

1 **This is a pre-referred version of the paper**  
2 **published in Cellulose (2016) 23:901-913**  
3 **DOI 10.1007/s10570-015-0848-z**

4  
5 **Application of chemometric analysis to**  
6 **infrared spectroscopy for the identification of**  
7 **wood origin**

8 Ara Carballo-Meilán†, Adrian M. Goodman ‡, Mark G. Baron\*, Jose Gonzalez-  
9 Rodriguez\*

10 † *Department of Chemical Engineering of the University of Loughborough,*  
11 *Loughborough, LE11 3TU, UK*

12 ‡ *School of Life Sciences of the University of Lincoln, Brayford Pool, Lincoln,*  
13 *LN6 7TS, UK*

14 \**School of Chemistry of the University of Lincoln, Brayford Pool, Lincoln, LN6*  
15 *7TS, UK*

16 Telephone: +441522886878

17 E-mail: [jgonzalezrodriguez@lincoln.ac.uk](mailto:jgonzalezrodriguez@lincoln.ac.uk)

18

19 Chemical characteristics of wood are used in this study for plant taxonomy classification based on  
20 the current Angiosperm Phylogeny Group classification (APG III System) for the division, class  
21 and subclass of woody plants. Infrared spectra contain information about the molecular structure  
22 and intermolecular interactions among the components in wood but the understanding of this  
23 information requires multivariate techniques for the analysis of highly dense datasets. This article  
24 is written with the purposes of specifying the chemical differences among taxonomic groups, and  
25 predicting the taxa of unknown samples with a mathematical model. Principal component analysis,  
26 t-test, stepwise discriminant analysis and linear discriminant analysis, were some of the chosen  
27 multivariate techniques. A procedure to determine the division, class, subclass, order and family of  
28 unknown samples was built with promising implications for future applications of Fourier  
29 Transform Infrared spectroscopy in wood taxonomy classification

30 *Plant taxonomy classification, Infrared spectroscopy, Multivariate analysis, Wood,*  
31 *Angiosperm, Gimnosperm*

## 32 **Introduction**

33 Trees belong to seed-bearing plants which are subdivided into two major  
34 botanical groupings: Gymnosperms (*Gymnospermae*) and Angiosperms  
35 (*Angiospermae* or flowering plants). Coniferous woods or softwoods belong to the  
36 first-mentioned category and hardwoods to the second group (Sjostrom, 1981).  
37 These groups are subdivided into class, subclass, orders, families, genera and  
38 species based on the current Angiosperm Phylogenetic System Classification  
39 (APG III System). Traditional methods of botanical classification include a  
40 taxonomic system based on structural and physiological connections between  
41 organisms and a phylogenetic system, based on genetic connections.  
42 The method of “chemical taxonomy” consists of the investigation of the  
43 distribution of chemical compounds in series of related or supposedly related  
44 plants (Erdtman, 1963). Taxonomically, the species are difficult to classify  
45 because there is great inter-species variability as well as narrow gaps between the  
46 morphological characteristics of different species (Gidman et al., 2003). The  
47 chemical composition of softwoods (gymnosperms) differs from that of  
48 hardwoods (angiosperms) in the structure and content of lignin and  
49 hemicelluloses. Generally speaking Gymnosperms have less hemicelluloses and  
50 more lignin (Martin, 2007). In hardwood the predominant hemicellulose is a  
51 partially acetylated xylan with a small proportion of glucomannan. In softwoods,  
52 the main hemicellulose is partially acetylated galactoglucomannan and  
53 arabinoglucuronoxylan (Barnett and Jeronimidis, 2003; Ek et al., 2009). The  
54 composition of xylans from various plants appears as well to be related to their  
55 belonging to evolutionary families (Ek et al., 2009). With regards to lignin,  
56 softwoods mainly contains only guaiacyl *lignin*, while hardwood contains both  
57 guaiacyl (G) and syringyl (S) lignin and the syringyl/guaiacyl (S/G) ratio varies  
58 among species (Barnett and Jeronimidis, 2003; Obst, 1982; Stewart et al., 1995;  
59 Takayama, 1997) (e.g. species of the same genus can show a large variation in the  
60 S/G ratio (Barnett and Jeronimidis, 2003)).  
61 Fourier transform infrared spectroscopy (FTIR) is a non-destructive technique  
62 suitable for representations of phylogenetic relationships between plant taxa, even  
63 those that are closely related (Shen et al., 2008). An advantage is that it can be  
64 applied in the analysis of wood without pre-treatment, thus avoiding the tedious  
65 methods of isolation which are normally required (Åkerholm et al., 2001; Obst,

66 1982). Infrared spectroscopy is quite extensively applied in plant cell wall  
67 analysis (Kacuráková et al., 2000). Furthermore, in combination with multivariate  
68 analysis, FTIR has been used for the chemotaxonomic classification of flowering  
69 plants, for example: the identification and classification of the *Camellia* genus  
70 using cluster analysis and Principal Component Analysis (PCA) (Shen et al.,  
71 2008); the taxonomic discrimination of seven different plants that belong to two  
72 orders and three families using a dendrogram based on PCA (Kim et al., 2004);  
73 and the differentiation of plants from different genera using cluster analysis  
74 (Gorgulu et al., 2007). In woody tissues, FTIR has been used to characterize  
75 lignin (Obst, 1982; Takayama, 1997), characterise soft and hardwood pulps using  
76 Partial Least-Squares analysis (PLS) and PCA (Bjarnestad and Dahlman, 2002).  
77 In addition, the interaction of wood polymers using Partial Least-Squares  
78 regression (Åkerholm et al., 2001) and differentiation of wood species using  
79 Partial Least-Squares regression (Hobro et al., 2010) has also been investigated.  
80 This paper reports on the chemical differences between wood samples using  
81 spectral data and multivariate analysis. To the best of our knowledge, this is the  
82 first time that unknown samples from trees have been successfully classified into  
83 division, class, subclass, as well as, order and family through a linear model based  
84 on the chemical features of wood using FTIR spectroscopy.

## 85 **Materials and Methods**

86 Branch material was collected from 21 tree species in Lincoln (Lincolnshire, UK).  
87 Five Gymnosperm trees and 16 Angiosperm trees (12 from Rosids class and 4  
88 from Asterids class) were analysed. Table 1 provides a detailed description of the  
89 samples. The samples were stored in a dry environment at ambient temperature  
90 conditions

### 91 **Sample preparation**

92 Sample preparation was reproduced in the same manner as described in detail in  
93 another publication (Carballo-Meilan et al., 2014). The dataset obtained from a  
94 PerkinElmer Spectrum 100 FTIR Spectrometer was integrated by 3500 variables  
95 and 252 observations recorded in pith, bark, rings and sapwood positions. Results  
96 from the ring dataset (101 observations) are shown in the present article.

## 97 **Multivariate techniques**

98 The data set was processed with Tanagra 1.4.39 software. A range of  
99 multivariable statistical methods were chosen to analyse spectra of the wood  
100 samples including: Principal Component Analysis (PCA), t-test, Stepwise  
101 Discriminant Analysis (STEPDISC) method, Partial-Least squares for  
102 Classification (C-PLS), Linear Discriminant Analysis (LDA) and PLS-LDA linear  
103 models. The statistical methodology from the previous research (Carballo-Meilan  
104 et al., 2014) was used in this work.

