Progression of the coefficient of determination ( $r_{cv}^2$ ) and the root-mean-square error of cross-validation (RMSECV) with rank (number of PLS vectors) for spectral range 7502–6098 + 5450–4246  $\text{cm}^{-1}$  for all species, all eucalypts, all eucalypts + subabul, all eucalypts + subabul + shisham. Dark circles represent  $r_{cv}^2$ , and open squares represent RMSECV

eliminated from the final model. Three samples of *E. grandis* and one sample each of *E. tereticornis* and *L. leucocephala* were removed as outliers, and the final model was constructed with 231 samples. Component range was 76.14–63.03 % with standard deviation of 2.586 %. First derivative plus multiplicative scatter correction in wave number range 7502–4246  $\text{cm}^{-1}$  with a factor of nine was found to be most appropriate. The model remained stable even when 30 % of the samples were left out with no outlier detected and RMSECV <1 % and  $r_{\text{cv}}^2$  more than 80 % with RPD above 2.57. The final model has RMSEE (root-mean-square error of estimation) of 0.688 %, RMSECV—0.792 %,  $r_{\text{cv}}^2$ —0.90, rank—9, RPD—3.26, SD/RMSECV—3.265, RMSECV/RMSEE—1.15, RMSEE/SD—0.266). The results of the final model are plotted in Figs. 5 and 6. When dividing the 231 samples into calibration and prediction set (116:115) and then interchanging the sets, the model statistics



**Fig. 4** Progression of the coefficient of determination ( $r_{\text{cv}}^2$ ) and the root-mean-square error of cross-validation (RMSECV) with rank (number of PLS vectors) for spectral range 7502–4246  $\text{cm}^{-1}$  for all species, all eucalypts, all eucalypts + subabul, all eucalypts + subabul + shisham. Dark circles represent  $r_{\text{cv}}^2$ , and open squares represent RMSECV

**Fig. 5** Plot of predicted holocellulose versus measured holocellulose by combined NIR model for *E. tereticornis*, *E. camaldulensis*, *E. grandis*, *E. hybrid*, *L. leucocephala*, *D. sissoo* and *P. deltoides* using milled samples



**Fig. 6** RMSECV,  $r^2_{cv}$  and RPD versus the percentage of samples left out during cross-validation. Dark circle indicates  $r^2_{cv}$  and RPD, and open circle indicates RMSECV

was comparable ( $r^2_{cv} = 0.88/0.89$ , RMSECV = 0.919/0.831 %,  $r^2_p = 0.90/0.90$ , RMSEP = 0.797/0.800 %, RPD = 3.21/3.26, RER = 16.44/15.72).

## Conclusion

NIR spectroscopy combined with multivariate analysis was used to develop combined PLSR-based NIR models for estimation of holocellulose of six common plantation species, namely *Eucalyptus tereticornis* (including few samples of *E. hybrid* of *E. tereticornis* × *E. Camaldulensis*), *E. grandis*, *E. camaldulensis*, *L. leucocephala* (subabul), *D. sissoo* (shisham) and *P. deltoides* (poplar). Un-extracted milled wood samples having particle size of 250–400  $\mu\text{m}$  (40–60 mesh size) were used for the study. Two combinations of spectral ranges, namely 7502–6098 + 5450–4246 and 7502–4246  $\text{cm}^{-1}$  with 1st derivative (9-point

smoothing) plus multiplicative scatter correction pretreatment were employed. Models were developed for individual species, combination of species (all eucalyptus species; all species barring poplar; all species barring shisham and poplar; all species barring poplar and subabul; all species barring poplar, subabul and *E. grandis*) and combined model for all species.

A model of individual species was constructed with four to six factors and explains 80 % of the variations and has <1 % of RMSECV. Except for *E. grandis* and *P. deltooides*, all models have RPD of more than 2.25. For all eucalyptus species (component range 63.03–76.14 %) in both the spectral ranges,  $r_{cv}^2$  of determination for cross-validation was 0.92/0.93, RMSECV—0.841/0.774 %, and RPD of 3.49/3.79, respectively, using eight factors with one outlier common in both the models belonging to *E. tereticornis*.

Methods of NIR models evaluation involved PLSR-1 (full cross-validation), PLSR-2 (leaving more than one sample out) and dividing the samples into calibration and prediction (test) sets and interchanging them from calibration to prediction sets. RPD, RER, RMSEP/RMSECV and SD/RMSECV were estimated for individual species models and combination species models.

