



# Predicting the chemical composition of juvenile and mature woods in Scots pine (*Pinus sylvestris* L.) using FTIR spectroscopy

Tomas Funda<sup>1,2,3</sup> · Irena Fundova<sup>1,4</sup> · András Gorzsás<sup>5</sup> · Anders Fries<sup>1</sup> · Harry X. Wu<sup>1,6,7</sup>

Received: 5 June 2019 / Published online: 8 February 2020  
© The Author(s) 2020

## Abstract

The chemical composition of wood is one of the key features that determine wood quality. The focus of this study was on identifying differences between juvenile and mature woods in Scots pine (*Pinus sylvestris* L.) and developing models for predicting the chemical composition of these two wood types. Chemical traits, determined by traditional wet chemistry techniques, included the proportion of lignin, polysaccharides and extractives. Partial least squares regression of Fourier transform infrared (FTIR) spectra was used for model building. The model performance was primarily evaluated by root mean squared error of predictions (RMSEP). High predictive power was attained for the content of lignin (RMSEP of 0.476 and 0.495 for juvenile and mature woods, respectively) and extractives (0.302 and 0.471), good predictive power for cellulose (0.715 and 0.696) and hemicelluloses in juvenile wood (0.719) and low predictive power for hemicelluloses in mature wood (0.823). A distinct band was observed at  $1693\text{ cm}^{-1}$ , and its intensity was strongly associated with the content of extractives ( $r=0.968$  and  $0.861$  in juvenile and mature woods, respectively). FTIR has proved suitable for the rapid, non-destructive, cost-efficient assessment of the chemical composition of juvenile and mature woods in Scots pine. The band at  $1693\text{ cm}^{-1}$  is to be further investigated to unravel its link with individual extractive components.

---

**Electronic supplementary material** The online version of this article (<https://doi.org/10.1007/s00226-020-01159-4>) contains supplementary material, which is available to authorized users.

---

✉ Harry X. Wu  
Harry.Wu@slu.se

Extended author information available on the last page of the article

## Introduction

Wood is an abundant and renewable natural composite material that has been utilized for a great variety of purposes such as production of paper, construction lumber, furniture or textiles. It has been receiving increasing interest as an alternative, carbon-neutral source of energy to fossil fuels (Baratieri et al. 2008; Morris et al. 2000) with a great potential to help balance net CO<sub>2</sub> emissions (Gnansounou et al. 2009) and mitigate environmental pollution (Acquah et al. 2016b). Its properties are determined by its anatomical and chemical structure, i.e. by the presence, extent and distribution of different types of wood tissue, by wood cell anatomy and by the chemical composition, in particular, the relative proportion of different chemical components and their allocation (Pereira et al. 2003).

Cellulose, hemicelluloses and lignin, polymeric macromolecules that are mainly allocated in plant cell walls, constitute the structural components of wood, with cellulose microfibrils and hemicellulosic chains being embedded in lignin (Rubin 2008). They jointly form ca 90–96% of total wood material (40–50%, 25–35% and 18–35% of the dry weight, respectively) (Pettersen 1984). The rest falls on a large and diverse group of extractives (Ekeberg et al. 2006; Shebani et al. 2008), comprising of both organic and inorganic compounds, among which the most important are pinosylvin, stilbenes, resin acids, fatty acids and sterols (Fries et al. 2000). While these four major components are common to all wood materials, their proportions vary among and within tree species and also depend on the age and part of a tree, the geographic location and soil conditions (Pettersen 1984).

The chemical composition of wood strongly influences wood quality (*sensu lato*) (Barnett and Jeronimidis 2009), but its determination under laboratory conditions using conventional methods is expensive, time-consuming and laborious and is therefore impractical in projects that involve large numbers of samples, such as genetic studies or breeding programs (Gebreselassie et al. 2017). High-throughput vibrational spectroscopy-based techniques, such as near-infrared (NIR; Tsuchikawa and Kobori 2015), Fourier transform infrared (FTIR; Rodriguez-Saona and Allendorf 2011) or Raman (Bowley et al. 2012) spectroscopy, have the potential to overcome phenotyping limitations (Conrad and Bonello 2016), as they can rapidly and inexpensively generate chemical fingerprints of various biological samples, including wood. The great advantage of these techniques is that chemical composition needs to be determined only for a small subset of samples, while those of the remaining ones will be predicted from the corresponding IR or Raman spectral profiles. Principal component regression (PCR), soft independent modelling of class analogy (SIMCA) and partial least squares regression (PLSR) belong to the most widely employed statistical methods in this context (Cozzolino 2014; Zhou et al. 2015).

FTIR spectroscopy is a powerful analytical tool for a rapid and accurate characterization of lignocellulosic biomass (Dokken et al. 2005) based on the presence of fundamental molecular vibrations (Acquah et al. 2016b). It can provide information, albeit indirect, about the chemical composition of a sample, including molecular structural details (Diem 1993; Gillgren and Gorzsás 2016). It has been successfully

applied as a surrogate to the traditional wet laboratory protocols in a number of forest tree species (Chen et al. 2010; Poletto et al. 2012), for the characterization of different types of biomass such as wood, slash, wood with bark (Acquah et al. 2016a, b), fibres (Åkerholm and Salmén 2001) and pulp (Bjarnestad and Dahlman 2002; Strunk et al. 2011) and for the identification of chemical changes occurring during particle and fibreboard production (Müller et al. 2009). Owing to the high sensitivity and the qualitative and (semi)quantitative nature of the analysis, it is capable of detecting minor differences in the chemical composition among samples (Rodríguez-Saona and Allendorf 2011). It is therefore particularly advantageous in studies where only small variations in the chemical composition are expected, for instance, when samples from different developmental stages of the same trees (juvenile versus mature wood) are analysed. Juvenile and mature woods differ in a number of key properties such as density, stiffness, tracheid length or the chemical composition, which strongly affect the wood's suitability for further utilization (Burdon et al. 2004; Ivkovic et al. 2013; Sykes et al. 2003). Since juvenile wood is considered to be inferior in many aspects, it is important to investigate the two types independently, especially if one takes into account the increasing proportion of juvenile wood in genetically improved trees as a consequence of shortened rotation periods (Pearson and Gilmore 1980). Although information regarding the samples' location within a tree (i.e. distance from the pith and bark) is scarcely provided, most published studies seem to have focused on mature wood only.

The objective of this study was to develop PLSR models for predicting the chemical composition of Scots pine (*Pinus sylvestris* L.) wood from FTIR spectra. Focus was placed on identifying differences between spectra obtained from juvenile and mature woods, and the predictive power of their models was compared. This could ideally pave the way for subsequent rapid screenings of individuals intended for inclusion in advanced cycles of tree breeding.

## Materials and methods

### Sample population

Samples for this study were taken from two Scots pine (*P. sylvestris* L.) full-sib progeny tests: “Skorped” (411-2-H72-Skorped-Y; latitude 63.3444° N, longitude 17.6417° E, altitude 330 m a.s.l.) and “Vännäsby” (411-3-V73-Vännäsby-AC; latitude 64.0250° N, longitude 19.8519° E, altitude 200 m a.s.l.). Seeds for both tests were sown in May 1972 in Skogforsk (the Forestry Research Institute of Sweden), Sävar. The tests were established by Skogforsk in October 1972 and May 1973, respectively, as part of a broader progeny test series. They were established on normal forest soils with intermediate fertility using completely randomized single-tree plot design. The study was conducted at the age of 43 years; thus, 30–35 years had passed since the trees reached breast height. The central part (hereafter classified as juvenile wood) had formed heartwood, which is—especially in Scots pine—very rich in extractives.

## Increment core collection

A subset of 40 trees on each test site was originally sampled for this study. In order to avoid jeopardizing the stability of trees, 5-mm-wide increment cores were collected, instead of standard 12-mm ones. This, unfortunately, led to insufficient amounts of wood material for some samples for subsequent analyses. Therefore, only 70 juvenile wood samples (36 from Skorped and 34 from Vännäsby) and 39 mature wood samples from Vännäsby were finally included in this study. The criterion for selection of trees was to cover as much phenotypic variation as possible in wood density and acoustic velocity from earlier measurements of ca 1200 trees using Resistograph IML-RESI PD300 (Instrumenta Mechanik Labor, Germany) and Hitman ST300 (Fibre-gen, Christchurch, NZ), respectively.

The increment cores were extracted from trees at breast height (ca 1.3 m above ground) using a 5-mm core borer (Haglöf, Långsele, Sweden) that was drilled through a stem bark to bark with a portable, battery-operated machine. After extraction, the cores were inserted into paper tubes and kept as such in a laboratory at ca +23 °C and 30% relative humidity until further processing. Each increment core was divided into a front and rear half at the pith; the front part corresponding to the side where the borer was attached to a stem. Each half was further split into juvenile and mature wood sections, which were represented by annual rings 2–6 as counted from the pith and 8–12 as counted from the bark, respectively. The outermost annual rings were removed to avoid contamination by material from earlier taken cores as well as to reduce the presence of the tree's own bark. The core sections (four pieces from each tree) were temporarily stored in locked 2.0-mL plastic tubes.