## 105 **Results and discussion**

### 106 **Wood spectra dataset**

107 The raw spectra of 16 wood samples that belong to the Angiosperm division and 5  
108 wood samples from the Gymnosperm division were statistically analysed. The  
109 sample size available for chemometric analysis in the division dataset was 29 and  
110 72 observations from Conifer and Angiosperm, respectively. From the total  
111 number of cases (101), 83 were assigned as training set and 18 as test set.  
112 Equivalent procedure was executed with class (74) and subclass (18) datasets; the  
113 former with 54 Rosids and 18 Asterids, and the later with 11 Euasterids I and 7  
114 Euasterids II. In the case of the class dataset, the sample was divided to give 60  
115 observations as training set and 14 as test set, and in the case of the subclass  
116 dataset 11 cases were assigned as training set and 7 as test set. Vibrational spectra  
117 from the growth rings of the wood samples are shown in Fig. 1-A, Fig. 2-A and  
118 Fig. 3-A for division, class and subclass dataset, respectively; the arrows indicate  
119 important bands in the discrimination of samples based on the STEPDISC results  
120 (See section below).

### 121 **Exploratory data analysis**

122 A PCA mathematical technique was applied to over 101 samples of individual  
123 spectra of trees to find the most relevant wavelengths, between the range 4000-  
124 500  $\text{cm}^{-1}$ , which contribute to sample discrimination between Gymnosperm versus  
125 Angiosperm divisions, Rosids versus Asterids classes and Euasterids I versus  
126 Euasterids II subclasses. The data set was standardized so each variable received  
127 equal weight in the analysis. PCA of the spectra of wood from division, class and

128 subclass dataset gave five main factor loading. Differences between groups, using  
129 the two first factors, led to poor structure of the data.

130 T-test was computed to determine which factors were more significant for  
131 differentiating groups. The factor rotated loading (FR) extracted from PCA were  
132 used for interpreting the principal components and to determine which variables  
133 are influential in the formation of PCs. Normality and homogeneity of variance  
134 was checked. Mann-Whitney test (i.e., non-parametric alternative to the t-test)  
135 was also performed, confirming the significance of the factors. The wavenumbers  
136 loading on those highlighted factors were chemically identified. In later  
137 computations, STEPDISC method confirmed the importance of those chemicals in  
138 the discrimination. The results of that probe showed that there are chemical  
139 differences between Gymnosperms and Angiosperms that were condensed only  
140 inside the fourth and fifth rotated factor (FR4 and FR5). The t-test was 2.902 with  
141 an associated probability of 0.00456 for FR4, and 4.6767 ( $p=0.000009$ ) for FR5.  
142 Then the null hypothesis may be rejected at the 99.54% and 99.99% levels for  
143 FR4 and FR5, respectively and, therefore, it is concluded that there is a significant  
144 difference in means due to the factor selected. A detailed band assignment of the  
145 factors highlighted in the t-test is presented in Table 2. Those factors seem to  
146 contain relevant meaning. The most highly correlated wavenumbers with those  
147 factors are 1762-1719, 1245-1220 and 1132-950 from FR4 and 2978-2832, 1713-  
148 1676 and 1279-1274 from FR5. As the STEPDISC method highlighted, it is  
149 highly likely that the C=O stretching in hemicelluloses and lignin, wavenumbers  
150 1730, 1712 and  $1684\text{ cm}^{-1}$  from feature selection (range 1762-1719  $\text{cm}^{-1}$  in FR4  
151 and 1713-1676  $\text{cm}^{-1}$  in FR5) play a key role in the classification.

152 In the case of Rosids vs. Asterids, the t-test emphasized FR3 and FR5 as main  
153 descriptors of the chemical differences between class. The result seems not  
154 significantly different with 95% probability for FR5 ( $t=1.7379$ ,  $p=0.0865$ ). The  
155 difference was only significant for FR3 at the 5% significant level since the p  
156 value was 0.00148 ( $t$  was 3.3062). Major contributors to the FR3 formation are  
157 wavenumber between 1171 and  $884\text{ cm}^{-1}$ , and  $2860\text{-}2847\text{ cm}^{-1}$ . The most highly  
158 correlated wavenumbers with FR5 are 1687-1385. Then the C-H ring in  
159 glucomannan, 874 and 872 (associated with FR3), and the C=O stretching and C-  
160 H deformation in lignin and carbohydrates, wavenumbers 1678, 1619, 1617, 1613  
161 and  $1438\text{ cm}^{-1}$  associated with FR5 are all important chemical signals for

162 differentiating Rosids from Asterids classes, based on PCA and STEPDISC  
163 analysis. With regards to the differences between Euasterids I and Euasterids II,  
164 FR4 was selected from the t-test analysis with a value of the probability greater  
165 than 0.05 ( $t=1.9179$ ,  $p=0.0731$ ). This factor is highly correlated with the  
166 wavenumbers 1763-1709 and 1245-1212  $\text{cm}^{-1}$ . Based on the feature selection  
167 procedure, it could be that 1769, 1701 and 1697  $\text{cm}^{-1}$  were significant for  
168 distinguishing among the subclass groups but the results were limited by the small  
169 sample size. The identity of the mentioned wavenumbers was associated with  
170 C=O stretching in hemicelluloses and lignin. The wavenumbers responsible for  
171 the classification between division, class, subclass, order and family are described  
172 in the next section (STEPDISC analysis).

173 A subset of wavenumbers from the STEPDISC method was used as input in PCA  
174 to emerge the underlined structure in division, class and subclass datasets. The  
175 scores extracted from PCA were used for interpreting the samples and the loading  
176 to determine which variables are in relation with the samples. The higher the  
177 loading of a variable, the more influence it has in the formation of the factor and  
178 vice versa. The score plot from division dataset (Fig. 1-B) showed that conifers  
179 were highly correlated with FR3, and the loading plot (Fig. 1-D) showed that the  
180 wavenumber 1684 could be related with conifers since it correlates more with its  
181 factor. A 3D plot (Fig. 1-C) with the individual observations is shown to highlight  
182 the underline structure of the dataset using the first three rotated factors. In the  
183 score plot from class dataset (Fig. 2-B), the Asterids sample correlated highly with  
184 FR2 and the Rosids sample better with FR1. The correlation plot (Fig. 2-D)  
185 suggested that the wavenumber 2031 is more highly correlated with FR2, and  
186 therefore would be more connected with the Asterids group. With respect to the  
187 subclass dataset, loading plot is shown in Fig. 3-B. In this case Euasterids I  
188 observations were positively correlated with FR2, and Euasterids II with FR1. The  
189 wavenumbers 1701, 1697 and 1769 were correlated with FR1, suggesting some  
190 closeness with Euasterids II.

### 191 **STEPDISC analysis**

192 Supervised approach, based on the Wilks' partial lambda, known as STEPDISC  
193 method was computed over the normalized wavenumbers to determine the most  
194 significant variables for the classification process. Groups based on the current

195 Angiosperm Phylogeny Group classification (APG III System) were used to find  
196 the discriminator wavenumbers. Forward strategy and computed statistic F to 3.84  
197 as statistical criterion for determining the addition of variables was chosen. The  
198 cut-off value selected as minimum conditions for selection of the variables was  
199 0.01 significant level to find the most relevant variables. Seven biomarkers (1730,  
200 1712, 1420, 3068, 1684, 1610, and 1512  $\text{cm}^{-1}$ ) were successfully found to  
201 discriminate between Angiosperms and Gymnosperms. The wavenumbers,  
202 arranged in a descendent order based on their F-values (i.e., the variable's total  
203 discriminating power, the greater contributor to the overall discrimination in the  
204 STEPDISC method will show a better F-value (Klecka, 1980)), have the  
205 following band assignment: 1730 (C=O stretching in acetyl groups of  
206 hemicelluloses (xylan/glucomannan) (Åkerholm et al., 2001; Bjarnestad and  
207 Dahlman, 2002; Gorgulu et al., 2007; Marchessault, 1962; McCann et al., 2001;  
208 Mohebbi, 2008, 2005; Rana et al., 2009; Stewart et al., 1995)), 1712(C=O stretch  
209 (unconjugated) in lignin (Hobro et al., 2010)), 1420 (aromatic ring vibration  
210 combined with C-H in-plane deformation lignin (Kubo and Kadla, 2005; Rhoads  
211 et al., 1987; Wang et al., 2009)), 3068 (C-H stretch aromatic (Larkin, 2011;  
212 Silverstein et al., 2005)), 1684 (C=O stretch in lignin (Coates, 2000; Silverstein et  
213 al., 2005; Sudiyani et al., 1999)), 1610 (aromatic skeletal vibration plus C=O  
214 stretching lignin (Kubo and Kadla, 2005; Wang et al., 2009)), and 1512 (aromatic  
215 skeletal vibration lignin(Hobro et al., 2010; Huang et al., 2008; Kubo and Kadla,  
216 2005; Wang et al., 2009)). It seems that differences between groups can be  
217 attributed to the lignin region. These spectral differences between hard and  
218 softwood lignin were observed in the fingerprint region between 1800 and 900  
219  $\text{cm}^{-1}$  by other authors (Pandey, 1999).

