

2023-12-29

# Optimisation of classification methods to differentiate morphologically-similar pollen grains from FT-IR spectra

Scoble, L

<https://pearl.plymouth.ac.uk/handle/10026.1/21812>

---

10.1016/j.revpalbo.2023.105041

Review of Palaeobotany and Palynology

Elsevier

---

*All content in PEARL is protected by copyright law. Author manuscripts are made available in accordance with publisher policies. Please cite only the published version using the details provided on the item record or document. In the absence of an open licence (e.g. Creative Commons), permissions for further reuse of content should be sought from the publisher or author.*



26 successful classification rate. This success demonstrates that using FT-IR microspectroscopy  
27 alongside spectral pre-processing and multivariate analysis can successfully identify and  
28 separate these moorland Poaceae species and has the clear potential to improve taxonomic  
29 resolution and classification of fossil pollen records. This will improve our understanding of  
30 how past land-use practice has shaped upland communities, provide more detailed  
31 ecologically-relevant palaeoecological information, and be utilised for the restoration and  
32 conservation of upland habitats.

33 **Keywords:** Pollen, Sporopollenin, FT-IR, Poaceae, Random forest

34

## 35 **1. Introduction**

36 Palynological research allows the reconstruction of past vegetation and environments to  
37 understand human impact and the development of a cultural landscape through time, creating  
38 insight into the response of ecosystems to anthropogenic impacts (Gaillard, et al., 2008).  
39 Semi-natural habitats such as moorlands, with diverse plant communities, are home to unique  
40 bird and insect species (Holden, et al., 2007). Palaeoecological research can aid in  
41 understanding of key long-term drivers causing changes in moorland vegetation community  
42 composition, and how different management regimes might be implemented to restore or  
43 maintain healthy environments (McCarroll, et al., 2017). Vegetation community composition  
44 and change may be a result of factors such as grazing or burning regimes and other  
45 management practices (Rowney, et al., 2023), or an indirect consequence of impacts such as  
46 twentieth-century nitrogen deposition that may have offered a competitive advantage to  
47 certain species in these nutrient-poor environments (Tomassen, et al., 2004). Palaeoecological  
48 data helps us understand how the landscape may have been of cultural significance to  
49 prehistoric communities (Davies & Bunting, 2010), and for conservation and restoration of  
50 blanket bogs and moorlands, such as *Calluna vulgaris*-dominated moorlands (Birks, 1996).

51         The application of palaeoecological methods to important conservation and  
52 management questions depends on the correct and detailed identification of the plant species  
53 comprising the vegetation communities of interest. Different research studies rely on the  
54 taxonomic resolution of pollen identification (Julier, et al., 2016) to increase the accuracy of  
55 their datasets. However, issues arise when pollen grains are indistinguishable using  
56 conventional light microscopy. Some morphologically similar taxa cannot be identified below  
57 family level, as there are no visible characteristics that allow them to be separated. This is a  
58 particular problem for Poaceae, a large and complex family whose pollen grains are  
59 morphologically indistinguishable through standard light microscopy. This results in coarse

60 taxonomic descriptions and resolution (Zimmerman, et al., 2016) which can lead to a loss of  
61 information and imprecise identification. Reliance on light microscopy thus presents  
62 limitations to the successful application of pollen analysis within upland contexts, where  
63 Poaceae pollen may represent more than 75 % of the pollen identified (Fyfe et al 2018). This  
64 leads to unanswered conservation and management questions particularly surrounding  
65 present day degraded mires and moorlands, where the long-term effects of animal grazing  
66 and land management practices on grassland communities are not fully recognised  
67 (Chambers, 2022). As an example, the conservation status of *Molinia caerulea* in  
68 environmentally sensitive areas has long been a subject of debate, with recent dominance  
69 linked to increased atmospheric nitrogen deposition, or different forms of more recent  
70 management practice (Chambers, et al., 1999). It has proved impossible to resolve the status  
71 of *Molinia caerulea* via light microscopy alone, and remains challenging even with the use of  
72 macrofossil analysis (the identification of grasses via their epidermis) as crucial features are  
73 not well preserved (Chambers, et al., 2013).

74         Research has demonstrated the successful use of Fourier-Transform Infra-Red  
75 spectroscopy (FT-IR) to identify and differentiate morphologically similar pollen taxa (Julier,  
76 et al., 2016), although the approach remains limited owing to the number of taxa for which  
77 measurements have been made. Infrared spectroscopy provides precise signatures of the  
78 biochemical composition of pollen (Zimmerman, et al., 2016). Pollen contains varying  
79 concentrations of specific lipids, proteins, carbohydrates and sporopollenin which are  
80 individual to each taxon, resulting from different dominant chemical functional groups in  
81 surface molecules due to their vibrational modes. These can all be identified using FT-IR  
82 spectroscopy, creating a spectrum consisting of numerous peak intensities (either  
83 transmittance or absorbance) (Kohler, et al., 2020). Furthermore, the use of FT-IR  
84 microspectroscopy (combination of FT-IR and microscopy) allows for focused measurements

85 on individual and clustered bioparticles, which can also be considered as a powerful tool for  
86 the characterisation of pollen grains.

87 Evidence of FT-IR's ability to identify and separate morphologically similar taxon can  
88 be seen in Julier et al. (2016), where 12 grass taxa from 8 subfamilies were identified across  
89 the grass phylogeny down to subfamily level, with an 80 % success rate. Jardine et al. (2019)  
90 classified eight domesticated and wild grasses based on the chemical signature of the pollen  
91 grains, achieving a 95% classification success rate when paired with k-nearest neighbour  
92 classification and leave-one-out cross validation. Zimmerman et al (2016) used FTIR  
93 microspectroscopy to classify singular pollen grains which included an optimising technique  
94 to prevent Mie-type scattering with a 95 % success rate, thus enabling better taxonomic  
95 resolution and classification. For wider context, Steemans et al (2010) used FT-IR  
96 microspectroscopy to demonstrate that cryptospores have similar spectra to that of trilete  
97 spores, which are composed of sporopollenin and characterised by "*absorption bands from*  
98 *aliphatic C-H in methylene (CH<sub>2</sub>) and methyl (CH<sub>3</sub>) groups, aromatic (C=C and C-H)*  
99 *groups and C=O groups of carboxylic acids*". Fraser et al (2012) analysed geologically  
100 unaltered sporopollenin from Pennsylvanian (310 million yr before present) cave deposits  
101 and demonstrated a strong chemical resemblance to extant relatives. Further comparisons  
102 indicated that the sporopollenin structure was similar across broader phylogenetic groups,  
103 with Fraser et al (2012) suggesting that "*land plant sporopollenin structure had remained*  
104 *stable since embryophytes invaded land*". Depciuch et al (2018) selected six *Betula* species  
105 to examine their chemical and morphological composition using FTIR. Their data showed  
106 that FTIR microspectroscopy could separate and manually characterise each individual  
107 chemical composition from most of the six *Betula* species, indicating that the technique can  
108 also identify morphologically similar tree taxa.

109 Whilst these studies have demonstrated the potential to distinguish morphologically similar  
110 pollen taxa, including Poaceae, more research is needed before such approaches can be  
111 considered suitable for application to the fossil pollen record to address questions relating to  
112 vegetation composition and change in moorland ecosystems. Firstly, it is necessary to  
113 demonstrate that key species can be separated, and to develop reference libraries for those  
114 species, and second, to develop classification approaches that can draw on reference libraries  
115 to automate the identification on unknown pollen grains. The aims of this research are  
116 therefore: 1) to test the extent to which FT-IR microspectroscopy can be used to separate the  
117 pollen of morphologically-similar moorland grasses; and 2) to assess the application of  
118 techniques including multivariate analysis and Random forest machine learning to determine  
119 species classification and separation.  
120

121 **2. Methods**

122

123 *2.1 Sample collection and preparation*

124 Five grass species were identified as important constituents of upland grassland  
125 communities in the UK, and chosen for analysis: *Agrostis capillaris*, *Anthoxanthum*  
126 *odoratum*, *Deschampsia cespitosa*, *Festuca ovina*, and *Molinia caerulea*. Four of the five  
127 species are widely distributed across the Northern Hemisphere, with the fifth (*Molinia*  
128 *caerulea* L.) abundant across Europe. The species are found across a wide variety of different  
129 moorland vegetation communities (Rodwell, 1998) . Fresh plant material for each species  
130 was collected from Northumberland, across four different locations (Figure 1). The *Agrostis*  
131 samples were not identified in the field beyond genus level, thus it is unclear which species  
132 were included in the sample. One bulk sample per specie was created by extracting four  
133 anthers from individual plant heads using tweezers, and delicately removing the pollen onto  
134 one half of a diamond anvil using a needle and scalpel. Pollen grains were compressed  
135 between the two halves of the anvil and then examined to see which half had the most sample  
136 on.



137

138 *Figure 1: Map of sample collection sites (Google Earth, 2023)*



139 *2.2 Chemical Analysis*

140 Individual bulk samples were examined using a Hyperion 1000 IR-enabled  
141 microscope with a 15x objective lens and liquid nitrogen-cooled MCT detector in absorbance  
142 mode, linked to a Bruker Vertex 70 (Bruker, Billerica, MA, USA) FT-IR bench unit. Fifty  
143 scans per bulk sample were taken with a background scan before the first scan and after every  
144 10<sup>th</sup>. Optimal scan rate and resolution (cm<sup>-1</sup>) were determined by preliminary method  
145 development (SM1), each scan consisted of 256 scans averaged with a resolution of 4 cm<sup>-1</sup>.  
146 Spectra were recorded between 4000 – 500 cm<sup>-1</sup> and scaled using Bruker OPUS vers.4  
147 software (Bruker, Billerica, MA, USA) for visual inspection.

148

149 *2.3 Spectral pre-processing*

150 In vibrational spectroscopy, spectroscopic data is generally pre-processed for data  
151 analysis (Kohler, et al., 2020). Pre-processing corrects the spectra by removing interfering  
152 atmospheric and instrumental effects. Due to the size and morphology of the samples,  
153 differences in the chemical compositions cannot be identified without pre-processing  
154 (Bassan, et al., 2010). Influential factors such as temperature, pressure and humidity can  
155 cause baseline drift (Yu, et al., 2013), affecting the overall accuracy of data analysis and  
156 classification. Therefore, baseline correction is used to set all baselines to zero absorption.  
157 The Extended Multiplicative Signal Correction (EMSC) model is regularly used in  
158 vibrational spectroscopy as a model-based pre-processing technique (Afseth & Kohler, 2012),  
159 aiding in correcting Mie scattering and peak positioning in FT-IR microspectroscopy  
160 (Bassan, et al., 2010). The model also allows for a reference spectrum to be included to aid  
161 baseline correction (Jardine, et al., 2021), with all corrected spectra resulting with the same  
162 baseline as the average (Afseth & Kohler, 2012). Raw spectra was baseline and EMSC  
163 corrected using the baseline (Liland, et al., 2010) and EMSC package (Martens & Stark,

164 1991; Liland, 2021) in R v.4.2.2 (R Core Team, 2022) with the mean spectrum of the dataset  
165 being used as the reference spectrum.