## Wood sample preparation

The core sections were cut into ca 2-mm-wide pieces and ground in a Retch MM400 ball mill (Retch GmbH, Haan, Germany) at the frequency of 30 Hz using metal jars with 12-mm metal balls. Two milling cycles of 40 s were conducted, with a ca 2 min gap between them to avoid sample overheating. Approximately 7 mg of wood powder was mixed with 390 mg of spectroscopy-grade KBr (Sigma-Aldrich, St. Louis, MO, USA) and manually finely ground using agate mortar and pestle. When not utilized immediately after grinding, the KBr-wood mixture powder was stored in locked 2.0-mL plastic tubes in a paper box with silica gel in a low-moisture environment.

## Analysis of chemical composition

Chemical compositional analysis of the 109 wood samples was performed in MoRe Research (Örnsköldsvik, Sweden) using the wood powder from ball milling (142–389 mg per sample;  $245.7 \pm 64.8$  mg SD). Carbohydrates were determined using protocol SCAN-CM 71:09 (Scandinavian Pulp, Paper and Board Testing Committee 2009) that involves hydrolysis of wood with sulphuric acid. The content

of glucose (*Glu*), xylose (*Xyl*), mannose (*Man*), galactose (*Gal*) and arabinose (*Ara*) as the principal monosaccharides was quantified by ion chromatography (IC) using Dionex ICS-5000 (Thermo Scientific Inc., Sunnyvale, CA, USA) and was subsequently used to calculate the proportion of cellulose (*Cel*) and hemicelluloses (*Hem*) in total carbohydrates as  $Glu - \frac{1}{3} Man$  and  $1 - Cel$ , respectively (Sjöström 1993). The remaining solid residue from the acid hydrolysis of carbohydrates was used for the determination of lignin content. Total lignin (*Lig*) was quantified as a sum of gravimetrically determined acid-insoluble (Klason) and spectrophotometrically determined acid-soluble lignin following the TAPPI protocols 222 om-02 (TAPPI 2002) and UM 250 (TAPPI 1991), respectively, with a slight modification to accommodate for small amounts of wood (< 1 g). The acid-soluble lignin was determined in a solution after filtering off the insoluble lignin, using a spectrophotometric method at wavelength 205 nm, while the acid-insoluble lignin, yielded from the filtering step, was dried and weighed. The analysis of non-volatile extractives (*Ext*) was performed in a small-scale extraction equipment to lower the amount of material needed, but it otherwise followed the gravimetric determination of extractives according to the Soxhlet extraction protocol SCAN-CM 67:03 (Scandinavian Pulp, Paper and Board Testing Committee 2003). Briefly, samples were treated with cyclohexane/acetone (9:1 ratio) at boiling temperature for 1 h and refluxed. The solution was filtered off and samples were washed with cyclohexane/acetone several times. The amount of extractives was weighed after drying. Samples were not extracted prior to lignin determination; therefore, lignin content was corrected for the extractive content after the analysis of extractives was performed.

## FTIR spectroscopic analysis

### Spectra acquisition

FTIR spectroscopic analysis was performed at the Vibrational Spectroscopy Core Facility at Umeå University, using a Bruker IFS 66v/S vacuum bench spectrometer (Bruker Optics, Ettlingen, Germany). Spectra from 70 juvenile and 39 mature wood samples were collected over the spectral range of 5200–400  $\text{cm}^{-1}$  at 4  $\text{cm}^{-1}$  spectral resolution (and a zero filling factor of 2). 128 scans per sample were co-added to obtain good signal-to-noise ratios. Background spectra were collected with the same settings, using pure KBr powder. Measurements were repeated when absorbance values were outside the 0.1 and 0.8 range and/or when a spectrum exhibited excess noise. 11% of the samples were replicated, and their superimposed spectral profiles were used to visually assess the levels of technical errors, including potential inconsistencies in instrument performance.

### Spectra standardization

Raw absorbance spectra were exported using OPUS 7.0 (Bruker Optics, Ettlingen, Germany) and standardized for subsequent prediction model calibration using the free, open-source graphical user interface developed at the Vibrational

Spectroscopy Core Facility at Umeå University (<https://www.umu.se/en/research/infrastructure/visp/downloads/>). The standardization procedure involved four steps: (1) cutting the spectral region to the 1870–770  $\text{cm}^{-1}$  range (“finger-print region”) to eliminate non-informative and hard-to-standardize regions (i.e. maximize information content and minimize noise); (2) baseline correction via asymmetrical least squares fitting (Eilers 2004), with parameters  $\lambda$  and  $p$  set to  $10^6$  and  $10^{-3}$ , respectively (Eilers and Boelens 2005); (3) normalization using two different methods: (a) total area normalization (TAN) performed over the 1870–770  $\text{cm}^{-1}$  range and (b) area minimum–maximum normalization (AMM), i.e. scaling the spectra according to a selected region corresponding to prominent features. These regions in the present case were 925–1145  $\text{cm}^{-1}$  (dominated by carbohydrate-associated bands), 1487–1553  $\text{cm}^{-1}$  (dominated by aromatic skeletal vibrations), and 1553–1633  $\text{cm}^{-1}$  (dominated by aromatics and extractives) (Dokken et al. 2005; Gierlinger 2018; Gorzsás and Sundberg 2014), referred to as AMM1, AMM2 and AMM3, respectively; and (4) smoothing, using a mild Savitzky–Golay filtering (Savitzky and Golay 1964) with first-order polynomial and a frame rate of 3, since original signal-to-noise ratios were already high.

## Model calibration

Models for predicting the chemical composition of wood from standardized FTIR spectra were developed using partial least squares regression (PLSR), a method that is based on the singular value decomposition of the  $X'Y$  matrix (i.e. predictor and response variables), with the objective to extract successive linear combinations of the predictor variables (aka latent variables or factors) to simultaneously explain as much variation in the predictor and response variables as possible. The computation was performed using the statistical package SAS 9.4 (PROC PLS, SAS Institute Inc., Cary, NC, USA) using the nonlinear iterative partial least squares (NIPALS) algorithm.

In total, 570 predictor variables were included in the PLS regression, each representing the intensity value at a given wavenumber in the 1870–770  $\text{cm}^{-1}$  range (at steps of ca 1.925  $\text{cm}^{-1}$ , resulting from the spectral resolution and the zero filling factor). The response chemical variables were comprised of the content of total lignin, cellulose, hemicelluloses and extractives as well as the five principle monosaccharides (glucose, xylose, mannose, galactose and arabinose) measured independently. Each of the response chemical variables was modelled separately. Raw spectra (RAW) and four sets of standardized spectra based on different normalization procedures were tested (TAN and AMM1–3), and three series of models were produced for each chemical variable: one for juvenile wood (based on 70 samples), one for mature wood (based on 39 samples) and one for pooled samples (based on 109 samples). All models were validated using a split-sample cross-validation test, in which groups of every seventh observation beginning with the first, second and so forth were excluded from calibration data sets.

The split-sample-validated root mean predicted residual error sum of squares (PRESS), also known as the root mean squared error of predictions (RMSEP), calculated as

$$\text{RMSEP} = \sqrt{\frac{\text{PRESS}}{n}} = \sqrt{\frac{\sum_{i=1}^n (y_i - \hat{y}_{i(i)})^2}{n}} \quad (1)$$

was used as the benchmark statistics during calibration. The symbol  $n$  denotes the number of observations,  $y_i$  is the  $i$ th observation for response variable  $y$  and  $\hat{y}_{i(i)}$  is the predicted value for the  $i$ th case; the second subscript indicates that the  $i$ th case was omitted when the regression model was fitted (Kutner et al. 2005). Van der Voet's randomization-based model comparison test (Van der Voet 1994) was then applied as the primary criterion for model selection, as it minimizes the number of retained factors through the exclusion of factors that have only a marginally higher RMSEP value than the absolute minimum, and thus reduces model complexity and the risk of overfitting. The cut-off probability for declaring a non-significant difference between factors was set at 0.1. The presence of outliers was visually assessed with the aid of diagnostics plots, which show the distance of each data point to the PLSR model with reference to both predictor and response variables. When an observation was found to lie dramatically far from other observations, it was removed from the calibration data set and the procedure was repeated.