220 With regards to class dataset, 10 biomarkers (2031, 1678, 1619, 1617, 1613, 784,  
221 771, 874, 872, and 1438  $\text{cm}^{-1}$ ) were found to successfully discriminate between  
222 the Rosids and Asterids classes within the Angiosperm division. Differences  
223 between groups can be attributed to C=O stretching in lignin and C-H deformation  
224 in carbohydrates and lignin, based on their literature assignments (in order of  
225 greater contribution to the overall discrimination): 2031 (-N=C=S (Donald L.  
226 Pavia & Gary M. Lampman & George S. Kriz & James A. Vyvyan, 2009; Larkin,  
227 2011)), 1678 (C=O stretching aryl ketone of guaiacyl (G) (Rhoads et al., 1987)),  
228 1619, 1617, 1613 (C-O stretching of conjugated or aromatic ketones, C=O

229 stretching in flavones (Hobro et al., 2010; Huang et al., 2008)), 784 (Out of plane  
230 CH bend (Silverstein et al., 2005)), 771 (out of plane N-H wagging primary and  
231 secondary amides in carbohydrates or OH out of plane bending (Marchessault,  
232 1962; Muruganatham et al., 2009; Peter Zugenmaier, 2007)), 874, 872 (C-H ring  
233 glucomannan (Åkerholm et al., 2001; Bjarnestad and Dahlman, 2002; Kacuráková  
234 et al., 2000; Marchessault, 1962)), and 1438 (C-H deformation in Lignin and  
235 carbohydrates (Mohebbi, 2005)). Thiocyanate was also seen by other authors to  
236 discriminate among Angiosperms (Rana et al., 2009).

237 The last probe was run over subclass dataset; 5 biomarkers (1769, 1697, 3613,  
238 3610, and 1701 cm<sup>-1</sup>) were found to successfully discriminate between Euasterids  
239 I and Euasterids II subclass from Asterids class. As mentioned before, C=O  
240 stretching in lignin and carbohydrates seems relevant for the classification. The  
241 greater contributor to the discrimination between subclass groups was the  
242 wavenumber 1769, attributed in the literature to C=O stretching in acetyl groups  
243 of hemicelluloses (xylan/glucomannan) (Åkerholm et al., 2001; Bjarnestad and  
244 Dahlman, 2002; Gorgulu et al., 2007; Marchessault, 1962; McCann et al., 2001;  
245 Mohebbi, 2008, 2005; Rana et al., 2009; Stewart et al., 1995), this contributor  
246 was followed in order of importance (the second greatest F-value) by 1697  
247 assigned to C=O stretching (Coates, 2000; Silverstein et al., 2005), 3613 and 3610  
248 (O-H stretching (Coates, 2000)), and lastly 1701 related to Conj-CO-Conj lignin  
249 (Hobro et al., 2010; Larkin, 2011).

250 STEPDISC method was run over different split datasets from ring dataset, the  
251 imbalance effect on the results was also checked; in such a way, the discriminator  
252 wavenumbers from the output of STEPDISC method were selected and used to  
253 construct linear regression models.

## 254 **Linear model and validation**

255 The next step after selecting the discriminator wavenumbers was to compute and  
256 compare several linear models: C-PLS, LDA and PLS-LDA. The discrete class  
257 attribute are the taxons based on the current taxonomic classification of trees and  
258 the continuous attributes are the discriminator wavenumbers filtered through the  
259 STEPDISC previous method. Wilks's lambda is a multivariate measure of group  
260 differences over the predictors (Klecka, 1980) and it was used to measure the  
261 ability of the variables in the computed classification function from LDA to



262 discriminate among the groups. Classification was done by using the classification  
263 functions computed for each group. Observations were assigned to the group with  
264 the largest classification score (Rakotomalala, 2005). LDA gave the lowest error  
265 in the classification and was for that reason the only one shown in this work.  
266 Bias-variance error rate decomposition was used to adjust the correct number of  
267 predictors in the model to the current sample size, as describe in our previous  
268 work (Carballo-Meilan et al., 2014). As shown in Fig. 4, the optimum model in  
269 division would have 4 wavenumbers instead of 7. In the case of the class model,  
270 the overfitting region showed up above 8 and underfitting below 7. Similar  
271 approach was taken for the subclass model where 4 wavenumber were selected as  
272 the optimum model. Table 3 shows the classification functions with their  
273 statistical evaluation for division, class and subclass datasets. The coefficients of  
274 the classification functions are not interpreted. Smallest lambda values (not  
275 shown) or largest partial F means high discrimination (Klecka, 1980). The  
276 significance of the difference was checked using Multivariate Analysis of  
277 Variance (MANOVA) and two transformations of its lambda, Bartlett  
278 transformation and Rao transformation (Rakotomalala, 2005). According to Rao's  
279 transformation (for small sample sizes,  $p < 0.01$ ), it can be concluded that there is  
280 a significant difference between groups in the three cases: division (Rao-F  
281 (7,75)=46.417,  $p=0.000$ ), class (Rao-F (7,75)=21.975,  $p=0.000$ ) and subclass  
282 (Rao-F (7,75)=35.028,  $p=0.000$ ). The discriminant functions scores were plotted  
283 in Fig. 5 to show the discrimination among division, class and subclass groups.  
284 The separation looks greater in the case of class and subclass.  
285 Validation of the model was done to evaluate the statistical and the practical  
286 significance of the overall classification rate and the classification rate for each  
287 group. Cross-validation (CV), bootstrap method, leave-one-out (LOO), Wolper  
288 and Kohavi bias-variance decomposition, and an independent test set which was  
289 not used in the construction of the model (test size appears in brackets in Table 3)  
290 were used in the validation procedure. The bootstrap value shown in Table 3 is the  
291 higher error obtained by the .632 estimator and its variant .632+. This error was  
292 seen to be preferred for Gaussian population and small training samples size  
293 ( $n \leq 50$ ) (Chernick, 2011). Error rate estimation is presented to evaluate the  
294 variance explained by the model; in division, 52% bias, 47% variance, 0.0671  
295 error rate; in class, 64% bias, 36% variance, 0.1552 error rate; and in subclass,

296 57% bias, 43% variance, 0.0950 error rate. The model seems stable with a low  
297 classification error. Further validation of the method was performed with an  
298 unknown piece of wood. The division, class, subclass and order were determined  
299 correctly. The samples were taken from a willow tree and belonged to  
300 Angiosperm > Rosids > Eurosids I > Malpighiales.