166 Derivatives of spectra can offer richer chemical information compared to raw spectra,  
167 as baseline effects are minimised while suppressed chemical signals are improved (Kohler, et  
168 al., 2020). Following the recommendations from Kohler et al. (2020), the raw spectral data  
169 was differentiated into second derivatives and EMSC performed afterwards. Derivatives of  
170 spectra can enhance noise (Jardine, et al., 2021); therefore, second derivative spectra were  
171 subject to Savitzky-Golay smoothing (window size of 15, polynomial of 2 and first degree)  
172 using the EMSC package (Martens & Stark, 1991) (Liland, 2021) in R v.4.2.2 (R Core Team,  
173 2022). Savitsky-Golay smoothing is an algorithm that estimates a spectrum by polynomial  
174 least-square fit, and defines a moving window which smooths the spectrum or derivated  
175 spectrum (Zimmerman & Kohler, 2013; Kohler, et al., 2020). Both the polynomial and the  
176 window size can influence the deviated curve, and ultimately the resulting spectrum and  
177 multivariate analysis. OriginLabs (OriginLab, Northampton, MA, USA) was used to plot the  
178 spectra.

#### 179 *2.4 Visual investigation and data analysis*

180 The mean and standard deviation of the non-differentiated spectra was calculated for  
181 each species using R v. 4.2.2 (R Core Team, 2022), and plotted for visual investigation  
182 (Figure 1) following Jardine's (2021) R script. Key absorption bands were chosen from  
183 previous research on sporopollenin chemistry and FT-IR Poaceae classification (Table 2)  
184 (Julier, et al., 2016; Jardine, et al., 2019; Kendel & Zimmermann, 2020; Zimmerman &  
185 Kohler, 2014; Steemans, et al., 2010; Fraser, et al., 2013; Fraser, et al., 2012; Watson, et al.,  
186 2007; Zimmerman, et al., 2017) for comparison against the average spectra (Figure 2 and  
187 SM3.1). Some absorption bands (e.g., the -OH band at 3300  $\text{cm}^{-1}$  and the  $\text{CH}_2$  bands at 2925  
188 and 2825  $\text{cm}^{-1}$ ) were omitted from data analysis as the bands offered no individual

189 classification information. Data analysis was conducted on both the non-differentiated and  
190 second derivative data in the spectral region of 1800-600  $\text{cm}^{-1}$  in R v. 4.2.2 (R Core Team,  
191 2022), where the biochemical signatures between species were compared and explored. For  
192 further investigation, the mean spectrum for each species were converted into their second  
193 derivatives.

194 Packages `vegan` (Oksanen, et al., 2020), `dendextend` (Galili, 2015) and `circlize` (Gu,  
195 2014) were used to perform hierarchical cluster analysis (HCA) (dendrogram) and principal  
196 component analysis (PCA) in R v. 4.2.2 (R Core Team, 2022) to visualise the non-  
197 differentiated and second derivative data. HCA and PCA were calculated using Euclidian  
198 distance to measure between-object distances, and classified samples into groups  
199 (Schumacker, 2016). Clear anomalies seen within the PCA were removed from the working  
200 dataset. PCA results and sample scores for each individual sample were extracted and plotted  
201 to visualise PC1 and PC2.

202 Loadings vectors for PC1 and PC2 were extracted to determine the importance of  
203 each absorbance band on each axis. A decision tree was created in Rstudio (Rstudio, 2020)  
204 using packages `rpart` (Therneau & Akinson, 2022) and `rpart.plot` (Milborrow, 2022) to  
205 identify and compare which specific wavenumbers were driving the species separation. The  
206 algorithm produces 'if-then' rules based on features in the dataset, resulting in a decision and  
207 outcome prediction. Rules were extracted to obtain the wavenumbers and absorbance units,  
208 then cross-checked with the original dataset to ensure the if-then rules were correct. The  
209 dataset was then split into training (80 %) and test (20 %) data, with the training dataset used  
210 to determine whether the wavenumbers used for the if-then rules varied every time a tree was  
211 created. A decision tree was run 100 times to ensure repeatability and rules were extracted.  
212 Comparisons between the original and trained decision tree were made, investigating which

213 variables were repeatedly used throughout the whole tree, and which were used regularly for  
214 the first broad split. Final comparisons were made against the PCA loading plots.

215         Using one decision tree for classification purposes can result in high variability and  
216 overfitting; therefore, Random forest (RF) was chosen to classify the non-differentiated  
217 dataset. RF is a supervised machine-learning algorithm using the collective wisdom of  
218 multiple decision trees to develop classification and regression models (Breiman, 2001).  
219 Classification trees are constructed by creating rules and decision points using training data  
220 that includes each sample's features. Samples move throughout each decision point until the  
221 terminal node is reached and classified. The trained model can then be used to predict classes  
222 of samples using the features alone. The ensemble method bagging can be used to reduce  
223 variance for more accurate predictions by setting the parameter *mtry* to the number of  
224 predictor variables (wavenumbers) within the dataset. Package randomForest (Wiener &  
225 Liaw, 2002) and the training dataset was used to produce and train the RF algorithm, with  
226 bagging being implemented with the argument *mtry* = 622. A confusion matrix using test data  
227 was produced to determine prediction accuracy, while variable importance indicated which  
228 variables (wavenumbers) would cause a greater loss in accuracy if excluded (Mean Decrease  
229 Accuracy), and which variables were most important in contributing to the homogeneity of  
230 the nodes, based off the mean decrease in Gini coefficient (MeanDecreaseGini).

231         Variable importance can be useful for variable reduction, where higher ranking  
232 variable can be used to build simpler models (Liaw & Wiener, 2002) while others that score  
233 lower are removed. By using variable importance measures, classification error rates can be  
234 kept at a similar level if low or reduced by only including important variables. Important  
235 variables were selected by running a RF loop within R v. 4.2.2 (R Core Team, 2022) and  
236 extracting the MeanDecreaseAccuracy (MDA) values. As *mtry*'s default is the square root of  
237 total variables for a classification model, the top 24 important variables for each data frame

238 were selected and combined into one data frame. A total of 240 variables were rearranged in  
239 ascending order from most important to least important, with the top 24 being selected again  
240 and replotted onto a dotchart. The MDA data was transformed into a boxplot to investigate  
241 which wavenumbers had greater within and between species variation. RF was re-run again  
242 with the refined dataset, a confusion matrix using test data was also produced to determine  
243 prediction accuracy. To test whether the simplified model could classify unlabelled samples,  
244 species labels were removed from the test data and predicted through a confusion matrix  
245 again.

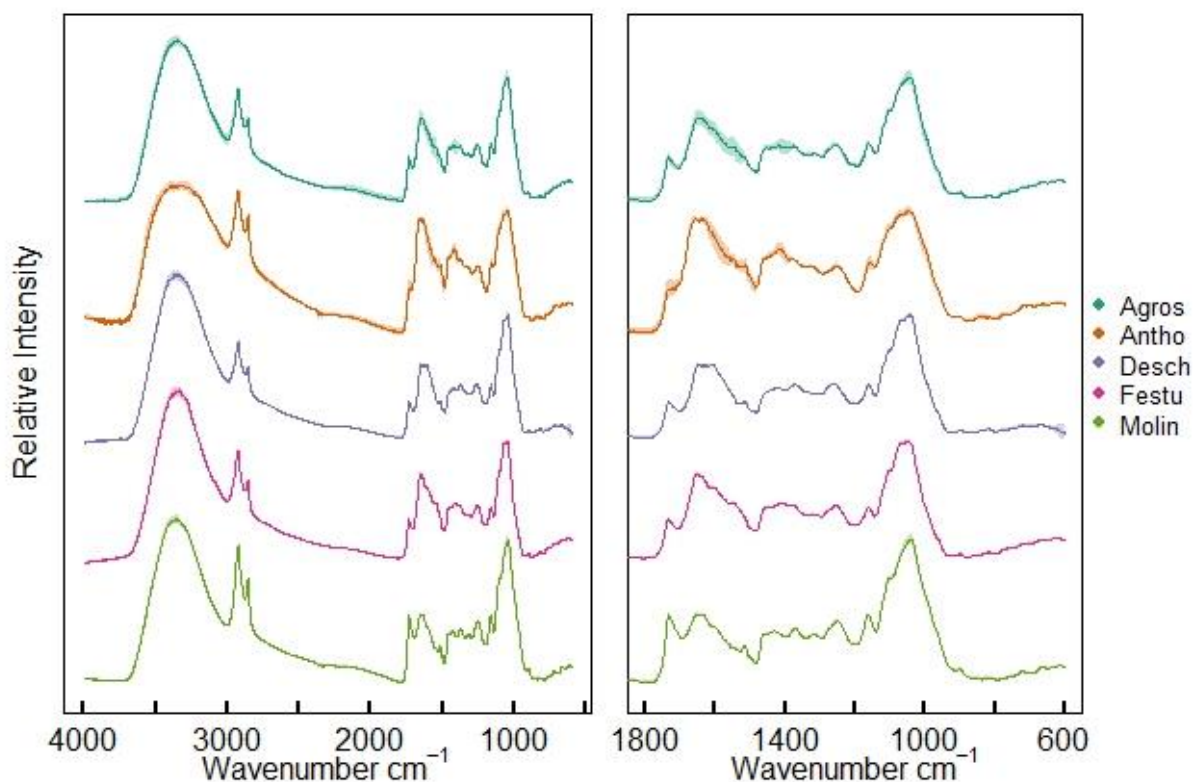
246 All R coding and additional figures can be found in the supplementary document SM2  
247 and SM3.