Simple linear regression (PROC REG in SAS) was applied to quantify the relationship between absorbance intensities at individual wavenumbers (570 variables) and four major chemical components (*Lig*, *Cel*, *Hem* and *Ext*) and thereby evaluate how the predictive power is distributed across the whole spectral range. In each regression model, only a single wavenumber was supplied as a predictor variable; therefore, the statistical significance level ( $\alpha=0.05$ ) was adjusted using Šidák correction  $1 - (1 - \alpha)^{1/m}$  to accommodate for the large number of independent hypotheses tested ( $m=570$ ).

## Results

### Analysis of chemical composition

All 109 samples included in the study provided estimates for the nine chemical variables despite the limited amounts of wood material used. The only exception was one sample of juvenile wood for which there was not enough powder left to perform the cyclohexane–acetone extraction analysis, and therefore, its extractive content could not be obtained. Chemical composition differed between juvenile and mature wood samples. While the proportion was similar for hemicelluloses and almost equal (ca 1% difference) for lignin between the two groups, there was on average nearly 9% less cellulose (and, correspondingly, glucose) and over 5% more extractives in juvenile wood than in mature wood. Variation in chemical composition among individual trees was comparable, but slightly higher

**Table 1** Descriptive statistics of the chemical composition of 70 juvenile and 39 mature wood samples

Component	Juvenile wood content (%)			Mature wood content (%)		
	Mean	SD	CV	Mean	SD	CV
Lignin	27.1	2.26	8.4	27.7	1.97	7.1
Cellulose	32.6	2.31	7.1	41.5	2.83	6.8
Hemicelluloses	22.9	1.61	7.0	21.9	1.02	4.6
Glucose	35.3	2.51	7.1	44.9	3.12	7.0
Mannose	7.9	0.81	10.2	10.1	1.04	10.3
Xylose	5.9	0.53	8.9	4.6	0.41	9.0
Galactose	4.7	1.39	29.8	2.6	1.48	57.0
Arabinose	1.8	0.22	12.3	1.3	0.11	9.0
Extractives	7.4 <sup>a</sup>	3.99 <sup>a</sup>	53.6 <sup>a</sup>	1.9	1.25	64.1

*SD* standard deviation, *CV* coefficient of variation

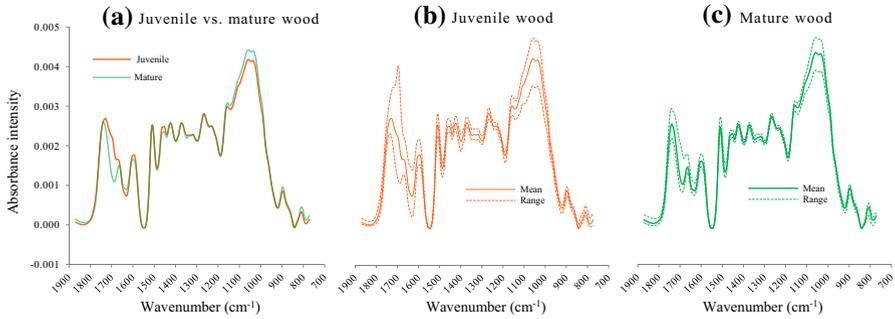
<sup>a</sup>Based on 69 samples

in juvenile wood, with the highest variation in galactose and extractive content (Table 1). The coefficient of variation for the extractives was higher in mature wood (64.1% vs. 53.6%), but the range of individual values was much greater in juvenile wood, extending from 2.0 to 19.8% as compared with only between 0.7 and 6.7% in mature wood.

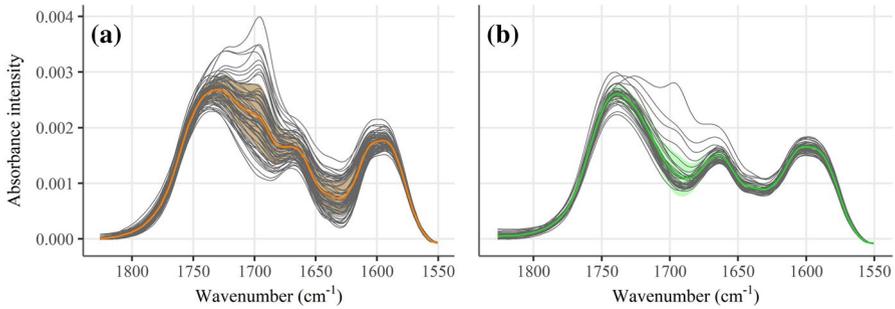
## FTIR spectra

All 70 juvenile and 39 mature wood samples provided interpretable spectra over the fingerprint region (1870–770  $\text{cm}^{-1}$ ). In this region, spectra have a complex pattern, containing a number of bands that are indicative of the presence of main chemical components of wood. Bands around 1739  $\text{cm}^{-1}$ , 1317  $\text{cm}^{-1}$ , 1157  $\text{cm}^{-1}$  and 897  $\text{cm}^{-1}$  have been earlier assigned to bending or stretching vibrations of different functional groups in polysaccharidic compounds, bands around 1595  $\text{cm}^{-1}$ , 1510  $\text{cm}^{-1}$ , 1270  $\text{cm}^{-1}$  and 1230  $\text{cm}^{-1}$  to those in aromatic compounds (in this case mainly lignin) and bands around 1465  $\text{cm}^{-1}$ , 1425  $\text{cm}^{-1}$ , 1375  $\text{cm}^{-1}$ , 1111  $\text{cm}^{-1}$  and 1030  $\text{cm}^{-1}$  to those in both polysaccharidic and aromatic compounds (see Chen et al. 2010; Acquah et al. 2016b; Poletto et al. 2012 for references and a detailed description of the corresponding functional groups and vibration modes in different pine wood samples). The samples here followed a similar absorbance pattern with all the above bands observed, albeit sometimes at slightly shifted wavenumbers (mostly within 2  $\text{cm}^{-1}$ ).

A summary of spectral profiles obtained for the juvenile and mature wood samples is presented in Fig. 1. On the left (a), average juvenile and mature wood spectra, constructed from mean absorbance intensity values, are compared. The two groups produced similar profiles although juvenile wood (b) exhibited a broader variation in absorbance intensity over nearly the whole spectral range than mature wood (c), with



**Fig. 1** Standardized FTIR spectra (following total area normalization) of juvenile and mature wood samples of Scots pine constructed from mean absorbance intensity values **(a)** and minimum and maximum values (dashed lines) to illustrate the absorbance intensity range at each wavenumber for 70 juvenile **(b)** and 39 mature **(c)** wood samples



**Fig. 2** Standardized FTIR spectra following total area normalization (grey lines) of 70 juvenile **(a)** and 39 mature **(b)** wood samples of Scots pine in spectral region from 1550 to 1830  $\text{cm}^{-1}$  that encompasses the most notable differences in absorbance intensities between the two wood groups. The orange and green lines and the highlighted areas around them represent sample means and their standard deviations for given wavenumbers, respectively

the most profound differences occurring between 1570 and 1750  $\text{cm}^{-1}$ , as shown in more detail in Fig. 2. Small differences could also be observed at 1595  $\text{cm}^{-1}$  and 1423  $\text{cm}^{-1}$ , but the absorbance intensities overlapped between the two groups. Similarly, although there was a difference in mean absorbance values between 1165 and 980  $\text{cm}^{-1}$ , with local maxima at 1037  $\text{cm}^{-1}$  and 1058  $\text{cm}^{-1}$ , the value range was large and overlapping in this region (Fig. 1). Apart from bands assigned to lignin and/or polysaccharides earlier, both juvenile and mature woods produced bands at 1453  $\text{cm}^{-1}$  and 1058  $\text{cm}^{-1}$ . Additionally, about half of the juvenile wood samples exhibited a clear extra band at 1693  $\text{cm}^{-1}$  (Fig. 2a), which was completely missing in all but three samples of mature wood (Fig. 2b).

## Prediction model calibration

### Full spectra

Using standardized FTIR spectra as predictor variables and the content of chemical components of wood as response variables, predictive PLSR models were developed separately for each response variable in both juvenile and mature woods. All 70 and 39 samples, respectively, were used in model calibration and for cross-validation. All four normalization methods (TAN and AMM1–3) were tested, but only one, which provided the lowest value of minimum RMSEP, was kept for a given variable (Table 2). Each of the four normalization methods performed best for at least one of the variables. TAN was superior for predicting five variables in juvenile wood (*Cel*, *Glu*, *Man*, *Xyl* and *Ext*) and three in mature wood (*Lig*, *Man* and *Gal*); AMM1 was best for *Gal* and *Ara* in juvenile wood; AMM2 for *Lig* in juvenile wood and *Hem*, *Ara* and *Ext* in mature wood, and AMM3 for *Hem* in juvenile wood and *Cel*, *Glu* and *Xyl* in mature wood.