## 301 **Conclusion**

302 A procedure was developed for the taxonomic classification of wood species  
303 using samples from different division, class and subclass. First, a STEPDISC  
304 method was used to select the predictor wavenumbers for classification. The  
305 chemical differences between taxonomic groups were attributed mainly to the  
306 differences in their lignin and hemicelluloses content, as well as some amide  
307 contribution. The results were also confirmed by a t-test applied on the output  
308 from PCA procedure. LDA, PLS-LDA and C-PLS linear models were computed  
309 to calculate the classification functions with the predictor variables as dependent  
310 variables and groups based on the APG III System as independent variables. LDA  
311 provided the lowest classification error based on different validation techniques  
312 such as bootstrap or LOO. For an unknown sample its division, class, subclass and  
313 order were successfully determined. This study demonstrates that spectra data  
314 obtained from wood samples have the potential to be used to discriminate trees  
315 taxonomically.

316 A scaffold for the taxonomic classification of woody plants has been produced. A  
317 procedure to statistically define differences among species and use them in a  
318 model that classifies unknown samples is possible. With additional work this may  
319 prove to be a useful tool to aid in the taxonomic classification of plants. Naturally  
320 the current models should only be applied to the species included in the model  
321 and, because of the differences in chemical composition among species, it is  
322 important that new models are developed to broaden its application.

## 323 **Acknowledgements**

324 This work was supported by Europracticum IV (Leonardo da Vinci Programme). We gratefully  
325 acknowledge to the Consello Social from Universidade de Santiago de Compostela (Spain)

326 **References**

- 327 Åkerholm, M., Salmén, L., Salme, L., 2001. Interactions between wood polymers studied by  
328 dynamic FT-IR spectroscopy. *Polymer (Guildf)*. 42, 963–969. doi:10.1016/S0032-  
329 3861(00)00434-1
- 330 Anchukaitis, K.J., Evans, M.N., Lange, T., Smith, D.R., Leavitt, S.W., Schrag, D.P., 2008.  
331 Consequences of a rapid cellulose extraction technique for oxygen isotope and radiocarbon  
332 analyses. *Anal. Chem.* 80, 2035–2041. doi:10.1016/j.gca.2004.01.006.Analytical
- 333 APG, I., 2003. An update of the Angiosperm Phylogeny Group classification for the orders and  
334 families of flowering plants: APG II. *Bot. J. Linn. Soc.* 141, 399–436. doi:10.1046/j.1095-  
335 8339.2003.t01-1-00158.x
- 336 Barnett, J.R., Jeronimidis, G., 2003. *Wood quality and its biological basis*. Blackwell, Oxford, p.  
337 226.
- 338 Bjarnestad, S., Dahlman, O., 2002. Chemical Compositions of Hardwood and Softwood Pulp  
339 Employing Photoacoustic Fourier Transform Infrared Spectroscopy in Combination with  
340 Partial Least-Squares Analysis. *Anal. Chem.* 74, 5851–5858. doi:10.1021/ac025926z
- 341 Carballo-Meilan, A., Goodman, A.M., Baron, M.G., Gonzalez-Rodriguez, J., 2014. A specific case  
342 in the classification of woods by FTIR and chemometric: Discrimination of Fagales from  
343 Malpighiales. *Cellulose* 21, 261–273. doi:10.1007/s10570-013-0093-2
- 344 Chen, J., Liu, C., Chen, Y., Chen, Y., Chang, P.R., 2008. Structural characterization and properties  
345 of starch/konjac glucomannan blend films. *Carbohydr. Polym.* 74, 946–952.  
346 doi:10.1016/j.carbpol.2008.05.021
- 347 Chernick, M.R., 2011. *Bootstrap Methods: A Guide for Practitioners and Researchers*. Wiley,  
348 Hoboken, N.J., p. 400.
- 349 Coates, J., 2000. Interpretation of infrared spectra, a practical approach. *Encycl. Anal. Chem.*  
350 10815–10837.
- 351 Donald L. Pavia & Gary M. Lampman & George S. Kriz & James A. Vyvyan, 2009. Introduction  
352 to spectroscopy. Brooks/Cole, Cengage Learning, Belmont, CA, p. 727.
- 353 Ek, M., Gellerstedt, G., Henriksson, G., 2009. *Wood Chemistry and Wood Biotechnology*. Walter  
354 de Gruyter, Berlin, p. 308.
- 355 Erdtman, H., 1963. Some aspects of chemotaxonomy. *Chem. Plant Taxon.* 89–125.
- 356 Gidman, E., Goodacre, R., Emmett, B., Smith, A.R., Gwynn-Jones, D., 2003. Investigating plant-  
357 plant interference by metabolic fingerprinting. *Phytochemistry* 63, 705–710.  
358 doi:10.1016/S0031-9422(03)00288-7
- 359 Gorgulu, S.T., Dogan, M., Severcan, F., 2007. The characterization and differentiation of higher  
360 plants by fourier transform infrared spectroscopy. *Appl. Spectrosc.* 61, 300–8.  
361 doi:10.1366/000370207780220903
- 362 Hobro, A., Kuligowski, J., Döll, M., Lendl, B., 2010. Differentiation of walnut wood species and  
363 steam treatment using ATR-FTIR and partial least squares discriminant analysis (PLS-DA).  
364 *Anal. Bioanal. Chem.* 398, 2713–22. doi:10.1007/s00216-010-4199-1

- 365 Huang, A., Zhou, Q., Liu, J., Fei, B., Sun, S., 2008. Distinction of three wood species by Fourier  
366 transform infrared spectroscopy and two-dimensional correlation IR spectroscopy. *J. Mol.*  
367 *Struct.* 883-884, 160–166. doi:10.1016/j.molstruc.2007.11.061
- 368 Kacuráková, M., Kauráková, M., Capek, P., Sasinkova, V., Wellner, N., Ebringerova, A., Kac, M.,  
369 2000. FT-IR study of plant cell wall model compounds: pectic polysaccharides and  
370 hemicelluloses. *Carbohydr. Polym.* 43, 195–203. doi:10.1016/S0144-8617(00)00151-X
- 371 Kim, S.W., Ban, S.H., Chung, H.J., Cho, S., Choi, P.S., Yoo, O.J., Liu, J.R., 2004. Taxonomic  
372 discrimination of flowering plants by multivariate analysis of Fourier transform infrared  
373 spectroscopy data. *Plant Cell Rep.* 23, 246–50. doi:10.1007/s00299-004-0811-1
- 374 Klecka, W.R., 1980. *Discriminant analysis*. Sage Publications, Beverly Hills, Calif., p. 71.
- 375 Kubo, S., Kadla, J.F., 2005. Hydrogen bonding in lignin: a Fourier transform infrared model  
376 compound study. *Biomacromolecules* 6, 2815–21. doi:10.1021/bm050288q
- 377 Larkin, P., 2011. *Infrared and Raman Spectroscopy; Principles and Spectral Interpretation*.  
378 Elsevier, Amsterdam ; Boston, p. 230.
- 379 Liang, C.Y., Marchessault, R.H., 1959. Infrared spectra of crystalline polysaccharides. II. Native  
380 celluloses in the region from 640 to 1700 cm.1. *J. Polym. Sci.* 39, 269–278.  
381 doi:10.1002/pol.1959.1203913521
- 382 Liang, C.Y., Marchessault, R.H., 1959. Infrared spectra of crystalline polysaccharides. II. Native  
383 celluloses in the region from 640 to 1700 cm.1. *J. Polym. Sci.* 39, 269–278.  
384 doi:10.1002/pol.1959.1203913521
- 385 Marchessault, R.H., 1962. Application of infra-red spectroscopy to cellulose and wood  
386 polysaccharides. *Pure Appl. Chem.* 5, 107–130. doi:10.1351/pac196205010107
- 387 Marchessault, R.H., Liang, C.Y., 1962. The infrared spectra of crystalline polysaccharides. VIII.  
388 Xylans. *J. Polym. Sci.* 59, 357–378. doi:10.1002/pol.1962.1205916813
- 389 Marchessault, R.H., Pearson, F.G., Liang, C.Y., 1960. Infrared spectra of crystalline  
390 polysaccharides. I. Hydrogen bonds in native celluloses. *Biochim. Biophys. Acta* 45, 499–  
391 507.
- 392 Martin, J.W., 2007. *Concise encyclopedia of the structure of materials*. Elsevier, Amsterdam ;  
393 Boston, p. 512.
- 394 McCann, M.C., Bush, M., Milioni, D., Sado, P., Stacey, N.J., Catchpole, G., Defernez, M.,  
395 Carpita, N.C., Hofte, H., Ulvskov, P., Wilson, R.H., Roberts, K., 2001. Approaches to  
396 understanding the functional architecture of the plant cell wall. *Phytochemistry* 57, 811–821.  
397 doi:10.1016/S0031-9422(01)00144-3
- 398 Mohebbi, B., 2005. Attenuated total reflection infrared spectroscopy of white-rot decayed beech  
399 wood. *Int. Biodeterior. Biodegradation* 55, 247–251. doi:10.1016/j.ibiod.2005.01.003
- 400 Mohebbi, B., 2008. Application of ATR Infrared Spectroscopy in Wood Acetylation. *J. Agric. Sci*  
401 10, 253–259.
- 402 Muruganantham, S., Anbalagan, G., Ramamurthy, N., 2009. FT-IR and SEM-EDS comparative  
403 analysis of medicinal plants, *Eclipta Alba Hassk* and *Eclipta Prostrata Linn*. *Rom. J. Biophys*  
404 19, 285–294.
- 405 Obst, J.R., 1982. Guaiacyl and Syringyl Lignin Composition in Hardwood Cell Components.  
406 *Holzforschung* 36, 143–152. doi:10.1515/hfsg.1982.36.3.143