### 248 3. Results

249

250 FT-IR spectra of the five different Poaceae species (Figure 2 and SM3.1) exhibited  
251 the characteristic absorbance peaks that have been reported in pollen studies using FT-IR  
252 spectroscopy (Table 2) (Julier, et al., 2016; Jardine, et al., 2019; Kendel & Zimmermann,  
253 2020; Zimmerman & Kohler, 2014; Steemans, et al., 2010; Fraser, et al., 2013; Fraser, et al.,  
254 2012; Watson, et al., 2007; Zimmerman, et al., 2017). They presented similar vibrational  
255 bands, with a broad -OH stretch at 3500-3000  $\text{cm}^{-1}$ , asymmetric  $\text{CH}_2$  stretches at 2925 and  
256 2845  $\text{cm}^{-1}$ , C=C stretch at  $\sim 1600 \text{ cm}^{-1}$  and C-OH and C-O-C stretches at  $\sim 1040 \text{ cm}^{-1}$  present  
257 in all spectra. Each spectrum signal represents vibrational modes of proteins, lipids,  
258 carbohydrates and sporopollenin (Zimmerman, et al., 2017; Zimmerman & Kohler, 2014;  
259 Jardine, et al., 2019; Bağcıoğlu, et al., 2015; Zimmerman, et al., 2015). Protein signals are  
260 represented at 1650  $\text{cm}^{-1}$  (secondary amide I C=O stretch) and  $\sim 1550 \text{ cm}^{-1}$  (amide II N-H  
261 deformation and C-N stretching) ; lipids at 2925  $\text{cm}^{-1}$  (asymmetric  $\text{CH}_2$  stretch), 2845  $\text{cm}^{-1}$   
262 (asymmetric  $\text{CH}_2$  stretch), 1740-1710  $\text{cm}^{-1}$  (C=O stretch), 1460-1450  $\text{cm}^{-1}$  ( $\text{CH}_2$  deformation)  
263 and 1400  $\text{cm}^{-1}$  and carbohydrates between 1200-1000  $\text{cm}^{-1}$  (C-O, C-OH and C-O-C  
264 stretches). Sporopollenin can be associated with bands at  $\sim 1600 \text{ cm}^{-1}$  (aromatic C=C stretch),  
265  $\sim 1515 \text{ cm}^{-1}$  (aromatic C=C stretch), 1161  $\text{cm}^{-1}$  (C-O stretch) and between  $\sim 900\text{-}800 \text{ cm}^{-1}$  (C-  
266 H bend); and amino acids at 1375  $\text{cm}^{-1}$  (symmetric  $\text{CH}_3$  bend) and 1325  $\text{cm}^{-1}$  (C-N bend).  
267 Variance across the spectra is obscure but can be seen with some shading in bands within the  
268 fingerprint region. For further investigation, spectral data between 1800-600  $\text{cm}^{-1}$  was used to  
269 calculate the second derivatives for each specie (SM3.2).

270 Second derivative spectra (SM3.2) revealed that while there is similarity across all  
271 five spectra, most structural change happened between 1800-1400  $\text{cm}^{-1}$ , where signals  
272 associated with protein and lipids are identified, and 1200-1000  $\text{cm}^{-1}$ , where signals  
273 associated with carbohydrates are identified. Broad absorbance bands have been suppressed,



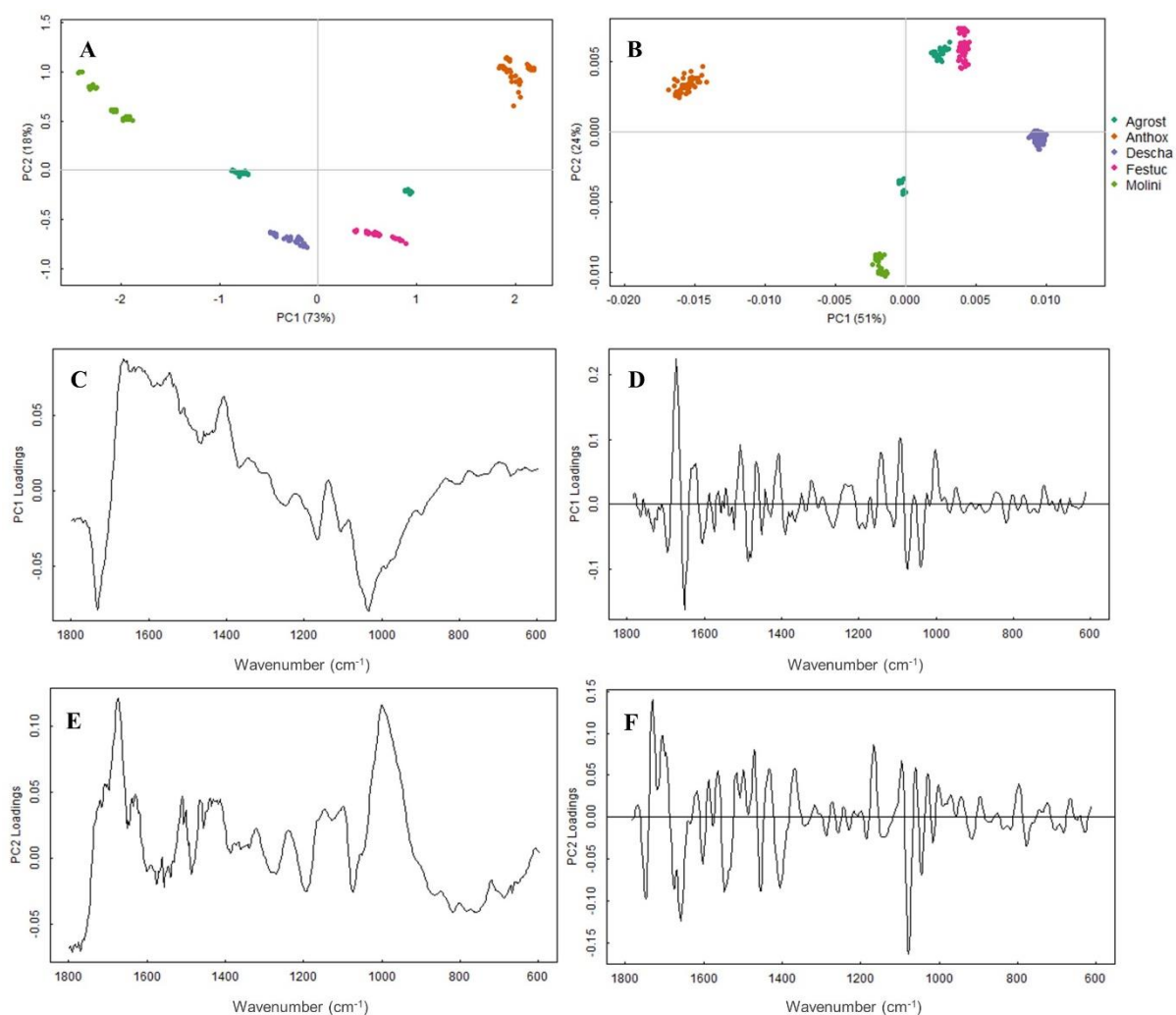
274

275 *Figure 2 (A): Stacked mean pre-processed FTIR spectra of the five chosen grass species for*  
 276 *the present study. (B) Fingerprint absorbance region of the FTIR spectra for each of the*  
 277 *grass species (1800-600  $\text{cm}^{-1}$ ). Shaded areas show  $\pm$  standard deviation about the mean.*

278

279 while peaks and shoulders have been enhanced. All species exhibit a strong downturned C=O  
 280 peak at  $1740\text{cm}^{-1}$ , this peak is present in the non-derivative spectra (Figure 2 and SM3.1), but  
 281 is more characteristic of a shoulder/weak peak. Peaks that are related to secondary structures  
 282 of proteins ( $1700\text{-}1600\text{cm}^{-1}$ ) are present across the five species, and all exhibit a C=O stretch  
 283 at  $\sim 1650\text{ cm}^{-1}$  (amide I). *Agrostis*, *Deschampsia cespitosa* and *Festuca ovina* have an  
 284 aromatic C=C stretch at  $1600\text{ cm}^{-1}$ , while *Anthoxanthum odoratum* and *Molinia caerulea*  
 285 display a C=O stretch at  $1630\text{cm}^{-1}$ . *Agrostis*, *Anthoxanthum odoratum* and *Festuca ovina*  
 286 display downturned symmetric peak at  $1550\text{ cm}^{-1}$  (amide II N-H deformation  
 287 and C–N stretching) and all exhibit peaks at  $\sim 1465\text{cm}^{-1}$  ( $\text{CH}_2$  deformation). *Agrostis*,

288 *Deschampsia cespitosa* and *Festuca ovina* present downturned strong peaks at  $\sim 1105\text{ cm}^{-1}$   
 289 (C-O-C stretch), whereas *Anthoxanthum odoratum* and *Molinia caerulea*'s peak exhibits a  
 290 slightly broad suppressed peak. Absorption bands related to sporopollenin at  $\sim 1515\text{ cm}^{-1}$  and  
 291  $\sim 1165\text{ cm}^{-1}$  are more pronounced as second derivatives in SM3.2 than Figure 2, with most  
 292 species apart from *Anthoxanthum odoratum* and *Molinia caerulea* exhibiting medium to  
 293 strong peaks at  $\sim 1060\text{ cm}^{-1}$ .



294  
 295 *Figure 3: (A): Non differentiated PCA and loading plots (C & E); (B) Second derivative*  
 296 *(Savitsky-Golay smoothed) and loading plots (D & F).*

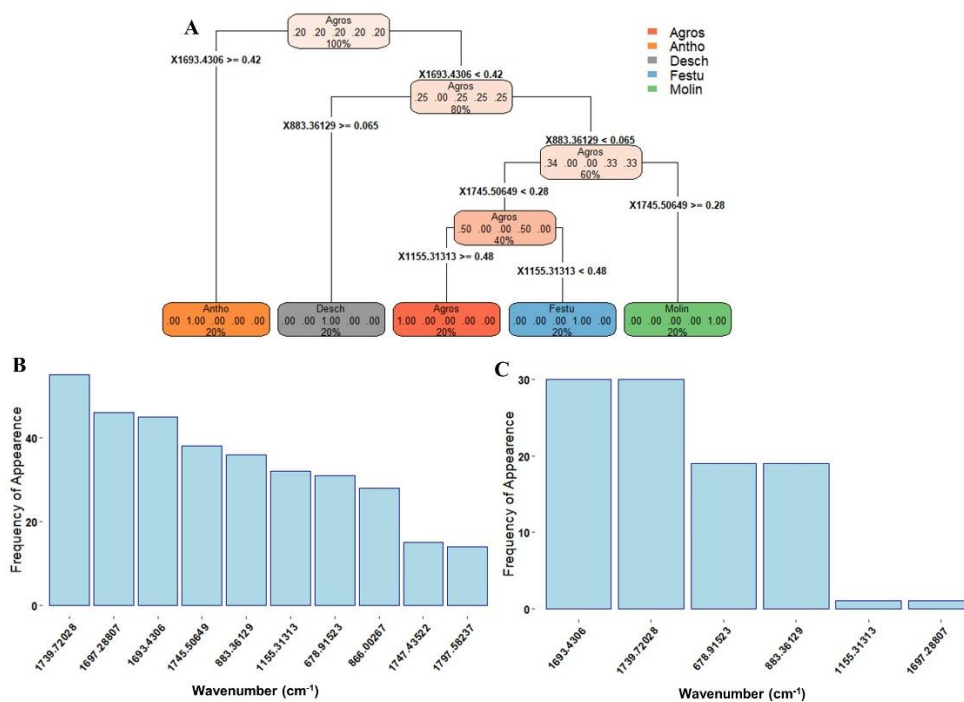
297 The first two components of the PCA of the non-differentiated spectra (Figure 3A) accounted  
 298 for 91 % of the variation, with PC1 contributing 73 % and PC2 contributing 18 %. There are