Final combined model in the component range 76.14–63.03 % (standard deviation 2.586 %) for all the species was developed using nine factors in the spectral range 7502–4246  $\text{cm}^{-1}$  using 1st derivative plus MSC by removing five samples found as outliers in all the evaluation steps and in most of the models. The model remained stable even when 30 % of the samples were left out. When dividing the samples into calibration and prediction set (116:115) and then interchanging the sets, the model statistics was almost identical ( $r_{cv}^2 = 0.88/0.89$ , RMSECV = 0.919/0.831 %,  $r_p^2 = 0.90/0.90$ , RMSEP = 0.797/0.800 %, RPD = 3.21/3.26, RER = 16.44/15.72).

**Acknowledgments** The authors are thankful to CSIR-NMITLI for providing funds for purchase of FT-NIR equipment and providing financial support to second author. Financial support to third author from DST is also acknowledged. Contribution to project partners of CSIR-NMITLI networking Institute for collecting samples of *L. leucocephala* (subabul) from Himanchal and Chattisgarh is acknowledged. Dr. Madhumita Gosh of Institute of Forest Genetics and Tree Breeding, Coimbatore, India, is acknowledged for providing some wood samples of *E. tereticornis*, *E. camaldulensis*, *E. grandis* and *E. hybrid*. WIMCO Ltd. is also acknowledged for providing some material of *Populus deltooides* from Rudrapur, Uttarakhand.

## References

- AACC (1999) Near-infrared methods—guidelines for model development and maintenance, American Association of Cereal Chemists (AACC). AACC Method 39-00 15
- Aenugu HPR, Kumar DS, Srisudharson Parthiban N, Ghogh SS, Banji D (2011) Near infrared spectroscopy—an overview. Int J ChemTech Res 3(2):825–836
- Alves A, Santos A, Rozenberg P, Paques Luc E, Charpentier JP, Schwanninger M, Rodrigues J (2012) A common near infrared—based partial least squares regression model for the prediction of wood density of *Pinus pinaster* and *Larix × eurolepis*. Wood Sci Technol 46:157–175
- Barton FE II (2004) Progress in near infrared spectroscopy: the people, the instrumentation, the applicants. In: Davies AMC, Garrido-VaroNear A (eds) Infrared spectroscopy proceedings of the 11th international conference, Cordoba, Spain. NIR publications, Chichester, pp 13–18
- Berzaghi P, Flinn PC, Dardenne P, Lagerholm M, Shenk JS, Westerhaus MO, Cowe IA (2002) Comparison of linear and non-linear near infrared calibration methods using large forage databases.