- 407 Pandey, K.K., 1999. A study of chemical structure of soft and hardwood and wood polymers by  
408 FTIR spectroscopy. *J. Appl. Polym. Sci.* 71, 1969–1975. doi:10.1002/(SICI)1097-  
409 4628(19990321)71:12<1969::AID-APP6>3.3.CO;2-4
- 410 Pandey, K.K., Vuorinen, T., 2008. Comparative study of photodegradation of wood by a UV laser  
411 and a xenon light source. *Polym. Degrad. Stab.* 93, 2138–2146.  
412 doi:10.1016/j.polymdegradstab.2008.08.013
- 413 Peter Zugenmaier, 2007. *Crystalline cellulose and derivatives: characterization and structures.*  
414 Springer, Berlin ; New York, p. 285.
- 415 Rakotomalala, R., 2005. “TANAGRA : un logiciel gratuit pour l’enseignement et la recherche.”
- 416 Rana, R., Langenfeld-Heysler, R., Finkeldey, R., Polle, A., 2009. FTIR spectroscopy, chemical and  
417 histochemical characterisation of wood and lignin of five tropical timber wood species of the  
418 family of Dipterocarpaceae. *Wood Sci. Technol.* 44, 225–242. doi:10.1007/s00226-009-  
419 0281-2
- 420 Rana, R., Sciences, F., 2008. Correlation between anatomical/chemical wood properties and  
421 genetic markers as a means of wood certification. *Niedersächsische Staats- und  
422 Universität Göttingen.* doi:978-3-9811503-2-2
- 423 Revanappa, S.B., Nandini, C.D., Salimath, P.V., 2010. Structural characterisation of pentosans  
424 from hemicellulose B of wheat varieties with varying chapati-making quality. *Food Chem.*  
425 119, 27–33. doi:10.1016/j.foodchem.2009.04.064
- 426 Rhoads, C.A., Painter, P., Given, P., 1987. FTIR studies of the contributions of plant polymers to  
427 coal formation. *Int. J. Coal Geol.* 8, 69–83. doi:10.1016/0166-5162(87)90023-1
- 428 Sekkal, M., Dincq, V., Legrand, P., Huvenne, J., 1995. Investigation of the glycosidic linkages in  
429 several oligosaccharides using FT-IR and FT Raman spectroscopies. *J. Mol. Struct.* 349,  
430 349–352.
- 431 Shen, J.B., Lu, H.F., Peng, Q.F., Zheng, J.F., Tian, Y.M., 2008. FTIR spectra of *Camellia* sect.  
432 *Oleifera*, sect. *Paracamellia*, and sect. *Camellia* (Theaceae) with reference to their taxonomic  
433 significance. *Plantsystematics.com* 46, 194–204. doi:10.3724/SP.J.1002.2008.07125
- 434 Silverstein, R.M., Webster, F.X., Kiemle, D., 2005. *Spectrometric identification of organic  
435 compounds.* Wiley, Hoboken, NJ, p. 502.
- 436 Sjostrom, 1981. *Wood chemistry: fundamentals and applications.* Academic Press, New York, p.  
437 293.
- 438 Stewart, D., Wilson, H.M., Hendra, P.J., Morrison, I.M., 1995. Fourier-Transform Infrared and  
439 Raman Spectroscopic Study of Biochemical and Chemical Treatments of Oak Wood  
440 (*Quercus rubra*) and Barley (*Hordeum vulgare*) Straw. *J. Agric. Food Chem.* 43, 2219–2225.  
441 doi:10.1021/jf00056a047
- 442 Sudiyani, Y., Tsujiyama, S., Imamura, Y., Takahashi, M., Minato, K., Kajita, H., Sci, W., 1999.  
443 Chemical characteristics of surfaces of hardwood and softwood deteriorated by weathering.  
444 *J. Wood Sci.* 45, 348–353.
- 445 Takayama, M., 1997. Fourier transform Raman assignment of guaiacyl and syringyl marker bands  
446 for lignin determination. *Spectrochim. Acta Part A Mol. Biomol. Spectrosc.* 53, 1621–1628.  
447 doi:10.1016/S1386-1425(97)00100-5

448 Wang, S., Wang, K., Liu, Q., Gu, Y., Luo, Z., Cen, K., Fransson, T., 2009. Comparison of the  
449 pyrolysis behavior of lignins from different tree species. *Biotechnol. Adv.* 27, 562–7.  
450 doi:10.1016/j.biotechadv.2009.04.010

451

## 452 **List of Figures**

453 Fig. 1 Average FTIR spectrum of division: Gymnosperm versus Angiosperm (A), score plot (B),  
454 3D plot (C) and loading plot (D) from Gymnosperm and Angiosperm dataset

455 Fig. 2 Average FTIR spectrum of class: Rosids versus Asterids (A), score plot (B), 3D plot (C) and  
456 loading plot (D) from Rosids and Asterids dataset

457 Fig. 3 Average FTIR spectrum of subclass: Euasterids I versus Euasterids II (A), score plot (B),  
458 2D plot (C) and loading plot (D) from Euasterids I and Euasterids II dataset