299 clear within-taxon groups spread out across the ordination space, although *Molinia caerulea*  
300 is spread further across both PC1 and PC2. *Agrostis capillaris* is evenly separated but is not  
301 overlapping any other species, while *Deschampsia cespitosa* and *Festuca ovina* have  
302 clustered together at the bottom. PC1 loading plot (Figure 3C) displays clear separation  
303 between the protein and carbohydrate region. PC1 has high positive loadings across the  
304 protein region (1700-1500 $\text{cm}^{-1}$ ) and high negative loadings in the lipid region (1750-1700 $\text{cm}^{-1}$ )  
305 and carbohydrate region (1200-900  $\text{cm}^{-1}$ ). Bands relating to sporopollenins at 1605, 1515,  
306 1171, 853, and 833  $\text{cm}^{-1}$  (Bağcıoğlu, et al., 2015) have positive loadings overall. PC2 loading  
307 plot (Figure 3E) has high positive loadings in the lipid region (1750-1730  $\text{cm}^{-1}$ ), low positive  
308 loadings in the secondary amide II region (1570-1515  $\text{cm}^{-1}$ ), high negative loadings in the  
309 secondary amide I region (1700-1600  $\text{cm}^{-1}$ ) and carbohydrate region (1200-1000  $\text{cm}^{-1}$ ). Low  
310 and high positive loadings at  $\sim 1500 \text{ cm}^{-1}$  and  $\sim 1165 \text{ cm}^{-1}$  are indicative of sporopollenin.

311 The first two components of the PCA of the second derivative spectra (Figure 3B)  
312 accounted for 67% of the variation, with PC1 contributing 51% and PC2 contributing 24%.  
313 Within taxon groupings are tighter with less overall spread across the ordination space.  
314 *Agrostis* is still separated but the majority of the data has clustered towards the top near  
315 *Festuca ovina*. PC1 loading plot (Figure 3D) exhibits more pronounced peaks, with high  
316 positive and negative secondary amide I region (1700-1600  $\text{cm}^{-1}$ ). Peaks related to the  
317 secondary amide II region (1570-1515  $\text{cm}^{-1}$ ) and carbohydrate region (1200–900  $\text{cm}^{-1}$ ) have  
318 low positive loadings, and sporopollenin at  $\sim 1600 \text{ cm}^{-1}$ ,  $\sim 1500 \text{ cm}^{-1}$  and 1161  $\text{cm}^{-1}$  have low  
319 negative loadings. PC2 loading plot (Figure 3F) exhibits stronger negative loadings within  
320 the secondary amide II region (1570-1515  $\text{cm}^{-1}$ ) region and carbohydrate region (1200–900  
321  $\text{cm}^{-1}$ ). An asymmetric stretch within the lipid's region (1750-1730  $\text{cm}^{-1}$ ) has high positive  
322 loadings, with peaks relating to sporopollenin at  $\sim 1600 \text{ cm}^{-1}$  and  $\sim 1161 \text{ cm}^{-1}$  having more  
323 prominent negative loadings.

324 The HCA for the non-differentiated data was split into five clusters (SM3.3). The first  
 325 division of the dendrogram separated *Molinia caerulea* from all other species; when five  
 326 clusters are chosen the *Molinia* separates further into two groups. *Anthoxanthum odoratum*  
 327 forms a distinct group while different species such as *Agrostis* and *Festuca ovina*, were  
 328 classified together. HCA for second derivative data (SM3.4) was split into five clusters again  
 329 and shows clear within-taxon groupings for all species. While *Agrostis* is not a single  
 330 coherent cluster, it has fallen into two distinct areas and has tighter clustering overall which is  
 331 exhibited in the PCA as well (Figure 3B).



332  
 333 *Figure 4: (A) Classification tree of pre-processed data, top decimals of terminal nodes*  
 334 *represent successful classification, with bottom percentage indicating how many samples*  
 335 *have been classified into that node. (B) Histogram of extracted wavenumbers used as*  
 336 *decision rules and the frequency of appearance throughout the entire tree. (C) Histogram of*  
 337 *extracted wavenumbers used as the first split rule for each tree and the frequency of*  
 338 *appearance.*

339

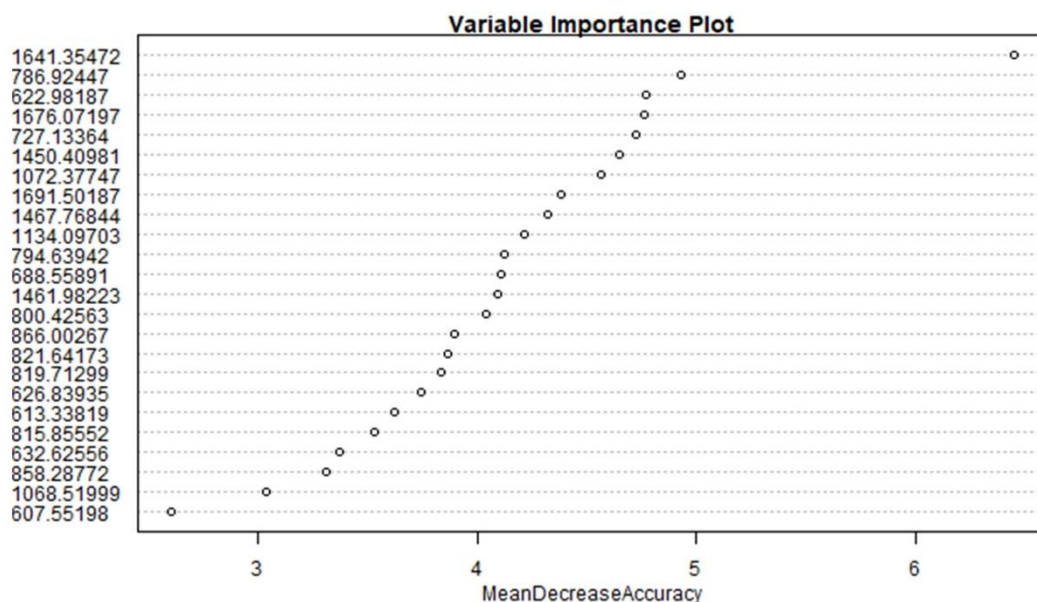
340 A decision tree for the pre-processed data had five terminal nodes (Figure 4A) with  
341 four predictors that overtook all other variables and were used as classification rules. Rules  
342 were cross checked with data set to ensure the algorithm was classifying correctly. All  
343 samples had 100 % successful classification indicated by the “1.00” and correct percentage at  
344 each terminal node. Figure 4B exhibits the top ten wavenumbers used by the looped decision  
345 tree model as classification rules by measuring the frequency of appearance. 1693.4306  $\text{cm}^{-1}$   
346 and 1745.50649  $\text{cm}^{-1}$  are within the top four, while 883.36129  $\text{cm}^{-1}$  and 1155.31313  $\text{cm}^{-1}$  has  
347 moved down to fifth and sixth. The majority of wavenumbers featured within the top ten are  
348 between 1800-1600  $\text{cm}^{-1}$  and 1200-800  $\text{cm}^{-1}$ , suggesting that absorbance bands found within  
349 the lipid (1750-1730  $\text{cm}^{-1}$ ), secondary amide I (1700-1600  $\text{cm}^{-1}$ ) and carbohydrate (1200-  
350 800 $\text{cm}^{-1}$ ) regions drive the discrepancy between the grass species. Figure 4C displays the top  
351 5 extracted wavenumbers used as the first split rule for each tree and the frequency of  
352 appearance, with 1693.4306  $\text{cm}^{-1}$ , 1739.72028  $\text{cm}^{-1}$ , 678.91523  $\text{cm}^{-1}$  and 883.36129  $\text{cm}^{-1}$   
353 within the top four.

354

355 Random forest (RF) classification performance was evaluated based on the training  
356 and test prediction accuracy (SM3.5) and out of bag estimate error rate (OOB). OOB is a  
357 method used for measuring the prediction error of machine learning models that use bagging.  
358 For the training and test dataset, the OOB was 0 % with 100 % accuracy. Figure 5 depicts the  
359 top 24 wavenumbers from the looped MeanDecreaseAccuracy (MDA) results. 1641.35472  
360  $\text{cm}^{-1}$  was identified as the most important, with roughly seven observations (samples) being  
361 misclassified if removed from the dataset. The top 24 variables were transformed into box  
362 plots (SM3.6) to investigate how the wavenumbers and corresponding spectral data differed  
363 between each species. *Agrostis* had the greatest within species variation at 1641.35472  $\text{cm}^{-1}$ ,

364 1461.98223  $\text{cm}^{-1}$  and 1450.40981  $\text{cm}^{-1}$ , indicating peaks varied more within the proteins and  
 365 lipids region. Wavenumbers with the greatest between species variation were seen within the  
 366 protein and lipids region, e.g, *Anthoxanthum odoratum* and *Molinia caerulea* had a 0.35  
 367 difference at 1641.35472  $\text{cm}^{-1}$  and 1467.76844  $\text{cm}^{-1}$ . The refined RF model had an error rate  
 368 of 0%, therefore species names were removed from test data and the trained RF model was  
 369 applied. Predicted names were compared against true species names (SM3.7) and showed  
 370 100 % prediction accuracy.