Following the models' diagnostics, up to four and up to three outliers were removed from the data sets of juvenile and mature wood samples, respectively, to further improve model fit. This treatment did not decrease the RMSEP substantially (it reached on average 0.613 vs. 0.683 and 0.606 vs. 0.691 for the two groups of samples, respectively), but  $R^2$  increased markedly in nearly all instances, with the highest difference being attained for *Hem* in juvenile wood (+0.463 for AMM1) and *Man* in mature wood (+0.312). Besides, while the full data set of 39 mature samples failed to produce significant models for *Hem* and *Xyl*, removing three and two outliers, respectively, enabled us to resolve the problem to a certain extent, giving raise to models with  $R^2$  of 0.658 and 0.790, albeit with relatively high RMSEP levels (0.823 and 0.845). The only variable for which we failed to develop a significant model was *Ara* in mature wood, as even after removing up to five outliers from the calibration data set, the minimum RMSEP did not drop below 1.

The overall predictive power of the present models was good, but highly variable both between juvenile and mature woods and among response variables within the two groups (Table 2). Excellent predictive powers (i.e. highly reliable predictions) were obtained for *Ext* in juvenile wood (RMSEP=0.302) and *Gal* in mature wood (0.311). Very good predictive powers were achieved for *Lig* and *Gal* in juvenile wood and *Ext* in mature wood (0.476, 0.483 and 0.471, respectively). Models for *Ara* in juvenile wood and *Hem* and *Xyl* in mature wood were considerably worse (0.812, 0.823 and 0.845, respectively). In order to improve the fit, juvenile and mature wood samples were pooled, and the models were recalibrated for all response variables (Table 2). As expected, the increased sample size ( $N=109$ ) as well as a larger variation within the samples led to a substantial improvement of the models, as the average RMSEP dropped down to 0.481 (versus 0.613 and 0.606 for models exclusive to juvenile and mature woods, respectively), with only three values slightly exceeding 0.5 (*Lig*, *Xyl*, *Ara*) and one staying near 0.8 (*Hem*) (Table 2). This indicates that models calibrated and cross-validated using a larger sample size might offer a greater statistical power for future predictions.

**Table 2** Partial least squares regression models' statistics for the chemical composition of wood using standardized FTIR spectra

Chemical component	Wood part	Normalization type <sup>a</sup>					All samples					Removed outliers				
		Wood part	Normalization type <sup>a</sup>	Min RMSEP <sup>b</sup>	Factors <sup>c</sup>	RMSEP	R <sub>x</sub> <sup>2</sup>	R <sub>y</sub> <sup>2</sup>	Removed outliers	Min RMSEP <sup>b</sup>	Factors <sup>c</sup>	RMSEP	R <sub>x</sub> <sup>2</sup>	R <sub>y</sub> <sup>2</sup>		
Lignin	J	AMM2	0.524	1 (5)	0.548	0.483	0.757	2	0.445	1 (6)	0.476	0.485	0.811			
Cellulose	J	TAN	0.715	7 (7)	0.715	0.840	0.806	0								
Hemicelluloses	J	AMM3	0.872	2 (2)	0.872	0.796	0.423	4 <sup>#</sup>	0.688	5 (7)	0.719	0.828	0.778			
Glucose	J	TAN	0.711	7 (7)	0.711	0.841	0.805	2	0.661	8 (8)	0.661	0.860	0.845			
Mannose	J	TAN	0.712	6 (7)	0.732	0.810	0.762	1	0.603	6 (10)	0.617	0.810	0.815			
Xylose	J	TAN	0.796	7 (8)	0.828	0.827	0.771	1	0.713	8 (10)	0.731	0.855	0.809			
Galactose	J	AMM1	0.526	7 (11)	0.588	0.891	0.902	1	0.472	9 (11)	0.483	0.929	0.934			
Arabinose	J	AMM1	0.781	3 (7)	0.848	0.767	0.536	1	0.754	5 (7)	0.812	0.857	0.679			
Extractives <sup>e</sup>	J	TAN	0.277	4 (9)	0.302	0.725	0.953	0								
Lignin	M	TAN	0.577	2 (3)	0.612	0.635	0.753	2	0.495	2 (2)	0.495	0.630	0.832			
Cellulose	M	AMM3	0.823	1 (1)	0.823	0.633	0.489	3	0.696	1 (1)	0.696	0.642	0.622			
Hemicelluloses	M	AMM2	1.067	0 (5)				3	0.823	5 (5)	0.823	0.894	0.658			
Glucose	M	AMM3	0.798	1 (1)	0.798	0.633	0.524	3	0.672	1 (1)	0.672	0.642	0.653			
Mannose	M	TAN	0.651	1 (5)	0.762	0.334	0.592	2	0.537	5 (5)	0.537	0.879	0.904			
Xylose	M	AMM3	1.102	0 (0)				2	0.845	7 (7)	0.845	0.957	0.790			
Galactose	M	TAN	0.354	6 (8)	0.408	0.904	0.958	1	0.311	7 (7)	0.311	0.916	0.976			
Arabinose	M	AMM2	1.036	0 (3)												
Extractives	M	AMM2	0.723	1 (2)	0.745	0.272	0.661	3	0.471	2 (2)	0.471	0.722	0.858			
Lignin	C	TAN	0.507	6 (6)	0.507	0.816	0.832	0								
Cellulose	C	AMM1	0.392	4 (6)	0.401	0.829	0.883	0								
Hemicelluloses	C	AMM1	0.867	3 (6)	0.899	0.787	0.402	4	0.756	4 (7)	0.799	0.817	0.568			
Glucose	C	TAN	0.388	4 (7)	0.407	0.750	0.880	0								
Mannose	C	AMM3	0.440	6 (8)	0.450	0.927	0.869	1	0.416	7 (7)	0.416	0.938	0.891			
Xylose	C	TAN	0.537	7 (11)	0.556	0.861	0.826	1	0.513	7 (11)	0.529	0.861	0.834			

Table 2 (continued)

Chemical component	Wood part	Normalization type <sup>a</sup>	All samples				Removed outliers				
			Min RMSEP <sup>b</sup>	Factors <sup>c</sup>	RMSEP	$R_x^2$	$R_y^2$	Min RMSEP <sup>b</sup>	Factors <sup>c</sup>	RMSEP	$R_x^2$
Galactose	C	AMM1	0.435	11 (11)	0.435	0.955	0.927	0			
Arabinose	C	AMM2	0.508	8 (9)	0.526	0.934	0.859	0			
Extractives <sup>f</sup>	C	AMM2	0.289	5 (9)	0.307	0.854	0.952	0			

The three groups of models are based on juvenile (J), mature (M) and pooled (C) wood samples (70, 39 and 109 samples, respectively)

<sup>a</sup>TAN, total area normalization; AMM1–3, area minimum–maximum normalization according to spectral regions 925–1145  $\text{cm}^{-1}$ , 1487–1553  $\text{cm}^{-1}$  and 1553–1633  $\text{cm}^{-1}$ , respectively

<sup>b</sup>Minimum root mean squared error of predictions

<sup>c</sup>Number of significant factors retained following Van der Voet's randomization-based model comparison test (number of factors corresponding to minimum RMSEP in parentheses);  $R_x^2$  and  $R_y^2$ : predictor and response variation explained

<sup>d</sup>Results not shown if identical to models based on all samples

<sup>e</sup>:<sup>f</sup>Sample size 69 and 108, respectively

<sup>g</sup>Using normalization AMM1

In most cases, the  $R^2$  values exceeded 0.8 (Table 2) (or a ratio of performance to deviation (RPD) equivalent of 2.24; see Minasny and McBratney 2013 for details). In juvenile wood, it ranged from 0.679 for *Ara* to 0.953 for *Ext* and reached on average  $0.826 \pm 0.081$  SD for the nine variables. The situation was similar in mature wood [average (AVG)  $0.787 \pm 0.130$  standard deviation (SD)], although  $R^2$ s did not reach 0.8 for four variables (*Cel*, *Hem*, *Glu* and *Xyl*). The models also tended to exploit most of the predictor variation, utilizing 48.5–92.9% (AVG  $79.9 \pm 12.9\%$  SD) and 63.0–95.7% (AVG  $78.5 \pm 14.0\%$  SD) of variation in juvenile and mature wood spectra, respectively.

The number of significant factors retained following Van der Voet's test ranged from 1 to 9 in juvenile wood and from 1 to 7 in mature wood. In ten instances, the number of factors was the same for the minimum RMSEP and Van der Voet's test selection criteria. In the remaining eight, two to five factors were non-significant and thus excluded from the models. Note that while three factors corresponding to the minimum RMSEP of 1.036 were obtained for *Ara* in mature wood, none of them was significant. The greatest difference (five factors) occurred for *Lig* and *Ext* in juvenile wood, but the  $R^2$  was only marginally lower due to the exclusion, reduced by 9% and 3% (0.811 vs. 0.902 and 0.953 vs. 0.978), respectively.