459 Fig. 4 Bias-variance decomposition from division, class and subclass models

460 Fig. 5 Boxplot of the discrimination function scores in division, class and subclass linear models

461

## 462 **List of Tables**

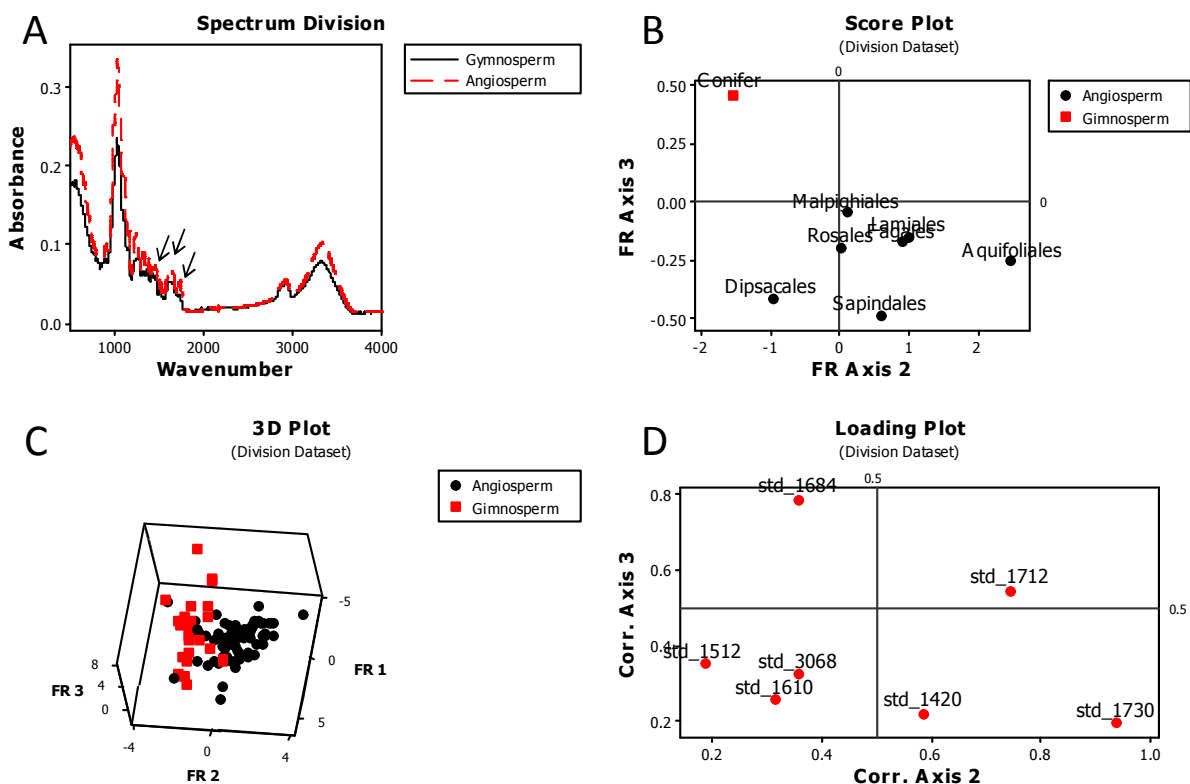
463 Table 1 Tree species based on APG III System Classification (APG, 2003)

464 Table 2 Band assignments of the third (FR3), fourth (FR4) and fifth (FR5) factor rotated loadings  
465 related to the variables obtained by PCA from ring dataset

466 Table 3 Classification functions for Gymnosperm, Rosids and Euasterids I, and validation from  
467 division, class and subclass models