371



372

373 *Figure 5: randomForest variable importance plot of the top 24 selected variables using the*  
 374 *entire dataset to train the algorithm*

375

#### 376 4. Discussion

377 The combination of visual investigations (Figure 2 and SM3.2) (Table 2) and data  
378 analysis (Figure 3, Figure 4, Figure 5, SM3.3, SM3.4 and SM3.6) of the non-differentiated  
379 and second derivative FT-IR spectra above demonstrates that using FT-IR microspectroscopy  
380 can successfully identify and separate morphologically similar moorland grass species (Table  
381 1). While the spectra present similar vibrational bands between species, the spectra  
382 themselves exhibit many of the same distinctive absorbance bands demonstrated in previous  
383 pollen studies employing FT-IR spectroscopy (Table 2) (Julier, et al., 2016; Jardine, et al.,  
384 2019; Zimmerman & Kohler, 2014; Kendel & Zimmermann, 2020; Steemans, et al., 2010;  
385 Fraser, et al., 2012; Fraser, et al., 2013; Watson, et al., 2007; Zimmerman, et al., 2017). Table  
386 2 showed that most variance across the non-differentiated spectra was within the fingerprint  
387 region, with most structural changes observed between 1800-1600  $\text{cm}^{-1}$  and 1200-1000  $\text{cm}^{-1}$ .  
388 This suggested that the moorland grass species could be separated based on the pollen grains'  
389 protein and carbohydrate chemical composition. Second derivatives of the data (SM3.2)  
390 indicated variations in the secondary amide I structures of proteins (1700-1600  $\text{cm}^{-1}$ ),  
391 secondary amide II (1570-1515  $\text{cm}^{-1}$ ) and carbohydrate regions 1200-1000  $\text{cm}^{-1}$ .  
392 Characteristics bands of sporopollenin were also more pronounced at 1600  $\text{cm}^{-1}$ , ~1515  $\text{cm}^{-1}$   
393 and ~1160  $\text{cm}^{-1}$  with varying absorbance for each species. While visual investigations could  
394 determine some differences between species, more subtle chemical differences were harder to  
395 detect.

396 Following the data analysis, sections of spectra between 4000-1800  $\text{cm}^{-1}$  were  
397 removed as they offered no varying chemical information, while the fingerprint region was  
398 extended to 1800-600  $\text{cm}^{-1}$  as previous literature has demonstrated that lipids found at 1730  
399  $\text{cm}^{-1}$  can vary between species (Fraser, et al., 2012; Jardine, et al., 2019; Julier, et al., 2016;  
400 Zimmerman, et al., 2017). The PCA score plots for non-differentiated and second derivative

401 spectra (Figure 3A and 3B) have clear within-taxon groupings and wider dispersion of some  
402 species across the ordination space. *Agrostis* has separated out into two separate clusters,  
403 while the other species exhibit tighter clustering, particularly when subjected to pre-  
404 processing. This suggested (i) variability between the scans from possible differences  
405 between background scans or (ii) that the PCA has identified two different *Agrostis* species.  
406 There are four species of *Agrostis* that are commonly found in British moorland and  
407 heathland communities: *A. curtisii*, *A. capillaris*, *A. stolonifera* and *A. canina* (Rodwell  
408 1991). The *Agrostis* samples were not identified in the field beyond genus level, and it is thus  
409 unclear which species were scanned. The separation of *Agrostis* sp. from other grasses via  
410 multivariate analyses is a possible positive outcome, although it is clear that this genus and  
411 background scans needs further detailed investigation at the species level.

412         Score plots for non-differentiated spectra (Figure 3A) show that *Anthoxanthum*  
413 *odoratum* and *Molinia caerulea* have mostly positive score values for PC1, indicating that  
414 these species have similar chemical composition, while *Festuca ovina*, *Deschampsia*  
415 *cespitosa* and *Agrostis* have negative scores. Loading plots for the non-differentiated spectra  
416 (Figure 3C and 3E) highlight that PC1 separation is driven by lipid-based (1750-1730 cm<sup>-1</sup>),  
417 protein-based (1700-1500 cm<sup>-1</sup>) and carbohydrate-based (1200-1000 cm<sup>-1</sup>) chemical  
418 compositions of the individual species, while PC2 is driven by protein-based (1700-1500 cm<sup>-1</sup>)  
419 and carbohydrate-based (1200-1000 cm<sup>-1</sup>) chemical compositions. The second derivative  
420 PCA (Figure 3B) exhibited tighter within-taxon groupings compared to Figure 3A, with  
421 *Anthoxanthum odoratum*, *Festuca ovina* and one group of *Agrostis* having more positive  
422 score values for PC1, but *Anthoxanthum odoratum* has negative score values for PC2.  
423 Loading plots for the second derivative spectra (Figure 3D and Figure 3F) are more  
424 distinguished compared to the non-differentiated loading plots (Figure 3C and 3E) with strong  
425 peaks between the protein (1700-1500 cm<sup>-1</sup>) and carbohydrate regions (1200-800 cm<sup>-1</sup>).

426 Loading plots for PC1 (Figure 3D) indicate separation is driven by protein-based (1700-1500  
427  $\text{cm}^{-1}$ ) and carbohydrate-based chemical compositions (1200-1000  $\text{cm}^{-1}$ ) of species, while PC2  
428 is driven by the lipid-based (1750-1730  $\text{cm}^{-1}$ ), secondary amide I region (1700-1600  $\text{cm}^{-1}$ )  
429 and particularly carbohydrate-based (1200-1000  $\text{cm}^{-1}$ ) chemical compositions found at  $\sim 1090$   
430  $\text{cm}^{-1}$ .

431 Each of these key regions represent different signatures for biochemicals or nutrients  
432 found within pollen. Within the lipids region, triglycerides are characterised by a strong C=O  
433 stretch at 1745  $\text{cm}^{-1}$ , a weaker stretch at  $\sim 1460$   $\text{cm}^{-1}$  (Bağcıoğlu, et al., 2015), and  
434 phospholipids between 1160-1150  $\text{cm}^{-1}$ ; gluten and chitin compounds are characterised in the  
435 protein region by two broad bands at 1650  $\text{cm}^{-1}$  (secondary amide I: C = O stretch) and 1550  
436  $\text{cm}^{-1}$  (secondary amide II: N-H deformation and C-N stretching) (Zimmerman, et al., 2015);  
437 and some carbohydrates, such as cellulose at 1107  $\text{cm}^{-1}$ , 1055  $\text{cm}^{-1}$  and 1028  $\text{cm}^{-1}$  and  
438 amylose at 1076  $\text{cm}^{-1}$  and 995  $\text{cm}^{-1}$  (Bağcıoğlu, et al., 2015). Combining the non-  
439 differentiated and second derivative loading plots results determined that the most variation  
440 was found in the cellulose and amylose content between species, with gluten and triglyceride  
441 content also influencing separation as well.

442 The resulting HCAs for both non differentiated and second derivative spectra (SM3.3  
443 and 3.4) exhibit tight clustering of species, with *Anthoxanthum odoratum*, *Deschampsia*  
444 *cespitosa* and *Molinia caerulea* being exclusively clustered into individual groups. While  
445 *Agrostis* is not a single coherent cluster, it has fallen into two distinct areas and has tighter  
446 clustering overall. The results correlate well with the PCAs (Figure 3A and 3B) and exhibit  
447 low variance and better separation amongst species. *Molinia caerulea* across both HCAs  
448 (SM3.3 and 3.4) exhibits tight clustering; however, the non-differentiated HCA (SM 3.3)  
449 clusters *Molinia* into one group but with two distinct branches. The replicate scans that are  
450 seen in the smaller group of the scaled *Molinia caerulea* (SM 3.3) are between 0-9 and 45-49,

451 indicating that the first and last few scans taken have more variance compared to the scans  
452 taken in between. This could be a result of background correction issues or instrument  
453 variation; therefore, investigations surrounding these factors would be beneficial in the  
454 future.

455         The non-differentiated PCA (Figure 3B) and the HCA (SM3.4) see clear separate  
456 within-taxon groupings when data is subject to the EMSC method. Pre-processing the data  
457 using EMSC has benefitted the species separation overall. As the species' chemical  
458 information is very similar between 1800-600  $\text{cm}^{-1}$ , using EMSC normalises the variations  
459 found within spectra, such as scaling, baselines, and replicate variation (Liland, 2021). This  
460 aids multivariate analysis and the overall within-taxon groupings by reducing noise and  
461 scattering that had likely resulted from atmospheric effects. While data analysis of the raw  
462 spectra exhibits positive results, classification of the species can be optimised when spectral  
463 data is subject to pre-processing.

464         Non-differentiated spectral data between 1800-600  $\text{cm}^{-1}$  was used for machine  
465 learning classification. Using a decision tree (Figure 4A) and extracting the rules using a  
466 looped model (Figure 4B and 4C) determined which variables (wavenumbers) and  
467 corresponding absorbance bands were driving the discrepancy and classification of each grass  
468 species, with wavenumbers featured between 1800-1600  $\text{cm}^{-1}$  and 1200-800  $\text{cm}^{-1}$  within the  
469 top ten and 1693.4306  $\text{cm}^{-1}$  and 1155.31313  $\text{cm}^{-1}$  within the top four. When compared to the  
470 non-differentiated PCA loading plots (Figure 3D and 3F), species plotting more positively on  
471 PC1 and PC2 had higher variance in absorbance bands found in similar regions as those used  
472 as rules in the decision tree. This also coincides with the differing chemical composition of  
473 lipids and carbohydrates content found within pollen, with triglycerides represented in the  
474 top four wavenumbers at 1741  $\text{cm}^{-1}$  and 1745  $\text{cm}^{-1}$  and phospholipids in the top 10 at 1153  
475  $\text{cm}^{-1}$ , 1149  $\text{cm}^{-1}$  and 1151  $\text{cm}^{-1}$ .



476 Using random forest (RF) (Donges, 2023) achieved 100% successful classification  
477 and prediction accuracy for the pre-processed data respectively. While the decision tree  
478 (Figure 4A) had 100% successful classification, it is prone to overfitting, something of which  
479 RF overcomes by bootstrapping samples (Petkovic, et al., 2018). While RF has classed each  
480 species using a randomised algorithm, it has not separated the two potential *Agrostis* species  
481 exhibited in the PCAs (Figure 3B). Therefore, it would be beneficial to investigate whether  
482 RF can class different species from the same genus, instead of the same family.

483 MeanDecreaseAccuracy (MDA) plot (Figure 5) highlighted wavenumbers between 1800-  
484 1600  $\text{cm}^{-1}$ , 1500-1400  $\text{cm}^{-1}$  and 1200-800  $\text{cm}^{-1}$ , suggesting that the wavenumbers found  
485 within the lipid (1750-1730  $\text{cm}^{-1}$ ), secondary amide I (1700-1600  $\text{cm}^{-1}$ ), carbohydrate (1200-  
486 900 $\text{cm}^{-1}$ ) and aromatic ring (800-600  $\text{cm}^{-1}$ ) regions have stronger influence on the RF  
487 classification than the rest of the dataset. By transforming the MDA data into a boxplot  
488 (SM3.6), it revealed *Agrostis* had greater within species variation within the protein  
489 (1691.35472 $\text{cm}^{-1}$ ) and lipids (1461.98223 and 1450.981  $\text{cm}^{-1}$ ) region, this was also found  
490 between species. While the combined decision tree rules (Figure 4B) and RF's MDA plot  
491 (Figure 5) display similar but different wavenumbers, they are different machine learning  
492 techniques that are classifying the grass spectra successfully overall. By refining the data set  
493 used to train RF, classification accuracy was optimised and OOB error continued to stay at  
494 0%. Furthermore, a 100% successful prediction (SM3.7) was also achieved when labels were  
495 removed from the test data set, suggesting that the specific 24 variables included in the  
496 overall model are driving the discrepancy between species. Optimising and using a machine  
497 learning technique that has managed to separate samples into their correct species  
498 successfully, demonstrates that FT-IR spectra can be used to separate and classify  
499 morphologically similar grass species if paired with multivariate data analysis.