- In: Davies AMC, Cho RK (eds) Near infrared spectroscopy: proceedings of the 10th international conference, NIR Publications, Chichester, UK, p 107
- Bokobza L (1998) Near infrared spectroscopy. *J Near Infrared Spectrosc* 6:3–17
- Dardenne (2004) Calibration transfer in near infrared spectroscopy. In: Davies AMC, Garrido-Varo A (eds) Near infrared spectroscopy: proceedings of the 11th international conference, NIR Publications, Chichester, UK, p 19
- Derkyi NSA, Adu-Amankwa B, Sekyere D, Darkwa NA (2011) Rapid prediction of extractives and polyphenolic contents in *Pinus caribaea* bark using infrared reflectance spectroscopy. *Int J of Appl Sci* 2(1):63–73
- DiFoggio R (1995) Examination of some misconceptions about near infrared analysis. *Appl Spectrosc* 49:67–75
- Ding L, Xiang YH, Huang AM, Zhang ZY (2009) Quantitative prediction of holocellulose, lignin and microfibril angle of Chinese Fir by BP-ANN and NIR spectrometry. *Spectrosc Spect Anal* 29(7):1784–1787
- Downes G, Meder R, Hardwood C (2010) A multi-site, multi-species near infrared calibration for the prediction of cellulose content in eucalyptus wood meal. *J Near Infrared Spectrosc* 18:381–387
- Fearn T (2002) Assessing calibrations: SEP, RPD, RER and R2. *NIR News* 13(6):12–14
- Fujimoto T, Kobori H, Tsuchikawa S (2012) Prediction of wood density independently of moisture conditions using near infrared spectroscopy. *J Near Infrared Spectrosc* 20:353–359
- Garbutt DCF, Donkin MJ, Meyer JH (1992) Near infrared reflectance analysis of cellulose and lignin in wood. *Pap S Afr* 2(4):45–48
- Geladi P (2002) Some recent trends in the calibration literature. *Chemometr Intell Lab* 60:211–224
- Geladi P, Kowalski BR (1986) Partial least squares regression: a tutorial. *Anal Chim Acta* 185:1–17
- Gong YM, Zhang W (2008) Recent progress in NIR spectroscopy technology and its application to the field of forestry. *Spectrosc Spect Anal* 28(7):1544–1548
- Haenlein M, Kaplan AM (2004) A beginner's guide to partial least square analysis. *Underst Stat* 3(4):283–297
- Hauksson JB, Bergqvist G, Bergsten U, Sjostrom M, Edlund U (2001) Prediction of basic wood properties for Norway spruce: interpretation of near infrared spectroscopy data using partial least squares regression. *Wood Sci Technol* 35:475–485
- He WM, Xue CY, Nie Y, Li YM (2010) Rapid prediction of wood cellulose, pentosan and Klason lignin contents using near infrared spectroscopy. *Trans China Pulp Pap* 25(3):9–12
- Hein PRG, Lima JT, Chaix G (2010) Effects of sample preparation on NIR spectroscopic estimation of chemical properties of *Eucalyptus urophylla* S.T. Blake wood. *Holzforschung* 64(1):45–54
- Hodge GR, Woodbridge WC (2004) Use of near infrared spectroscopy to predict lignin content in tropical and subtropical pines. *J Near Infrared Spectrosc* 12:381
- Hodge GR, Woodbridge WC (2010) Global near infrared models to predict lignin and cellulose content of pine wood. *J Near Infrared Spectrosc* 18:367–380
- Hou S, Li L (2010) Rapid characterization of woody biomass digestibility and chemical composition using near-infrared spectroscopy. *J Integr Plant Biol* 00(00):1–10
- Huang AM, Jiang ZH, Li GY (2007) Determination of holocellulose and lignin content in Chinese fir by near infrared spectroscopy. *Spectrosc Spect Anal* 27(7):1328–1331
- Inagaki T, Schwanninger M, Kato R, Kurata Y, Thanapase W, Puthson P, Tsuchikawa S (2012) *Eucalyptus camaldulensis* density and fiber length estimated by near-infrared spectroscopy. *Wood Sci Technol* 46:143–155
- Ishizuka S, Sakai Y, Tanaka-Oda A (2012) Quantifying lignin and holocellulose content in conifers decayed wood using near-infrared reflectance spectroscopy. *J For Res* 19(1):233–237
- Jouan-Rimbaud D, Bouveresse E, Massart DL, de Noord OE (1999) Detection of prediction outliers and inliers in multivariate calibration. *Anal Chim Acta* 388(3):283
- Kothiyal V, Raturi A (2011) Estimating mechanical properties and specific gravity for five year old *Eucalyptus tereticornis* having broad moisture content range by NIR spectroscopy. *Holzforschung* 65(5):757–762
- Kothiyal V, Raturi A, Kaler Naithani S (2012) Klason lignin estimation in *Leucaena leucocephala* by near infrared spectroscopy for selection of superior material for pulp and paper. *J Indian Acad Wood Sci* 9(2):105–114
- Kothiyal V, Raturi A, Jaideep, Dubey YM (2014) Enhancing the applicability of near infrared spectroscopy for estimating specific gravity of green timber from *Eucalyptus tereticornis* by