468

469 **Figures**

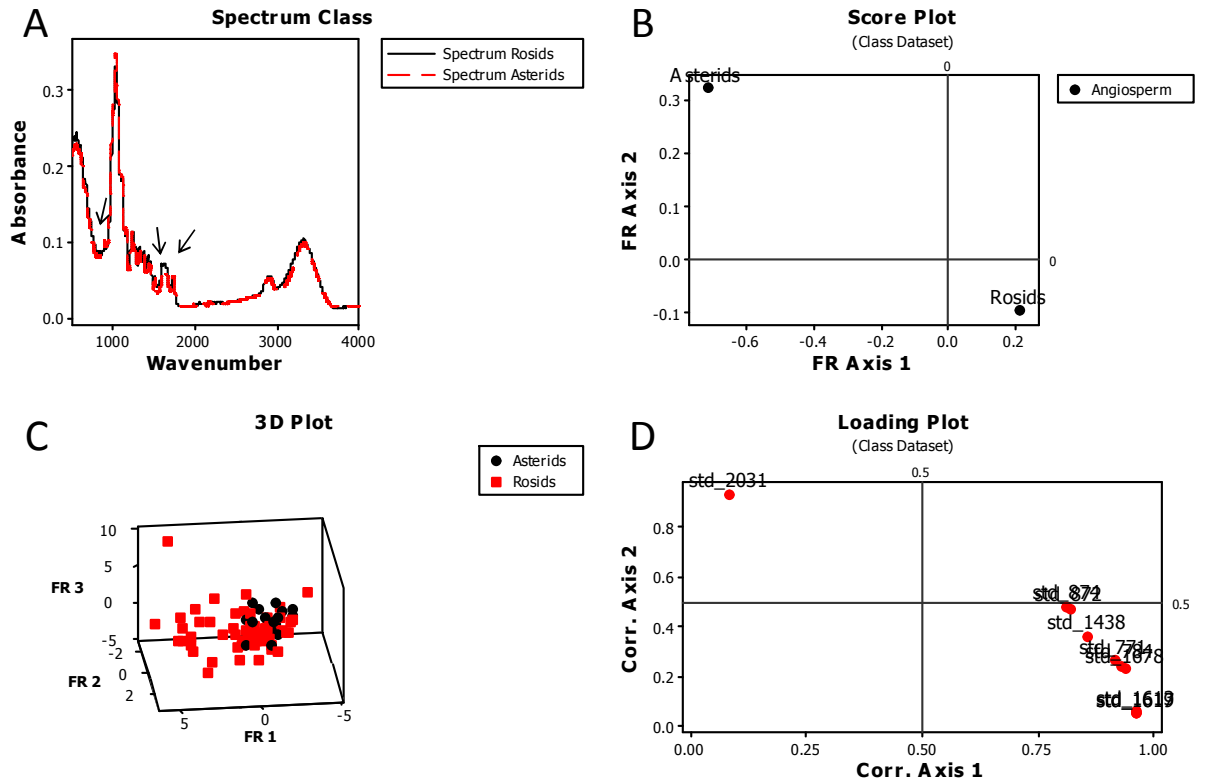


470

471 Fig. 1 Average FTIR spectrum of division: Gymnosperm versus Angiosperm (A),

472 3D plot (C) and loading plot (D) from Gymnosperm and Angiosperm dataset

473

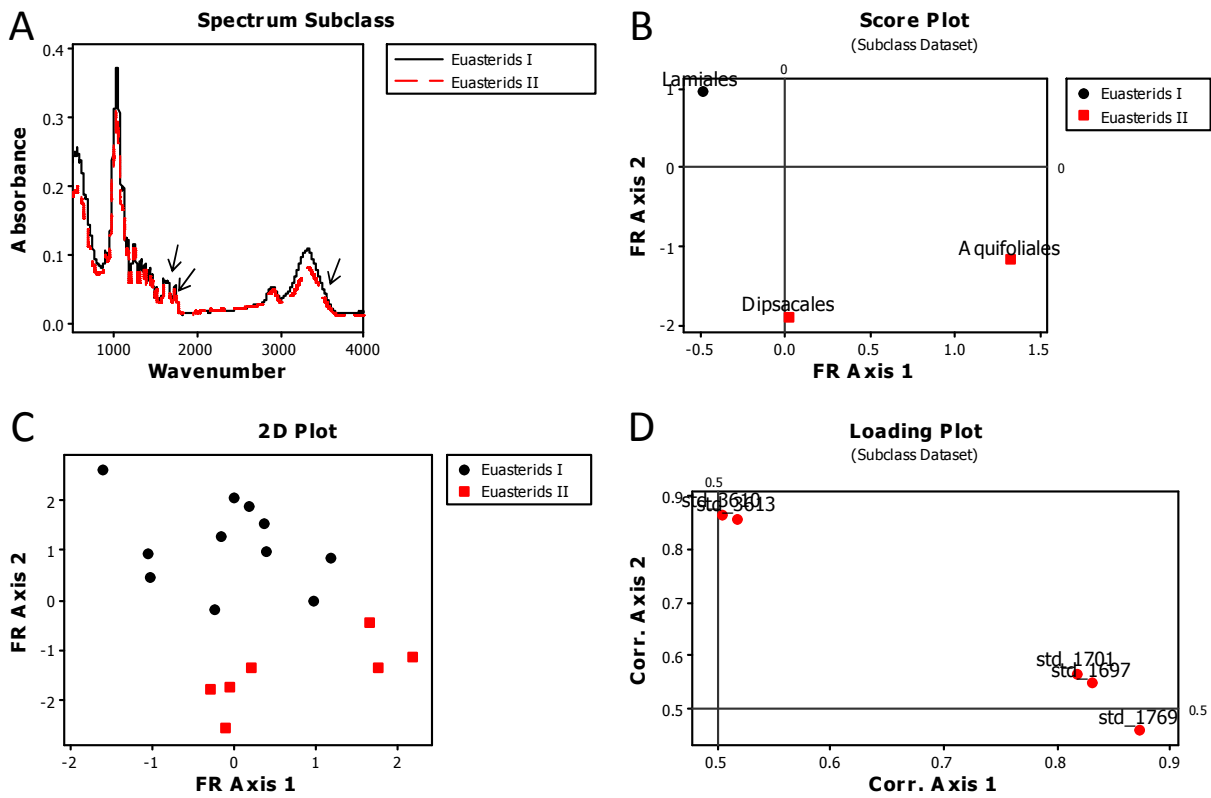


474

475

476

Fig. 2 Average FTIR spectrum of class: Rosids versus Asterids (A), score plot (B), 3D plot (C) and loading plot (D) from Rosids and Asterids dataset



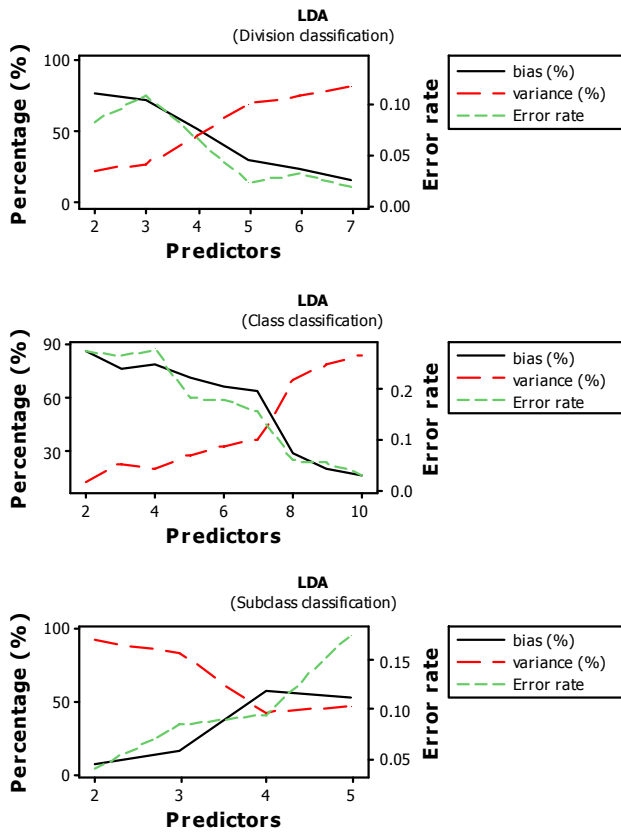
477

478

479

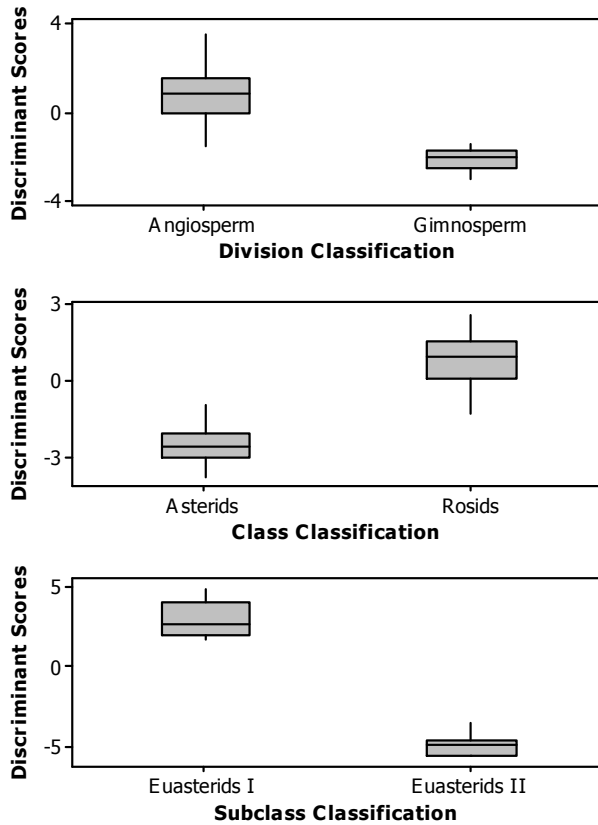
Fig. 3 Average FTIR spectrum of subclass: Euasterids I versus Euasterids II (A), score plot (B), 2D plot (C) and loading plot (D) from Euasterids I and Euasterids II dataset





480  
481  
482

Fig. 4 Bias-variance decomposition from division, class and subclass models



483  
484

Fig. 5 Boxplot of the discrimination function scores in division, class and subclass linear models

485 **Tables**

486 Table 1 Tree species based on APG III System Classification (APG, 2003)

Division	Class	Subclass	Order	Family	Genus	Specie	Common name
Gymnosperms	Pinophyta	Pinopsida	Pinales	Taxaceae	Taxus L.	<i>Taxus baccata</i>	Yew
				Pinaceae	Pinus L.	<i>Pinus sylvestris</i>	Scot Pine (3 varieties)
					Larix	<i>Larix decidua</i>	Larch
Angiosperms	Rosids	Eurosids I	Rosales	Moraceae	Ficus	<i>Ficus carica</i>	Fig
				Ulmaceae	Ulmus L.	<i>Ulmus procera</i>	Elm
			Fagales	Betulaceae	Alnus M.	<i>Alnus glutinosa</i>	Black Alder
					Corylus L.	<i>Corylus avellana</i>	Hazel
					Betula L.	<i>Betula pubescens</i>	Birch
			Fagaceae	Castanea	<i>Castanea sativa</i>	Sweet Chestnut	
				Fagus L.	<i>Fagus sylvatica</i>	Beech	
				Quercus	<i>Quercus robur</i>	English Oak	
			Malpighiales	Salicaceae	Populus	<i>Populus</i>	Poplar
					Populus	<i>Poplar nigra</i>	Black Poplar
					Salix	<i>Salix fragilis</i>	Willow
			Eurosids II	Sapindales	Sapindaceae	Acer	<i>Acer pseudoplatanus</i>
Asterids	Euasterids I	Lamiales	Oleaceae	Fraxinus L.	<i>Fraxinus excelsior</i>	Ash (2 varieties)	
		Aquifoliales	Aquifoliaceae	Illex L.	<i>Illex aquifolium</i>	Holly	
			Dipsacales	Adoxaceae	Sambucus	<i>Sambucus nigra</i>	Elder

487  
488

489 Table 2 Band assignments of the third (FR3), fourth (FR4) and fifth (FR5) factor rotated loadings  
 490 related to the variables obtained by PCA from ring dataset