500 While this research has shown that moorland grasses can be differentiated, the results  
501 are based on modern pollen material from one region. Pollen grain composition can differ  
502 ecologically by being exposed to different temperatures, humidity, light, and nutrients  
503 (Zimmerman, et al., 2017; Pacini & Franchi, 2020); therefore, future research on these  
504 species should be focused on investigating spatial and environmental variation found within  
505 the spectra. While the results demonstrate random forest can be used as a classifier and the  
506 optimised, the variable importance indicated that proteins, lipids and mostly carbohydrates  
507 were used to classify each species. Modern pollen contains internal material such as  
508 cytoplasm and intine (Julier, et al., 2016), therefore the signal for each of the species above  
509 represents the whole pollen grain and not just the sporopollenin. Though this information is  
510 beneficial for modern pollen, it is not as useful for classifying unknown fossil material and  
511 requires further work to optimise techniques for this type of sample. Research comparing  
512 modern and fossil pollen have indicated that peaks found at  $\sim 1740\text{ cm}^{-1}$  (C=O stretch within  
513 the lipids region),  $\sim 1650\text{ cm}^{-1}$  (amide I within the proteins region) and  $\sim 1550\text{ cm}^{-1}$  (secondary  
514 amide II within the protein region) in modern pollen were absent in fossil pollen, likely  
515 resulting from degradation within the peatbog and the loss of pollenkit and intines (Wang, et  
516 al., 2023). Fossil pollen or pollen samples that have been chemically processed may lose non-  
517 sporopollenin pollen components (Julier, et al., 2016), therefore making modern spectral  
518 information less efficient for classification. Sporopollenins are robust grain wall biopolymers  
519 based off of phenylpropanoids such as *p*-coumaric, ferulic and sinapic acids (Zimmerman,  
520 2010; Bağcıoğlu, et al., 2015), their chemical signatures have provided information on the  
521 past and the possibility to identify fossil pollen (Fraser, et al., 2012; Lomax, et al., 2012;  
522 Fraser, et al., 2013; Jardine, et al., 2021). Thus, there is an opportunity for further research  
523 conducting replicate measurements across different species and taxonomic groups has been  
524 recommended (Jardine, et al., 2021), therefore future investigations surrounding the species

525 above should focus more on the sporopollenin chemistry to aid in better classification at

526 genus and species level.

527

528 **5. Conclusions**

529 This study demonstrated that using FT-IR microspectroscopy alongside spectral pre-  
530 processing and multivariate analysis can successfully identify and separate morphologically  
531 similar pollen taxa, specifically four species and one genus from the Poaceae family that are  
532 common across moorland communities. Using a pre-processing method, further multivariate  
533 analysis on the spectral data and optimising a machine learning algorithm has led to a 100%  
534 successful classification rate of species overall. This has the clear potential to improve  
535 taxonomic resolution and classification of fossil pollen records, particularly as grasses can  
536 represent up to 75% of pollen identified in moorland and upland pollen sequences. Applying  
537 an improved taxonomic resolution will improve our understanding of how past land-use  
538 practice has shaped upland communities, enable the provision of much more detailed  
539 ecologically-relevant palaeoecological information, and can be utilised for the restoration and  
540 conservation of upland habitats. Whilst this study has demonstrated the potential of FT-IR  
541 microspectroscopy for moorland grass identification, the next steps in this frontier will be to  
542 develop spectra from species across a wider spatial range (particularly the *Agrostis* species, as  
543 shown in this study), to investigate species sporopollenin chemistry through single grain  
544 analysis, and to further develop the statistical approaches that will enable the routine  
545 separation of the FT-IR spectra.

546

547 **6. Acknowledgement**

548 We would like to thank Northumbria Wildlife Trust (Geoff Dobbins) for collecting and  
549 sending samples to the University of Plymouth, Matthew Kent for discussions about R coding  
550 and project outlines, Faidra Katsi for discussions about project outline and research, Billy  
551 Simmonds for all technical and instrument training support and Jamie Quinn for cartographic  
552 support.

553 **References**

- 554 Afseth, K. N. & Kohler, A., 2012. Extended multiplicative signal correction in vibrational  
555 spectroscopy, a tutorial. *Chemometrics and Intelligent Laboratory Systems* , Volume 117, pp.  
556 92-99.
- 557 Bağcıoğlu, M., Zimmerman, B. & Kohler, A., 2015. A Multiscale Vibrational Spectroscopic  
558 Approach for Identification and Biochemical Characterization of Pollen. *PLoS ONE*, 10(9).
- 559 Bassan, P. et al., 2010. RMieS-EMSC correction for infrared spectra of biological cells:  
560 Extension using Mie theory and GPU computing. *J. Biophoton*, Volume 3, pp. 609-620.
- 561 Birks, H. J. B., 1996. Contributions of Quaternary palaeoecology to nature conservation.  
562 *Journal of Vegetation Science*, Volume 7, pp. 89-98.
- 563 Breiman, L., 2001. Random forests. *Machine learning*, Volume 45, pp. 5-32.
- 564 Brenchley, J. P. & Harper, T. A., 1998. Investigating the history of the biosphere. In:  
565 *Palaeoecology: Ecosystems, environments and evolution*. London: Chapman and Hall, an  
566 input of Thomson Science, pp. 1-6.
- 567 Bush, B. M., 2002. On the interpretation of fossil Poaceae pollen in the lowland humid  
568 neotropics. *Palaeogeogr. Palaeoclimatol. Palaeoecol.*, pp. 5-17.
- 569 Chambers, F. M., 2022. The use of paleoecological data in mire and moorland conservation.  
570 *Past Global Changes Magazine*, 30(1), pp. 16-17.
- 571 Chambers, M. F. et al., 2013. Long-term ecological study (palaeoecology) to chronicle habitat  
572 degradation and inform conservation ecology: an exemplar from the Brecon Beacons, South  
573 Wales. *Biodiversity and Conservation*, Volume 22, pp. 719-736.
- 574 Chambers, M. F., Mauquoy, D. & Tood, A. P., 1999. Recent rise to dominance of *Molinia*  
575 *caerulea* in environmentally sensitive areas: new perspectives from palaeoecological data.  
576 *Journal of Applied Ecology*, 36(5), pp. 719-733.

577 Davies, A. L. & Bunting, M. J., 2010. Applications of Palaeoecology in Conservation. *The*  
578 *Open Ecology Journal*, Volume 3, pp. 54-67.

579 Delcourt, H. R., 1987. The impact of prehistoric agriculture and land occupation on natural  
580 vegetation. *Trends Ecol. Evolut.*, Volume 2, pp. 39-44.

581 Depciuch, J., Kasprzyk, I., Drzymata, E. & Parlinska-Wojtan, M., 2018. Identification of  
582 birch pollen species using FTIR spectroscopy. *Aerobiologia*, Volume 34, pp. 525-538.

583 Donges, N., 2023. *Random Forest: A Complete Guide for Machine Learning*. [Online]  
584 Available at: <https://builtin.com/data-science/random-forest-algorithm#what>  
585 [Accessed 6 June 2023].

586 Faegri, K. & Iverson, J., 1989. *Textbook of Pollen Analysis*. IV ed. New Jersey: The  
587 Blackburn Press.

588 Fraser, T. W. et al., 2012. Evolutionary Stasis of Sporopollenin Biochemistry Revealed by  
589 Unaltered Pennsylvanian Spores. *New Phytologist*, Volume 196, pp. 397-401.

590 Fraser, T. W. et al., 2013. Changes in Spore Chemistry and Appearance With Increasing  
591 Maturity. *Review of Palaeobotany and Palynology*, Volume 201, pp. 41-46.

592 Gaillard, J. M. et al., 2008. Human impact on terrestrial ecosystems, pollen calibration and  
593 quantitative reconstruction of past land-cover. *Vegetation History and Archaeobotany*,  
594 Volume 17, pp. 415-418.

595 Gaillard, J.-M. et al., 2008. The use of modelling and simulation approach in reconstructing  
596 past landscapes from fossil pollen data: a review and results from the POLLANDCAL  
597 network. *Vegetation History and Archaeobotany*, 17(5), pp. 419-443.

598 Galili, T., 2015. dendextend: an R package for visualizing, adjusting, and comparing trees of  
599 hierarchical clustering. *Bioinformatics*.

600 Gu, Z., 2014. circlize implements and enhances circular visualization in R. *Bioinformatics*.

601 Holden, J. et al., 2007. Environmental change in moorland landscapes. *Earth-Science*  
602 *Reviews*, 82(1-2), pp. 75-100.

603 Jardine, E. P., 2021. *Data and code for "Sporopollenin chemistry and its durability in the*  
604 *geological record: an integration of extant and fossil chemical data across the seed plants"*.  
605 [Online]  
606 Available at: <https://doi.org/10.6084/m9.figshare.11382102.v1>  
607 [Accessed 28 September 2023].

608 Jardine, E. P. et al., 2019. Chemotaxonomy of domesticated grasses: a pathway to  
609 understanding the origins of agriculture. *Micropalaeontol*, Volume 38, pp. 83-95.

610 Jardine, E. P. et al., 2021. Sporopollenin chemistry and its durability in the geological record:  
611 an integration of extant and fossil chemical data across the seed plants. *Palaeontology*, 64(2),  
612 pp. 285-305.

613 Julier, C. M. A. et al., 2016. Chemotaxonomy as a tool for interpreting the cryptic diversity of  
614 Poaceae pollen. *Review of Palaeobotany and Palynology*, Volume 235, pp. 140-147.

615 Kendel, A. & Zimmermann, B., 2020. Chemical Analysis of Pollen by FT-Raman and FTIR  
616 Spectroscopies. *Frontiers in Plant Science*, Volume 11.

617 Kohler, A. et al., 2020. Model-Based Pre-Processing in Vibrational Spectroscopy. In:  
618 *Comprehensive Chemometrics*. s.l.:Elsevier, pp. 83-100.

619 Kuhn, M., 2008. Building Predictive Models in R Using the caret Package. *Journal of*  
620 *Statistical Software*, 28(5), pp. 1-26.