### Diagnostic band positions

Absorbance intensities at 13 band positions (wavenumbers) listed in the *Full spectra* section are considered to be good indicators of the presence of lignin and cellulose in wood materials. To test how much information they carry in relation to the two major wood components in the present samples, the models were run calibrated with all 570 variables above, using only intensities at these selected 13 wavenumbers (9 for polysaccharides and 9 for lignin, of which 5 were shared), but keeping the respective normalization selected during calibration for each response variable. The results did not differ markedly from the models where all 570 wavenumbers were included. Only a small decrease in  $R^2$  and increase in RMSEP in *Lig* and a slightly bigger change for *Cel*, in particular in juvenile wood, were observed. It confirms that the 13 selected variables capture most of the variation pertaining to the respective wood components (Table 3). However, when the nine wavenumbers assigned to cellulose were applied to predict the content of hemicelluloses, we failed to obtain reliable models for either of the two wood groups. In juvenile wood, the  $R^2$  was less than a half of that attained by all predictors and the RMSEP reached nearly 1. In mature wood, no significant factors at all were obtained. On the other hand, the content of extractives attained a high predictability using only these bands plus the three unassigned bands at  $1058 \text{ cm}^{-1}$ ,  $1453 \text{ cm}^{-1}$  and  $1693 \text{ cm}^{-1}$ .  $R^2$  remained nearly equal in juvenile wood and even increased by 1% in mature wood when all other predictors were removed; RMSEP slightly improved (0.263 vs. 0.302 and 0.435 vs. 0.471 for juvenile and mature woods, respectively).

**Table 3** Comparison of partial least squares regression models' statistics for four major chemical components of wood, using all 570 wavenumbers in the fingerprint region as predictor variables versus individual bands only (9, 9, 13 and 16 variables for lignin, cellulose, hemicelluloses and extractives, respectively)

Wood part	Juvenile wood						Mature wood					
	All			Individual bands			All			Individual bands		
Statistic	$R^2$	$F$	RMSEP	$R^2$	$F$	RMSEP	$R^2$	$F$	RMSEP	$R^2$	$F$	RMSEP
Lignin	0.812	1	0.476	0.775	1	0.522	0.860	3	0.527	0.802	3	0.556
Cellulose	0.806	7	0.715	0.598	3	0.733	0.622	1	0.696	0.612	1	0.705
Hemicelluloses	0.778	5	0.719	0.343	2	0.932	0.658	5	0.823	N/A	N/A	NA <sup>a</sup>
Extractives	0.953	4	0.302	0.951	4	0.263	0.858	2	0.471	0.871	2	0.435

$R^2$ , coefficient of determination;  $F$ , number of significant factors based on Van der Voet's randomization-based model comparison test; RMSEP, root mean squared error of predictions

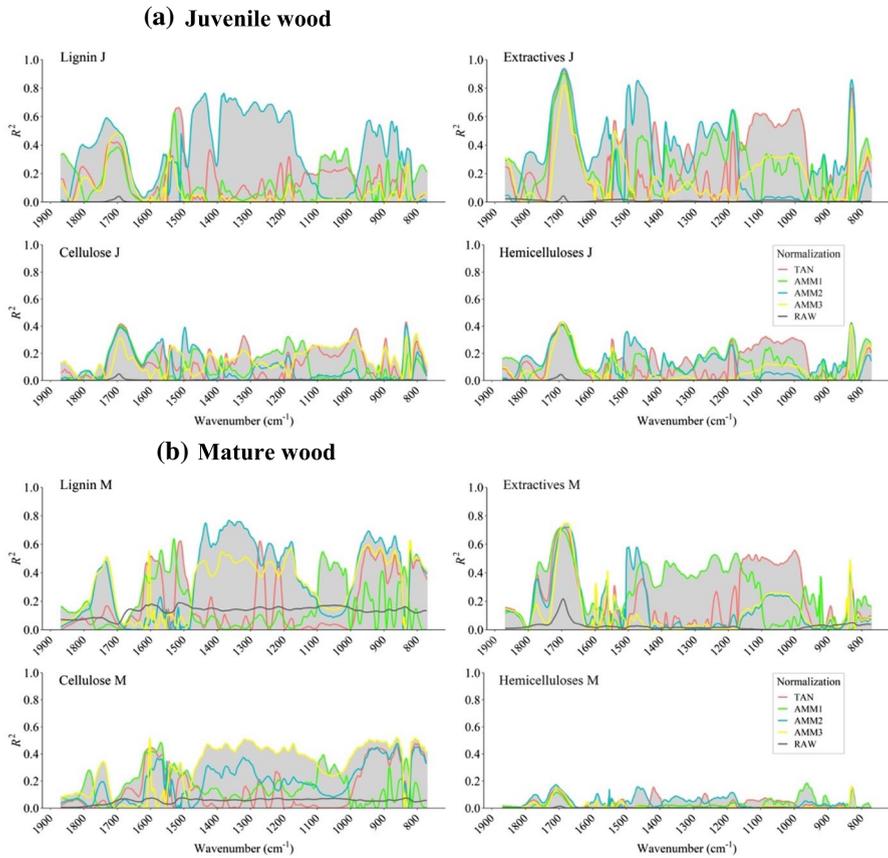
<sup>a</sup>Minimum RMSEP = 1.058

### Regression using individual wavenumbers

The relationship between the intensities at individual wavenumbers and the content of the four major wood components (*Lig*, *Cel*, *Hem* and *Ext*) was further investigated. Figure 3 shows  $R^2$  values computed using univariate regression analyses (570 for each response variable) plotted against the respective wavenumbers. One set of graphs was constructed for each of the components in both juvenile and mature woods. Each graph contains colour lines representing one of the four normalization methods plus raw data, and the grey area illustrates the highest  $R^2$  for a given wavenumber. Statistical significance of each relationship was evaluated based on  $p$  values, where the  $\alpha$  level was adjusted to  $9 \times 10^{-5}$  following Šidák correction.

For all four response variables, the distribution of  $R^2$  across the whole spectral range was greatly dependent on the normalization method used (Fig. 3), although the pattern remained more or less consistent between juvenile and mature woods. For instance, in variable *Lig*, AMM2 produced a large region of  $R^2$  values near 0.6 between 1170 and 1480  $\text{cm}^{-1}$ , with a narrow but deep depression at 1400  $\text{cm}^{-1}$ , as well as in two shorter regions around 950  $\text{cm}^{-1}$  and 1730  $\text{cm}^{-1}$ . The other three normalizations were inferior for *Lig* over most of the spectral range, especially in juvenile wood. Similarly, TAN produced a longer region of high  $R^2$  (also ca 0.6) between 980 and 1130  $\text{cm}^{-1}$  and a number of narrow but clearly separated peaks at shorter wavenumbers for the variable *Ext* (and with smaller  $R^2$ s also for *Hem* in juvenile wood). AMM1 was superior between 1150 and 1460  $\text{cm}^{-1}$  for *Ext* and AMM3 over a large region from 850 to 1450  $\text{cm}^{-1}$  in *Cel*, both in mature wood.

The most distinct and, at the same time, most consistent  $R^2$  hotspot across all normalization methods and the two types of wood material was found for variable *Ext* near 1693  $\text{cm}^{-1}$ , with local maxima ranging from 0.822 and 0.938 for juvenile wood and 0.688 and 0.741 for mature wood (all  $p$  values < 0.001) (Fig. 3). This finding corresponds well to the present observations of individual FTIR spectra described earlier (presented in Fig. 1), which revealed highly variable absorbance intensities



**Fig. 3** Relationship between absorbance intensities at individual wavenumbers and the content of major wood components (lignin, cellulose, hemicelluloses and extractives) in juvenile **(a)** and mature **(b)** wood based on 70 and 39 samples, respectively.  $R^2$  values are computed using univariate regression (570 for each response variable) plotted against the respective wavenumbers. Each colour line represents one normalization method (TAN and AMM1–3), while the grey line shows results using raw spectra for comparison. The light grey area illustrates the highest  $R^2$  for a given wavenumber

at this wavenumber, particularly in juvenile wood. On the other hand, *Hem* exhibited a weak relationship with individual frequencies across the whole spectral range regardless of the normalization method used, with the highest  $R^2$  values barely exceeding 0.4 in juvenile wood (local maxima occurred at  $1693\text{ cm}^{-1}$  and  $831\text{ cm}^{-1}$  with  $R^2$ s of 0.425 and 0.431, respectively) and not even reaching 0.2 in mature wood. The corresponding correlation coefficients that reveal the direction of the above-described relationships are presented in Figure S1 in Online Source.