FR	$\nu$ (cm <sup>-1</sup> )	Literature assignments and band origin
<i>Division</i>		
4	1762-1719	1740-1730, 1725 C=O stretching in acetyl groups of hemicelluloses (Åkerholm et al., 2001; Bjarnestad and Dahlman, 2002; Gorgulu et al., 2007; Marchessault and Liang, 1962; Marchessault, 1962; McCann et al., 2001; Mohebbi, 2008, 2005; Rana et al., 2009; Stewart et al., 1995)
	1245-1220	1245-1239 C-O of acetyl stretch of lignin and xylan 1238-1231 common to lignin and cellulose, S ring breathing with C-O stretching C-C stretching and OH in-plane bending (C-O-H deformation) cellulose, C-O-C stretching in phenol-ether bands of lignin(Åkerholm et al., 2001; Anchukaitis et al., 2008; Bjarnestad and Dahlman, 2002; Hobro et al., 2010; CY Y Liang and Marchessault, 1959; Marchessault, 1962; Pandey and Vuorinen, 2008; Rhoads et al., 1987)
	1132-950	1125,1123,1113 aromatic C-H in-plane deformation syringyl in lignin(Kubo and Kadla, 2005; Rhoads et al., 1987; Wang et al., 2009) 1110,1112 antisymmetrical in-phase ring stretch cellulose(CY Y Liang and Marchessault, 1959) 1090, 1092 C-C glucomannan(Kacuráková et al., 2000; McCann et al., 2001) 1090 antisymmetric $\beta$ C-O-C hemicelluloses(Sekkal et al., 1995) 1064 C=O stretching glucomannan(Gorgulu et al., 2007) 1059,1033 C-O stretch (C-O-H deformation) cellulose(CY Y Liang and Marchessault, 1959; Rhoads et al., 1987) 1030 aromatic C-H in-plane deformation guaiacyl plus C-O(Kubo and Kadla, 2005; Rhoads et al., 1987; Wang et al., 2009) 1034,941,898 C-H, ring glucomannan(Åkerholm et al., 2001; Bjarnestad and Dahlman, 2002; Gorgulu et al., 2007; Kacuráková et al., 2000; McCann et al., 2001)
5	2978-2832	2957 2922, 2873, 2852 CH <sub>3</sub> asymmetric and symmetric stretching: mainly lipids and proteins with a little contribution from proteins, carbohydrates, and nucleic acids(Gorgulu et al., 2007) 2945,2853 CH <sub>2</sub> antisymmetric stretching cellulose(Marchessault and Liang, 1962; Marchessault et al., 1960) 2853 CH <sub>2</sub> symmetric stretching xylan(Marchessault and Liang, 1962; Marchessault et al., 1960) 2940 (S), 2920(G), 2845-2835(S), 2820(G) C-H stretching (methyl and methylenes) lignin(Rhoads et al., 1987)
	1713-1676	1711 C=O stretch (unconjugated) in lignin(Hobro et al., 2010) Conj-CO-Conj(Larkin, 2011)
	1279-1274	1282,1280 C-H bending (CH <sub>2</sub> -O-H deformation) cellulose(CY Y Liang and Marchessault, 1959; Rhoads et al., 1987)
<i>Class</i>		
3	2860-2847	2852 CH <sub>2</sub> symmetric stretching: mainly lipids with a little contribution from proteins, carbohydrates, and nucleic acids(Gorgulu et al., 2007) 2853 CH <sub>2</sub> stretching xylan and cellulose(Marchessault and Liang, 1962; Marchessault et al., 1960)
	1171-884	1168-1146 C-O-C antisymmetric stretching in cellulose and xylan; and characteristic pectin band(Gorgulu et al., 2007; CY Y Liang and Marchessault, 1959; Marchessault and Liang, 1962; Marchessault, 1962; Mohebbi, 2005; Pandey and Vuorinen, 2008; Rana and Sciences, 2008; Rhoads et al., 1987; Sekkal et al., 1995) 1129-1088 out-of-plane ring stretch in cellulose and glucomannan, aromatic C-H in plane syringyl and C-O-C antisymmetric stretching hemicelluloses(Kubo and Kadla, 2005; C. Y. Liang and Marchessault, 1959; Sekkal et al., 1995; Wang et al., 2009) 1076-883 C-O-C symmetric stretching in hemicelluloses and celluloses; C-O stretch glucomannan and celluloses; and aromatic C-H deformation guaiacyl, amorphous cellulose and glucomannan(Bjarnestad and Dahlman, 2002; Gorgulu et al., 2007; Kacuráková et al., 2000; Kubo and Kadla, 2005; CY Y Liang and Marchessault, 1959; Mohebbi, 2005; Pandey and Vuorinen, 2008; Rana et al., 2009; Rhoads et al., 1987; Sekkal et al., 1995; Wang et al., 2009)

---

5	2929-2927	2922 CH <sub>2</sub> asymmetric stretching: mainly lipids with a little contribution from proteins, carbohydrates, and nucleic acids(Gorgulu et al., 2007)
	1687-1385	1683-1512 C-O ketones, flavones and glucuronic acid; amides in proteins; water; OH intramolecular H-bonding glucomannan; lignin skeletal(Chen et al., 2008; Gorgulu et al., 2007; Hobro et al., 2010; Huang et al., 2008; Kubo and Kadla, 2005; CY Y Liang and Marchessault, 1959; Marchessault and Liang, 1962; Rana and Sciences, 2008; Revanappa et al., 2010; Wang et al., 2009)

---

*Subclass*

---

4	1763-1709	1740-1730, 1725 C=O stretching in acetyl groups of hemicelluloses (Åkerholm et al., 2001; Bjarnestad and Dahlman, 2002; Gorgulu et al., 2007; Marchessault and Liang, 1962; Marchessault, 1962; McCann et al., 2001; Mohebbi, 2008, 2005; Rana et al., 2009; Stewart et al., 1995)
	1245-1212	1245-1239 C-O of acetyl stretch of lignin and xylan 1238-1231 common to lignin and cellulose, S ring breathing with C-O stretching C-C stretching and OH in-plane bending (C-O-H deformation) cellulose, C-O-C stretching in phenol-ether bands of lignin(Åkerholm et al., 2001; Anchukaitis et al., 2008; Bjarnestad and Dahlman, 2002; Hobro et al., 2010; CY Y Liang and Marchessault, 1959; Marchessault, 1962; Pandey and Vuorinen, 2008; Rhoads et al., 1987)

---

491  
492

493 Table 3 Classification functions for Gymnosperm, Rosids and Euasterids I, and validation from  
 494 division, class and subclass models

Classification functions		Statistical Evaluation	
Descriptors	LDA	F(1,5)	p-value
Division			
1730	3.3377	21.52445	0.000015
1712	-3.0887	9.14461	0.003414
1684	0.7958	1.6519	0.202655
1512	-2.9963	46.30463	0.000000
constant	-1.1877	-	
Class			
1678	-2.80427	23.71985	0.000011
1619	25.07698	14.33562	0.000398
1617	-22.13934	10.37686	0.002203
1438	0.917706	2.02774	0.160424
874	-1.413472	6.36166	0.014761
784	-6.00400	14.4103	0.000386
771	6.421311	21.53428	0.000024
constant	-0.52498		
Subclass			
3614	179.3411	4.59063	0.08504
3610	-224.9511	7.89394	0.037565
1768	58.8748	5.71739	0.062302
1701	-102.0568	6.67082	0.049265
constant	-22.1101	-	
Validation and test (ring samples)			
	Division	Class	Subclass
CV	0.0400	0.0900	0.0000
.632+	0.0508	0.0899	0.0513
Bootstrap			
LOO	0.0396	0.1081	0.0000
Train test	0.0452	0.0435	0.0500
Independent test (size)	0.0556(18)	0.2143(14)	0.0000(7)
Error rate	0.0671	0.1552	0.0950

495

496