621 Liaw, A. & Wiener, M., 2002. Classification and Regression by randomForest. *R News*,  
622 Volume 2, pp. 18-22.

623 Liland, H. K., 2021. *EMSC: Extended Multiplicative Signal Correction, R Package*. [Online]  
624 Available at: <https://CRAN.R-project.org/package=EMSC>  
625 [Accessed 11 January 2023].



626 Liland, H. K., Almøy, T. & Mevik, H., 2010. Optimal Choice of Baseline Correction for  
627 Multivariate Calibration of Spectra. *Applied Spectroscopy*, Volume 64, pp. 1007-1016.

628 Martens, H. & Stark, E., 1991. Extended multiplicative signal correction and spectral  
629 interference subtraction: new preprocessing methods for near infrared spectroscopy. *J Pharm*  
630 *Biomed Anal*, 9(8), pp. 625-35.

631 McCarroll, J., Chambers, M. F., Webb, C. J. & Thom, T., 2017. Application of palaeoecology  
632 for peatland conservation at Mossdale Moor, UK. *Quaternary International*, Volume 432, pp.  
633 39-47.

634 Milborrow, S., 2022. *Plot 'rpart' Models: An Enhanced Version of 'plot.rpart'*. [Online]  
635 Available at: <https://CRAN.R-project.org/package=rpart.plot>  
636 [Accessed 23 May 2023].

637 Moore, D. P., Webb, A. J. & Collnison, E. M., 1991. Basis and Application . In: 2nd, ed.  
638 *Pollen Analysis*. Oxford: Blackwell Scientific Publications, pp. 1-4.

639 Oksanen, J. et al., 2020. *vegan: Community Ecology Package. R package version*. [Online]  
640 Available at: <https://CRAN.R-project.org/package=vegan>  
641 [Accessed 30 January 2023].

642 Pacini, E. & Franchi, G. G., 2020. Pollen biodiversity: why are pollen grains different despite  
643 having the same function? A review. *Botanical Journal* , 193(2), pp. 141-164.

644 Petkovic, D., Altman, R., Wong, M. & Vigil, A., 2018. Improivng the explainability of  
645 Random Forest classifier - user centered approach. *Biocomputing* , pp. 204-215.

646 Pigott, D. C. et al., 1991. *British Plant Communities Volume 2: Mires and Heaths*.  
647 Cambridge: Cambridge University Press.

648 Pigott, D. C. et al., 1992. *British Plant Communities Volume 3: Grasslands and Montane*  
649 *Communities*. Cambridge: Cambridge Publishing Press.

650 R Core Team, 2022. *R: a language and environment for statistical computing*. Vienna: R  
651 Foundation for Statistical Computing.

652 Rodwell, S. J., 1998. *British Plant Communities. Volume 2. Mires and Heaths*. Cambridge:  
653 Cambridge University Press.

654 Rowney, M. F. et al., 2023. Historical anthropogenic disturbances explain long-term  
655 moorland vegetation dynamics. *Ecology and Evolution*, 13(3).

656 Rstudio, T., 2020. *RStudio: Integrated Development for R*. [Online]  
657 Available at: <http://www.rstudio.com/>

658 Schumacker, E. R., 2016. Multidimensional Scaling. In: *Using R With MULTIVARIATE*  
659 *STATISTICS*. California: Sage Publications, pp. 231-232.

660 Simpson, G. L. & Oksanen, J., 2021. *analogue: Analogue matching and Modern Analogue*  
661 *Technique transfer function models*. [Online]  
662 Available at: <https://cran.r-project.org/package=analogue>  
663 [Accessed 17 March 2023].

664 Sobol, K. M. & Finkelstein, A. S., 2018. Predictive pollen-based biome modeling. *PLOS*,  
665 13(8), pp. 1-29.

666 Speiser, L. J., Miller, E. M., Tooze, J. & Ip, E., 2019. A comparison of random forest variable  
667 selection methods for classification prediction modeling. *Expert Systems with Applications*,  
668 134(15), pp. 93-101.

669 Steemans, P. et al., 2010. FTIR Characterisation of the chemical composition of Silurian  
670 miospores (cryptospores and trilete spores) from Gotland, Sweden. *Review of Palaeobotany*  
671 *and Palynology*, Volume 162, pp. 577-590.

672 Therneau, T. & Akinson, B., 2022. *Recursive Partitioning and Regression Trees*. [Online]  
673 Available at: <https://CRAN.R-project.org/package=rpart>  
674 [Accessed 23 May 2023].

675 Tomassen, M. B. H. et al., 2004. Expansion of invasive species on ombrotrophic bogs:  
676 desiccation or high N deposition?. *Journal of applied Ecology*, 41(1), pp. 139-150.

677 Watson, S. J. et al., 2007. Rapid Determination of Spore Chemistry Using Thermochemolysis  
678 Gas Chromatography-mass Spectrometry and micro-Fourier Infrared Spectroscopy.  
679 *Photochemical & Photobiological Sciences*, Volume 6, pp. 689-694.

680 Wickham, H., 2007. Reshaping Data with the reshape Package. *Journal of Statistical*  
681 *Software*, Volume 21, pp. 1-20.

682 Wickham, H., 2016. *ggplot2: Elegant Graphics for Data Analysis.*, New York: Springer-  
683 Verlag.

684 Wickham, H. & Girlich, M., 2022. *tidyr: Tidy Messy Data*. [Online]  
685 Available at: <https://CRAN.R-project.org/package=tidyr>  
686 [Accessed 13 January 2023].

687 Wiener, M. & Liaw, A., 2002. Classification and Regression by randomForest. *R News*, 2(3),  
688 pp. 18-22.

689 Yu, S., You, X. & Mou, Y., 2013. *A New Approach for Spectra Baseline Correction using*  
690 *Sparse Representation*. s.l., s.n.

691 Zimmerman, B. et al., 2017. a high-throughput FTIR spectroscopy approach to assess  
692 adaptive variation in the chemical composition of pollen. *Ecology and Evolution* , 7(24), pp.  
693 10839-10849.

694 Zimmerman, B. & Kohler, A., 2013. Optimizing Savitzky–Golay Parameters for Improving  
695 Spectral Resolution and Quantification in Infrared Spectroscopy. *Applied Spectroscopy*,  
696 67(8), pp. 892-902.

697 Zimmerman, B. & Kohler, A., 2014. Infrared Spectroscopy of Pollen Identifies Plant Species  
698 and Genus as Well as Environmental Conditions. *PLOS One*, 9(4), pp. 1-12.

699 Zimmerman, B. et al., 2016. Analysis of Allergenic Pollen by FTIR Microspectroscopy.  
700 *analytical chemistry*, Volume 88, pp. 803-811.

701 Zimmerman, B., Tkalčec, Z., Mešić, A. & Kohler, A., 2015. Characterizing Aeroallergens by  
702 Infrared Spectroscopy of Fungal Spores and Pollen. *PLoS ONE*, 10(4).

703

704

705 **Tables**706 *Table 1: Sample name, location of collection and correlating legend*

707	Sample name	Location	Legend
708	<i>Agrostis</i>	Holywell Pond	A
709	<i>Anthoxanthum odoratum</i>	Holywell Pond	A
710	<i>Deschampsia cespitosa</i>	Whitelee Moor	D
711	<i>Festuca ovina</i>	Whitelee Moor	C
712	<i>Molinia caerulea</i>	Milburn	B

713 *Table 2: Observed chemical absorption bands for each species.* Interpretation key: “(as)” =  
 714 asymmetric stretch, “(s)” = symmetric stretch, “(b)” = bending, “(d)” = deformation, “sh” =  
 715 shoulder, + signs = intensity of absorbance, “/” = absence of band and “~” = varying band  
 716 position.

Group	Wavenumber (cm-1)	<i>Agrostis</i>	<i>Anthoxanthum odoratum</i>	<i>Deschampsia cespitosa</i>	<i>Festuca ovina</i>	<i>Molinia caerulea</i>
-OH	3500-3000	+++	++	++	++	++
CH <sub>2</sub> (as)	2925	++	+	+	+	++
	2845	+	+	+	+	+
C=O	1740-1710	+	sh	+	+	+
C=O (amide I)	1650	++	++	+	++	+
C=C	~1600	sh	/	/	sh	/
N-H(b/d) (amide II)	~1550	sh	+	/	sh	/
C=C	~1515	/	/	1517 (sh)	/	1517 (sh)
CH <sub>2</sub> (d)	1460-1450	sh	/	/	sh	sh
	~1400	+	+	/	sh	/
CH <sub>3</sub> (s)(b)	1375	sh	/	sh	/	+
C-N(b)	~1325	/	sh	/	/	Sh
C-O	~1260	+	+	+	+	+
C-OH/C-O-C	1160	+	sh	+	+	+
	~1040	++	++	++	++	++
C-H(b)	~900	sh	sh	/	/	sh
	~800	/	sh	/	/	/
	~750	/	sh	/	/	sh
	~700	/	/	/	/	/

717

718 **SM1: Parameter Experiment**

719 Laura Scoble <sup>a</sup>, Simon J. Ussher <sup>a</sup>, Mark F. Fitzsimons <sup>a</sup>, Lauren Ansell <sup>b</sup>, Matthew Craven <sup>b</sup>,  
720 Ralph M. Fyfe <sup>a</sup>

721 <sup>a</sup> *School of Geography, Earth and Environmental Sciences, University of Plymouth,*  
722 *Plymouth, PL4 8AA, UK.*

723 <sup>b</sup> *School of Engineering, Computing and Mathematics, University of Plymouth, Plymouth,*  
724 *PL4 8AA, UK.*

725 **1. Introduction**

726 A previous critical literature review demonstrated inconsistencies in how FT-IR  
727 microspectroscopy had been applied across different studies, in particular the parameters used  
728 to generate spectra (particularly the scan rate and resolution). To address the impact of this  
729 inconsistency and address a knowledge gap in identifying the best approach, experiments  
730 were undertaken using replicate measurements from a single bulk sample from *Molinia*  
731 *caerulea*. This section presents the results from this methodological experimentation and  
732 suggests recommendations for standardised practice.

733 Two experiments were conducted, the first focussed on scan rate and the second on resolution  
734 ( $\text{cm}^{-1}$ ). The set-up variable was five different scan rates (16,32,64,128 and 256) and three  
735 resolutions (2, 4 and  $8\text{cm}^{-1}$ ).

736

## 737 **2. Methods**

### 738 *2.1 Sample preparation*

739 Fresh *Molinia caerulea* was collected from Northumbria Wildlife Trust, UK and used  
740 to create a bulk sample. Pollen grains were obtained by extracting four anthers from  
741 individual heads using tweezers and delicately scrapped out onto one half of a diamond anvil  
742 slide using a needle and scalpel. Pollen grains were compressed between the two halves of  
743 the anvil and then examined to see which half had the most sample.