## Discussion

### Prediction model evaluation

Using FTIR spectra as predictor variables and chemical composition (content of major components or their groups) as response variables, predictive PLSR models for juvenile and mature woods were developed using 70 and 39 samples, respectively. Their performance was evaluated based on the minimum value of the root mean square error of prediction (RMSEP), a measure of the average accuracy of prediction of new observations, i.e. of the difference between the true and estimated values.  $R^2$  was also used as a secondary measure (Table 2). While  $R^2$  is a good indicator of how well a model fits actual data from which it is constructed, the reliability and predictive ability of a model are generally better assessed with the aid of appropriate cross-validation statistics. Aside from the RMSEP, the most widely used statistics for evaluating the performance of predictive models are the standard error of prediction (SEP), which measures the precision of the predictions (i.e. the difference between repeated measurements); the ratio of performance to deviation (RPD), which is the ratio of the standard error in prediction to the standard deviation; and the bias, which is the average difference between the predicted and real values, indicating under- or overestimation (Acquah et al. 2015; Chen et al. 2010; Kutner et al. 2005; Zhou et al. 2015).

In the present study, the RMSEP statistics varied across the studied chemical components, from very high for extractives in juvenile wood and galactose in mature wood, through high to moderate for most variables in both groups down to very low for hemicelluloses, despite  $R^2$ s being high across variables and wood types (Table 2). The standardization of raw spectra (consisting of four steps: (1) trimming the spectral region; (2) baseline correction; (3) normalization; and (4) smoothing), generally had a positive effect on model performance, although the minimum values of RMSEP and the corresponding  $R^2$  attained following standardization were not necessarily superior to those obtained using raw spectra. For some variables, the predictive models yielded a slightly lower RMSEP when raw spectra were utilized and thus performed apparently better than those constructed using their standardized counterparts. However, with the exception of *Cel* and *Glu* in juvenile wood and *Hem* and *Man* in mature wood, the numbers of retained factors were greater with raw spectra (on average by 2.9 and 2.3 factors per response variable in juvenile and mature woods, respectively). This is likely to be the result of accounting for the baseline shift along with the other standardization steps. Thus, the models obtained with raw spectra were more complex, possessing an increased risk of overfitting. The unstructured patterns of  $R^2$  and  $r$  coefficients obtained from raw spectra (grey lines in Figs. 3 and S1, respectively), which follow nearly straight lines along the whole spectral region, confirm that raw spectra are suboptimal and that a suitable pre-processing procedure, conducted with the aim of making spectra compatible with one another, is desired when FTIR-based predictive models are constructed (see Conrad and Bonello 2016 for a review). For instance, a first derivative treatment (Owen 1995) substantially decreased RMSEP values across response variables in a study by

Zhou et al. (2015), in particular in extractives (from 1.19 to 0.34) and in lignin (from 1.05 to 0.50), with superior results reported by Acquah et al. (2016b) as well. In the present case, the standardization accentuated differences in IR absorbance intensities among individuals at most wavenumbers for the four variables *Lig*, *Cel*, *Hem* and *Ext* in both juvenile and mature woods (Figs. 3 and S1). Therefore, untreated spectra were not used for further calibration of the models and one normalization method among TAN and AMM1–3 was chosen that performed best for each respective response variable (Table 2). The differences between the methods in terms of the minimum RMSEP were often just marginal (Table S1), on average only 1.2% and 2.0% between the first- and second-best methods and 5.1% and 4.4% between the best and worst methods in juvenile and mature woods, respectively. Thus, model performances would not be severely affected if only one normalization method was applied to all response variables and both wood types. It cannot be excluded that normalization of FTIR spectra according to other regions than the three tested in the present study, or other normalization types (e.g. point maximum or offset) could lead to even higher predictive power of the calibration models. However, these were not tested, as the results were satisfying (except for arabinose in mature wood) with these common and spectroscopically accepted procedures. The performance of the four normalization methods remained consistent between models constructed using full data sets and after removing outliers. The only exception was variable *Hem* in juvenile wood, where AMM3 provided the best fit using all 70 samples, but turned to be inferior when four outliers were removed. In this case, AMM1 attained 17% lower RMSEP and 66% higher  $R^2$  than AMM3. As to the trimming step, excluding data outside of the fingerprint region (i.e. 5200–1870 and 770–400  $\text{cm}^{-1}$ ) did not compromise the predictive ability of the models, similarly to what was reported by Acquah et al. (2016b). Nearly all models' fit further improved when all 109 samples (representing the two wood types) were pooled prior to calibration.

The present models had high predictive powers for extractives and lignin, moderate for cellulose and low for hemicelluloses (along with some of their structural monosaccharides), which is in congruence with other studies utilizing FTIR spectroscopy for similar purposes. For instance, Zhou et al. (2015) obtained robust models for extractives and lignin (RMSEP = 0.34 and 0.50, respectively) and acceptable for cellulose (0.80), but the predictive ability for hemicelluloses was problematic (1.90), despite the  $R^2$  being very high for this component (0.929), indicating a good fit of actual observations. Similarly, the models reported by Acquah et al. (2016b) were good for the first three components (RPD = 2.83, 2.04 and 1.61, respectively), but unreliable for hemicelluloses (along with mannose and galactose), with RPD values remaining below 1.0. Results from some other studies using NIR spectroscopy (Acquah et al. 2015; Jones et al. 2006), which could be considered as an alternative or complementary method to FTIR, also showed that predicting hemicelluloses with reasonably high accuracies using rapid, non-destructive spectroscopic techniques remains a challenge. In these studies, predictions of hemicelluloses were poor too, with RPD values barely reaching 1.0.

## Individual wavenumber diagnostics

Spectral profiles produced by the present wood samples were similar to those reported in earlier studies (e.g. Chen et al. 2010; Müller et al. 2009; Pandey and Pitman 2003; Poletto et al. 2012). Most of the bands were observed at or near (usually  $\pm 2$   $\text{cm}^{-1}$ ) to previously assigned wavenumbers. However, hereby three extra bands at 1453  $\text{cm}^{-1}$ , 1058  $\text{cm}^{-1}$  and 1693  $\text{cm}^{-1}$  (Figs. 2, 3) are reported, which have not been discussed in connection with major wood components. The 1453  $\text{cm}^{-1}$  band is related to C–H vibrations (e.g. in methyl or methylene groups), but is rather unspecific and can be seen in both polysaccharidic and aromatic compounds (Dokken et al. 2005; Gorzsás and Sundberg 2014). The band at 1058  $\text{cm}^{-1}$  is most likely to originate from sugar ring vibrations of polysaccharides (particularly from glucosidic residues) (Gorzsás and Sundberg 2014; Toole et al. 2004; Wilson et al. 2000).

The band at 1693  $\text{cm}^{-1}$  produced the highest variation in absorbance intensities among the present samples. This was particularly true for juvenile wood, and the high and significant correlations between absorption intensities at this wavenumber and the content of extractives (Figure S1), approaching 1.0 and exceeding 0.8 in juvenile and mature woods, respectively, indicate a strong potential association. This hypothesis is in accordance with results obtained from wet chemistry analyses presented in Table 1, where a higher content of extractives was found in juvenile than in mature wood (the difference being on average ca 5% or 3.8-fold). In addition, a much larger variation in this trait was observed among individual juvenile wood samples (range 2.0–19.8% versus 0.7–6.7% in mature wood samples). Furthermore, while the four normalization methods tested in the present study differed greatly in terms of the predictive power of PLSR models for most regions of the FTIR spectra in all four major wood chemical components, the band at 1693  $\text{cm}^{-1}$  consistently provided a high and positive correlation with the content of extractives regardless of the normalization used (0.907–0.968 and 0.830–0.861 for juvenile and mature woods, respectively). Raw spectra performed poorly as single-band predictors of the chemical composition over the whole spectral regions. Even the 1693  $\text{cm}^{-1}$  band, which produced global correlation maxima in both juvenile and mature woods for the extractive content in raw spectra too, yielded much lower values than from standardized spectra (only 0.206 and 0.460, respectively). To verify the apparent relationship between extractive content and the 1693  $\text{cm}^{-1}$  band, we turned to the only three mature wood samples which, like many of the juvenile samples, exhibited this extra band (Fig. 2b). It was found that these were indeed relatively rich in extractives, in essence being the only ones with extractive content exceeding 4% (6.7%, 5.0% and 4.8% versus the average value of  $1.9\% \pm 1.25$  SD). It is therefore suggested that this band could be utilized as a strong indicator of the presence of extractives in wood, at least of the group of extractives *sensu lato*.