- developing composite calibration using both radial and tangential face of wood. *Eur J Wood Prod* 72(1):11–20
- Lu J, McClure WF (1998) Effect of random noise on the performance of NIR calibrations. *J Near Infrared Spectrosc* 6:77–87
- Pasikatan MC, Steele JL, Spillman CK, Haque E (2001) Near infrared reflectance spectroscopy for online particle size analysis of powders and ground materials. *J Near Infrared Spectrosc* 9:153–164
- Pasquini C (2003) Near infrared spectroscopy: fundamentals, practical aspects and analytical applications. *J Braz Chem Soc* 14:198–219
- Raturi A, Kothiyal V, Uniyal KK, Semalty PD (2012) Development and evaluation of models for specific gravity of *Eucalyptus tereticornis* wood by Fourier transformed near infrared spectroscopy and partial least squares regression analysis. *J Indian Acad Wood Sci* 9(1):40–45
- Rinnan A, Berg FVD, Engelsen SB (2009) Review of the most common pre-processing techniques for near-infrared spectra. *Trends Anal Chem* 28(10):1201–1222
- Rodrigues J, Alves A, Pereira H, da Silva Perez D, Chantre G, Schwanninger M (2006) NIR PLSR results obtained by calibration with noisy, low-precision reference values: are the results acceptable? *Holzforschung* 60(2):402–408
- 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* 38:14–35
- Schimleck LR, Evans R, Ilic J (2001) Application of near infrared spectroscopy to a diverse range of species demonstrating wide density and stiffness variation. *IAWA* 22(4):415–429
- Schimleck LR, Evans R, Ilic J (2003) Application of near infrared spectroscopy to the extracted wood of a diverse range of species. *IAWA* 24(4):429–438
- Schimleck LR, Sturzenbecher R, Jones PD, Evans R (2004) Development of wood property calibrations using near infrared spectra having different spectral resolutions. *J Near Infrared Spectrosc* 12:55–61
- Schimleck LR, Sturzenbecher R, Mora C, Jones PD, Daniels RF (2005) Comparison of *Pinus taeda* L. wood property calibration based on NIR spectra from the radial-longitudinal and radial transverse faces of wood strips. *Holzforschung* 59:214–218
- Schimleck LR, Hodge GR, Woodbridge W (2010) Toward global calibrations for estimating the wood properties of tropical, sub-tropical and temperate pine species. *J Near Infrared Spectrosc* 18:355–365
- Schwanninger M, Rodrigues JC, Fackler K (2011a) A review of band assignments in near infrared spectra of wood and wood components. *J Near Infrared Spectrosc* 19:287–308
- Schwanninger M, Rodrigues JC, Gierlinger N, Hinterstoisser B (2011b) Determination of lignin content in Norway spruce wood by Fourier transformed near infrared spectroscopy and partial least square regression. Part 1: wavenumber selection and evaluation of the selected range. *J Near Infrared Spectrosc* 19(5):319–329
- Schwanninger M, Rodrigues JC, Gierlinger N, Hinterstoisser B (2011c) Determination of lignin content in Norway spruce wood by Fourier transformed near infrared spectroscopy and partial least square regression. Part 2: development and evaluation of the final model. *J Near Infrared Spectrosc* 19(5):331–341
- So CL, Via BK, Groom LH, Schimleck LR, Shupe TF, Kelley SS, Rials TG (2004) Near infrared spectroscopy in the forest products industry. *For Prod J* 54(3):6–16
- Starr C, Morgan AG, Smith DB (1981) An evaluation of near-infrared reflectance analysis in some plant-breeding programs. *J Agric Sci* 97:107–118
- Tsuchikawa S (1998) Non-traditional applications of near infrared spectroscopy based on the optical characteristic models for a biological material having cellular structure. *J Near Infrared Spectrosc* 6:41–46
- Tsuchikawa S (2007) A review of recent near infrared research for wood and paper. *Appl Spectrosc Rev* 42:43–71
- Tsuchikawa S, Schwanninger M (2013) A review of recent near infrared research for wood and paper (Part 2). *Appl Spectrosc Rev* 48:560–587
- Williams P, Norris K (2004) Near-Infrared technology in the agricultural and food industries. St. Paul, American Association of Cereal Chemists, p 26
- Wise LE, Murphy M, Addieco AAD (1946) Chlorite holocellulose, its fractionation and bearing on summative wood analysis and studies on the hemicelluloses. *Pap Trade J* 122(2):35–43
- Workman JJ (1999) Review of process and non-invasive near-infrared and infrared spectroscopy: 1993–1999. *Appl Spectrosc Rev* 34:1–89
- Yao S, Pu JW (2009) Application of near infrared spectroscopy in analysis of wood properties. *Spectrosc Spect Anal* 29(4):974–978



- Yao S, Wu GF, Jiang YF, Fu XD, Lu HK, Su M, Pu JW (2010a) Extending hemicelluloses content calibration of *Acacia* spp. using NIR to new sites. *Spectrosc Spect Anal* 30(5):1206–1209
- Yao S, Xing M, Zhou S, Wu G, Jiang Y, Pu J (2010b) The accuracy of near infrared prediction of hemicelluloses content arising from varying introduced error. *J Near Infrared Spectrosc* 18(6):397–402
- Yao S, Wu GF, Zhou SK, Jiang YF, Jin XJ, Zhao Q, Pu JW (2011) The influence of reference data noise on the NIR prediction results. *Spectrosc Spect Anal* 31(5):1216–1219
- Yeh TF, Chang H, Kadla JF (2004) Rapid prediction of solid wood lignin content using transmittance near infrared spectroscopy. *J Agric Food Chem* 52:1435–1439
- Yeniav O, Goktas A (2002) A comparison of Partial least squares regressions with other prediction methods. *Hacet J of Math Stat* 31:99–111