The collective predictive power of the band positions that had been earlier assigned to lignin and/or cellulose (or at least to polysaccharidic compounds) in wood was overall high and, with the exception of hemicelluloses, approached values attained by the whole fingerprint spectral region for the respective components. Thus, these band positions can be collectively regarded as reliable predictors of lignin and cellulose content [descriptions of the chemical bonds and their

respective vibration modes for these compounds have been assigned, e.g. by Hergert (1971), Faix (1992) and Popescu et al. (2009)]. However, other positions/regions are required for raising the predictive power for hemicellulose content to reasonable levels. Unfortunately, in the authors' experience the reliability of the models was questionable for hemicellulose content even when all variables (all 570 wavenumbers in the fingerprint region) were included in the PLS regression. Acquah et al. (2016b) reported that extending the number of predictor variables included in the PLSR models from the fingerprint region to full spectra did not bring any notable improvement in predicting this component either; in fact, it resulted in increased cross-validation errors and lower RPDs for all models except for lignin.

On the other hand, the content of extractives seems to be highly predictable from these bands alone (plus the three bands discussed earlier, in particular the band at  $1693\text{ cm}^{-1}$ ). This result may seem surprising as these bands were previously assigned to wood components other than extractives. One key factor can be the effect of normalization: Since spectra are scaled, all intensity values (and therefore the deduced concentrations) are relative, not absolute (i.e. proportions, not absolute concentrations). This in turn means that extractives are explained indirectly, i.e. when different polysaccharides and lignin are removed, the remaining proportions are assigned to extractives. Thus, precise estimations of lignin and cellulose indirectly help the precision of extractives. Furthermore, when many collinear variables are removed whereby the predictive power for the main components improves, the remaining proportion indirectly becomes more accurate as well. The impact of this "spillover" effect may, however, be limited, especially in cases when normalization is not performed by the total area. Thus, the possibility that certain bands previously exclusively assigned to lignin and cellulose are in fact not as diagnostic as previously believed cannot be excluded, i.e. they can originate from a number of other compounds included in the large group of extractives too.

## Conclusion

This study confirms that FTIR spectroscopy can serve as an effective tool for rapid, non-destructive identification of chemical compositional differences between juvenile and mature woods in Scots pine. The partial least squares regression models constructed using standardized FTIR spectra provided high statistical power for predicting the content of lignin, cellulose and extractives in juvenile and mature wood samples. On the other hand, it appears that FTIR spectroscopy would benefit from a complementary technique to improve the prediction accuracy for the content of hemicelluloses and some of their structural monosaccharides.

It was observed that the band at  $1693\text{ cm}^{-1}$ , in particular among juvenile wood samples, was strongly associated with the content of wood extractives. It might be of practical interest to investigate this feature in more detail, as wood extractives play an important role in many end-uses and serve, for instance, as natural preservatives of wood materials, natural fungicides (Hart 1981; Pearce 1996), or might be utilized as an appealing source of bioenergy, for example for vehicles (Panithasan et al. 2019).

**Acknowledgements** The authors thank Skogforsk (the Forestry Research Institute of Sweden) for access to the progeny tests, the Chemical Biological Centre (KBC) and the Department of Chemistry at Umeå University for the support to the Vibrational Spectroscopy Core Facility and Sandu Seres and members of the Forest Genetics Lab at Umeå Plant Science Centre, SLU, for help with sample preparation. Financial support from FORMAS, Grant Number 2014-00171, is greatly appreciated. Open access funding was provided by the Swedish University of Agricultural Sciences.

## Compliance with ethical standards

**Conflict of interest** The authors declare no conflict of interest. The funder had no role in the design of the study; in the collection, analyses or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

## References

- Acquah GE, Via BK, Fasina OO, Eckhardt LG (2015) Non-destructive prediction of the properties of forest biomass for chemical and bioenergy applications using near infrared spectroscopy. *J Near Infrared Spectrosc* 23:93–102. <https://doi.org/10.1255/jnirs.1153>
- Acquah GE, Via BK, Billor N, Fasina OO, Eckhardt LG (2016a) Identifying plant part composition of forest logging residue using infrared spectral data and linear discriminant analysis. *Sensors* 16:1375. <https://doi.org/10.3390/s16091375>
- Acquah GE, Via BK, Fasina OO, Eckhardt LG (2016b) Rapid quantitative analysis of forest biomass using Fourier transform infrared spectroscopy and partial least squares regression. *J Anal Methods Chem*. <https://doi.org/10.1155/2016/1839598>
- Åkerholm M, Salmén L (2001) Interactions between wood polymers studied by dynamic FT-IR spectroscopy. *Polymer* 42:963–969. [https://doi.org/10.1016/s0032-3861\(00\)00434-1](https://doi.org/10.1016/s0032-3861(00)00434-1)
- Baratieri M, Baggio P, Fiori L, Grigianti A (2008) Biomass as an energy source: thermodynamic constraints on the performance of the conversion process. *Biores Technol* 99:7063–7073. <https://doi.org/10.1016/j.biortech.2008.01.006>
- Barnett J, Jeronimidis G (2009) *Wood quality and its biological basis*. Wiley, Hoboken
- Bjarnestad S, Dahlman O (2002) Chemical compositions of hardwood and softwood pulps employing photoacoustic Fourier transform infrared spectroscopy in combination with partial least-squares analysis. *Anal Chem* 74:5851–5858. <https://doi.org/10.1021/ac025926z>
- Bowley HJ, Gerrard DL, Loudon JD, Turrell G (2012) *Practical Raman spectroscopy*. Springer, Berlin
- Burdon RD, Kibblewhite RP, Walker JCF, Megraw RA, Evans R, Cown DJ (2004) Juvenile versus mature wood: a new concept, orthogonal to corewood versus outerwood, with special reference to *Pinus radiata* and *P-taeda*. *For Sci* 50:399–415
- Chen H, Ferrari C, Angiuli M, Yao J, Raspi C, Bramanti E (2010) Qualitative and quantitative analysis of wood samples by Fourier transform infrared spectroscopy and multivariate analysis. *Carbohydr Polym* 82:772–778. <https://doi.org/10.1016/j.carbpol.2010.05.052>
- Conrad AO, Bonello P (2016) Application of infrared and Raman spectroscopy for the identification of disease resistant trees. *Front Plant Sci* 6:1152. <https://doi.org/10.3389/fpls.2015.01152>

- Cozzolino D (2014) Use of infrared spectroscopy for in-field measurement and phenotyping of plant properties: instrumentation, data analysis, and examples. *Appl Spectrosc Rev* 49:564–584. <https://doi.org/10.1080/05704928.2013.878720>
- Diem M (1993) Introduction to modern vibrational spectroscopy. Wiley, New York
- Dokken KM, Davis LC, Marininkovic NS (2005) Use of infrared microspectroscopy in plant growth and development. *Appl Spectrosc Rev* 40:301–326. <https://doi.org/10.1080/05704920500230898>
- Eilers PHC (2004) Parametric time warping. *Anal Chem* 76:404–411. <https://doi.org/10.1021/ac034800e>
- Eilers PHC, Boelens HFM (2005) Baseline correction with asymmetric least squares smoothing. Leiden University Medical Centre Report 11
- Ekeberg D, Flaete PO, Eikenes M, Fongen M, Naess-Andresen CF (2006) Qualitative and quantitative determination of extractives in heartwood of Scots pine (*Pinus sylvestris* L.) by gas chromatography. *J Chromatogr A* 1109:267–272. <https://doi.org/10.1016/j.chroma.2006.01.027>
- Faix O (1992) Fourier transform infrared spectroscopy. In: Lin SY, Dence CW (eds) *Methods in lignin chemistry*. Springer, Berlin
- Fries A, Ericsson T, Gref R (2000) High heritability of wood extractives in *Pinus sylvestris* progeny tests. *Can J For Res* 30:1707–1713. <https://doi.org/10.1139/cjfr-30-11-1707>
- Gebreselassie MN, Ader K, Boizot N, Millier F, Charpentier JP, Alves A, Simoes R, Rodrigues JC, Bodineau G, Fabbrini F, Sabatti M, Bastien C, Segura V (2017) Near-infrared spectroscopy enables the genetic analysis of chemical properties in a large set of wood samples from *Populus nigra* (L.) natural populations. *Ind Crops Prod* 107:159–171. <https://doi.org/10.1016/j.indcrop.2017.05.013>
- Gierlinger N (2018) New insights into plant cell walls by vibrational microspectroscopy. *Appl Spectrosc Rev* 53:517–551
- Gillgren T, Gorzsás A (2016) A one-pot set-up for real-time reaction monitoring by FTIR spectroscopy. *Wood Sci Technol* 50:567–580. <https://doi.org/10.1007/s00226-016-0801-9>
- Gnansounou E, Dauriat A, Villegas J, Panichelli L (2009) Life cycle assessment of biofuels: energy and greenhouse gas balances. *Biores Technol* 100:4919–4930. <https://doi.org/10.1016/j.biortech.2009.05.067>
- Gorzsás A, Sundberg B (2014) Chemical fingerprinting of *Arabidopsis* using Fourier transform infrared (FT-IR) spectroscopic approaches. In: Sanchez-Serrano J, Salinas J (eds) *Arabidopsis protocols. Methods in molecular biology*, vol 1062. Humana Press, Totowa, pp 317–352
- Hart JH (1981) Role of phytostilbenes in decay and disease resistance. *Annu Rev Phytopathol* 19:437–458. <https://doi.org/10.1146/annurev.py.19.090181.002253>
- Hergert HL (1971) Infrared spectra. In: Ludwig CH, Sarkanen KV (eds) *Lignins: occurrence, formation, structure and reactions*. Wiley-Interscience, New York
- Ivkovic M, Gapare W, Wu H, Espinoza S, Rozenberg P (2013) Influence of cambial age and climate on ring width and wood density in *Pinus radiata* families. *Ann For Sci* 70:525–534. <https://doi.org/10.1007/s13595-013-0290-z>
- Jones PD, Schimleck LR, Peter GF, Daniels RF, Clark A III (2006) Nondestructive estimation of wood chemical composition of sections of radial wood strips by diffuse reflectance near infrared spectroscopy. *Wood Sci Technol* 40:709–720. <https://doi.org/10.1007/s00226-006-0085-6>
- Kutner MH, Nachtsheim CJ, Neter J, Li W (2005) *Applied linear statistical models*. McGraw-Hill, New York
- Minasny B, McBratney A (2013) Why you don't need to use RPD. *Pedometron* 33:14–15
- Morris DR, Steward FR, Gilmore CA (2000) Comparative analysis of the consumption of energy of two wood pulping processes. *Energy Convers Manag* 41:1557–1568. [https://doi.org/10.1016/S0196-8904\(99\)00187-9](https://doi.org/10.1016/S0196-8904(99)00187-9)
- Müller G, Schöpfer C, Vos H, Kharazipour A, Polle A (2009) FTIR-ATR spectroscopic analyses of changes in wood properties during particle- and fibreboard production of hard- and softwood trees. *BioResources* 4:49–71
- Owen AJ (1995) Uses of derivative spectroscopy. UV-visible spectroscopy, application note. Agilent Technologies, pp 1–8
- Pandey KK, Pitman AJ (2003) FTIR studies of the changes in wood chemistry following decay by brown-rot and white-rot fungi. *Biodeterior Biodegrad* 52:151–160. [https://doi.org/10.1016/S0964-8305\(03\)00052-0](https://doi.org/10.1016/S0964-8305(03)00052-0)
- Panithasan MS, Gopalakichenin D, Venkadesan G, Veeraraagavan S (2019) Impact of rice husk nanoparticle on the performance and emission aspects of a diesel engine running on blends of pine oil-diesel. *Environ Sci Pollut Res* 26:282–291. <https://doi.org/10.1007/s11356-018-3601-y>

- Pearce RB (1996) Antimicrobial defences in the wood of living trees. *New Phytol* 132:203–233. <https://doi.org/10.1111/j.1469-8137.1996.tb01842.x>
- Pearson RG, Gilmore RC (1980) Effect of fast growth-rate on the mechanical-properties of loblolly-pine. *For Prod J* 30:47–54
- Pereira H, Graça J, Rodrigues JC (2003) Wood chemistry in relation to quality. In: Barnett JR, Jeronimidis G (eds) *Wood quality and its biological basis*. Blackwell Publishing Ltd., Oxford, pp 53–86
- Pettersen RC (1984) The chemical composition of wood. *Chem Solid Wood* 207:57–126
- Poletto M, Zattera AJ, Santana RMC (2012) Structural differences between wood species: evidence from chemical composition, FTIR spectroscopy, and thermogravimetric analysis. *J Appl Polym Sci* 126:E336–E343. <https://doi.org/10.1002/app.36991>
- Popescu C-M, Singurel G, Popescu M-C, Vasile C, Argyropoulos DS, Willför S (2009) Vibrational spectroscopy and X-ray diffraction methods to establish the differences between hardwood and softwood. *Carbohydr Polym* 77:851–857
- Rodríguez-Saona LE, Allendorf ME (2011) Use of FTIR for rapid authentication and detection of adulteration of food. *Ann Rev Food Sci Technol* 2:467–483. <https://doi.org/10.1146/annurev-food-022510-133750>
- Rubin EM (2008) Genomics of cellulosic biofuels. *Nature* 454:841–845. <https://doi.org/10.1038/nature07190>
- Savitzky A, Golay MJE (1964) Smoothing + differentiation of data by simplified least squares procedures. *Anal Chem* 36:1627–1639. <https://doi.org/10.1021/ac60214a047>
- Scandinavian Pulp, Paper and Board Testing Committee (2003) Protocol SCAN-CM 67:03: content of extractable lipophilic matter. Scandinavian Pulp, Paper and Board Testing Committee, Stockholm
- Scandinavian Pulp, Paper and Board Testing Committee (2009) Protocol SCAN-CM 71:09. Pulps—carbohydrate composition. Scandinavian Pulp, Paper and Board Testing Committee, Stockholm
- Shebani AN, van Reenen AJ, Meincken M (2008) The effect of wood extractives on the thermal stability of different wood species. *Thermochim Acta* 471:43–50. <https://doi.org/10.1016/j.tca.2008.02.020>
- Sjöström E (1993) *Wood chemistry: fundamentals and applications*. Academic Press, San Diego
- Strunk P, Oman T, Gorzsas A, Hedenstrom M, Eliasson B (2011) Characterization of dissolving pulp by multivariate data analysis of FT-IR and NMR spectra. *Nord Pulp Pap Res J* 26:398–409
- Sykes R, Isik F, Li BL, Kadla J, Chang HM (2003) Genetic variation of juvenile wood properties in a loblolly pine progeny test. *Tappi J* 2:3–8
- TAPPI (1991) TAPPI UM 250: acid-soluble lignin in wood and pulp. TAPPI test methods. Tappi Press, Atlanta
- TAPPI (2002) TAPPI T 222 om-02: acid-insoluble lignin in wood and pulp. TAPPI test methods. Tappi Press, Atlanta
- Toole GA, Kacurakova M, Smith AC, Waldron KW, Wilson RH (2004) FT-IR study of the *Chara corallina* cell wall under deformation. *Carbohydr Res* 339:629–635. <https://doi.org/10.1016/j.carres.2003.11.010>
- Tsuchikawa S, Kobori H (2015) A review of recent application of near infrared spectroscopy to wood science and technology. *J Wood Sci* 61(3):213–220. <https://doi.org/10.1007/s10086-015-1467-x>
- Van der Voet H (1994) Comparing the predictive accuracy of models using a simple randomization test. *Chemom Intell Lab Syst* 25:313–323. [https://doi.org/10.1016/0169-7439\(94\)85050-x](https://doi.org/10.1016/0169-7439(94)85050-x)
- Wilson RH, Smith AC, Kacurakova M, Saunders PK, Wellner N, Waldron KW (2000) The mechanical properties and molecular dynamics of plant cell wall polysaccharides studied by Fourier-transform infrared spectroscopy. *Plant Physiol* 124:397–405. <https://doi.org/10.1104/pp.124.1.397>
- Zhou CF, Jiang W, Cheng QZ, Via BK (2015) Multivariate calibration and model integrity for wood chemistry using Fourier transform infrared spectroscopy. *J Anal Methods Chem*. <https://doi.org/10.1155/2015/429846>

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

## Affiliations

**Tomas Funda**<sup>1,2,3</sup>  · **Irena Fundova**<sup>1,4</sup> · **András Gorzsás**<sup>5</sup> · **Anders Fries**<sup>1</sup> · **Harry X. Wu**<sup>1,6,7</sup>

- <sup>1</sup> Umeå Plant Science Centre, Department of Forest Genetics and Plant Physiology, Swedish University of Agricultural Sciences, Umeå, Sweden
- <sup>2</sup> Department of Genetics and Breeding, Faculty of Agrobiological and Natural Resources, Czech University of Life Sciences Prague, Prague, Czech Republic
- <sup>3</sup> Key Laboratory of Forest Genetics and Biotechnology, Co-Innovation Center for the Sustainable Forestry in Southern China, Nanjing Forestry University, Nanjing, China
- <sup>4</sup> Skogforsk (The Forestry Research Institute of Sweden), Sävar, Sweden
- <sup>5</sup> Department of Chemistry, Umeå University, Umeå, Sweden
- <sup>6</sup> Beijing Advanced Innovation Center for Tree Breeding by Molecular Design, Beijing Forestry University, Beijing, China
- <sup>7</sup> National Research Collections Australia, CSIRO, Canberra, Australia