### 744 *2.2 Chemical Analysis*

745 The Bruker Vertex 70 FT-IR bench unit with infrared microscopy on the Hyperion  
746 1000 was used to take ten replicate scans for each different parameter per experiment.  
747 Spectra recording was conducted between 4000 – 500 cm<sup>-1</sup> and generated using Bruker  
748 OPUS vers.4 software. Scans were exported as .csv files and manipulated within R v. 3.1.4  
749 (Team, 2022). Packages ggplot2 (Wickham, 2016) and tidyr (Wickham & Girlich, 2022)  
750 were used to plot spectra. Average spectra were created to plot the second derivatives in  
751 Origin (OriginLab, Northampton, MA, USA); for the purpose of these results, a smoother  
752 was not used.

### 753 *2.3 Data Analysis*

754 Data analysis was conducted on both scan rate and resolution data in R v.3.1.4 (Team,  
755 2022), and focused on a specific wavenumber (scan rate and resolution: 1654 cm<sup>-1</sup>) to  
756 compare and evaluate whether there was a significant statistical difference between each  
757 parameter. Mean absorbance units were calculated for each and plotted onto a boxplot for  
758 visual analysis. Null hypothesis stated (H<sub>0</sub>): all mean values were equal; the alternative  
759 hypothesis stated (H<sub>1</sub>): not all mean values were equal. If all mean values were equal then  
760 there was no significant difference between the scan rates/resolution and changing the

761 number didn't influence the overall spectrum. However, if all mean values weren't equal then  
762 it was concluded that there was a significant difference between the scan rates/resolution,  
763 which indicated that changing the number influenced the overall spectrum.

764 A one-way ANOVA model was used to determine whether the mean values across  
765 each parameter were equal ( $P = <0.05$ ), which provided quantification of whether increasing  
766 scan rate/resolution was significant and affected the overall spectrum. Tukey honestly  
767 significant difference (HSD) test was performed for pairwise comparison between means. A  
768 confidence level of 95% ( $>0.05$ ) was used, the p adj value (p-value) indicated whether there  
769 was a statistically significant difference between each pair or not. TukeyHSD test results  
770 were then manipulated, packages dplyr (Wickham, et al., 2023), multcomp (Hothorn, et al.,  
771 2008), emmeans (Lenth, 2023) and stringr (Wickham, 2022) was used to plot the data as a  
772 Compact Letter Display boxplot.

773 Full parameter methodology flowchart can be seen below (Figure SM1).

774



775

778

779

780

781

10 replicate scans per variable

Spectra recording: 4000-500 (cm-1)

10 replicate scans per variable

782

783

784

785

786

787

788

789

790

791

792

793

794

795

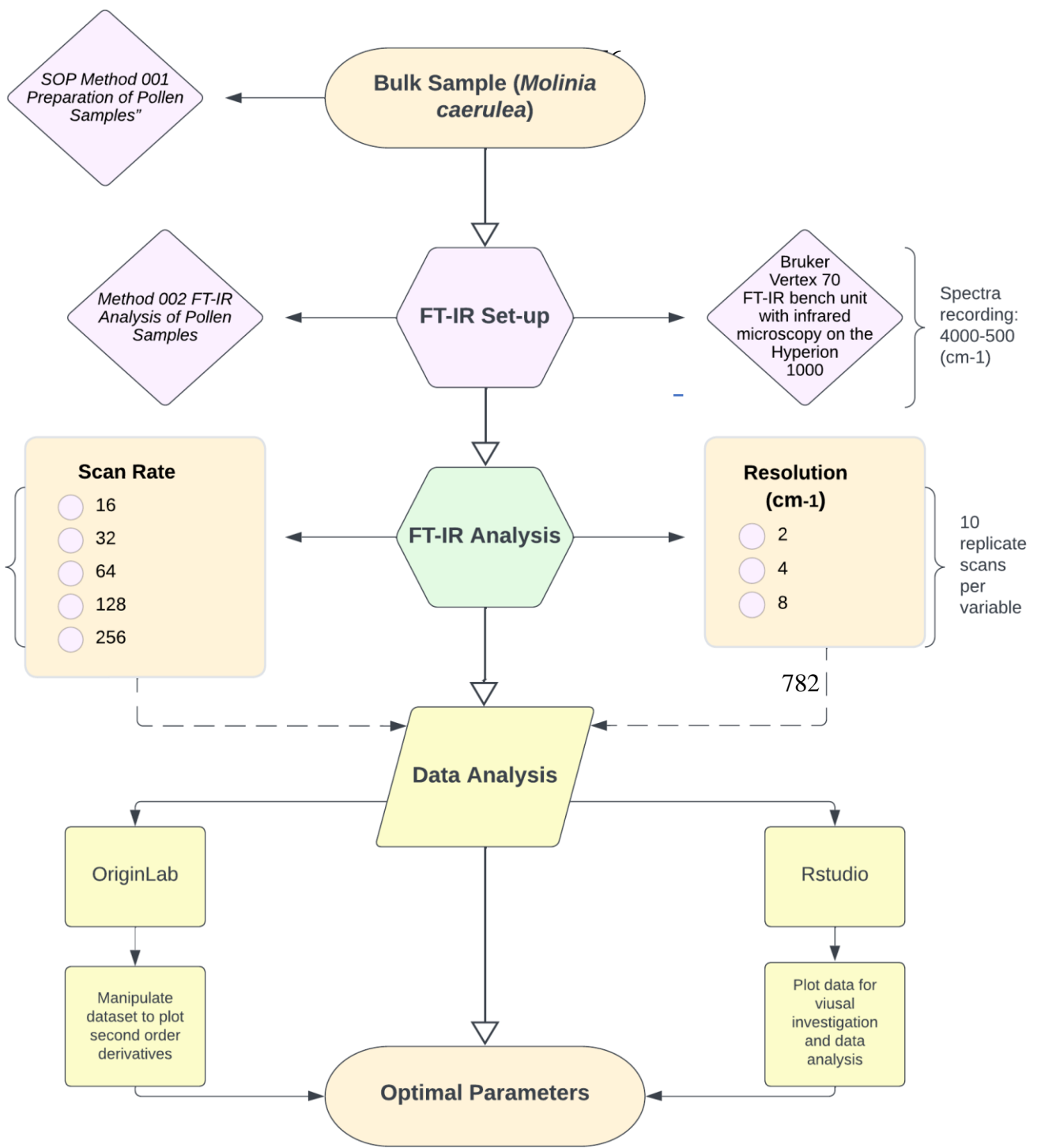


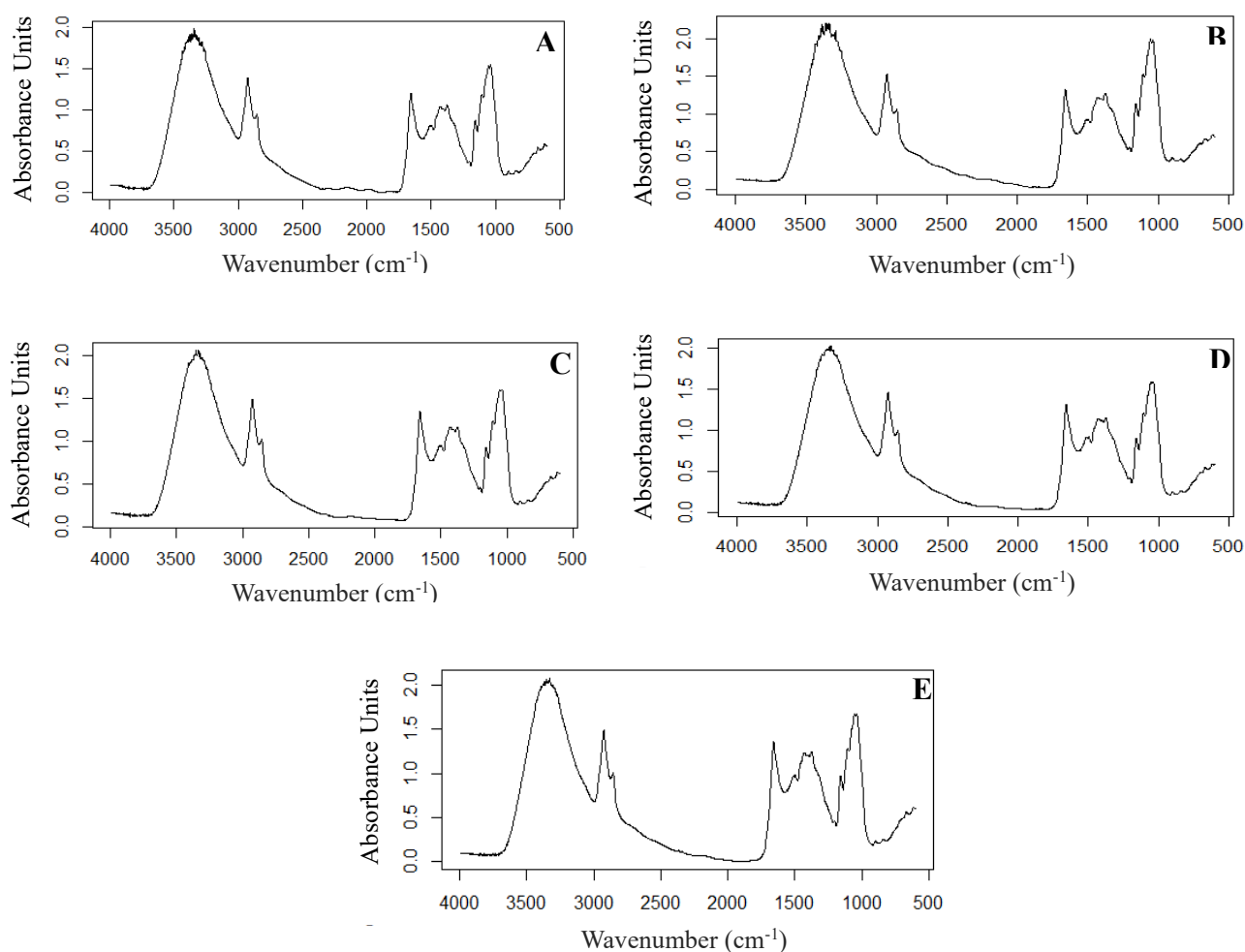
Figure SMI.1: Parameter methodology flowchart

(Scoble, 2023)

### 796 3. Results

#### 797 3.1 Scan rate

798 Five different scan rates are presented in Figure SM1.2: 16 (A), 32 (B), 64 (C), 128  
799 (D) and 256  $\text{cm}^{-1}$  (E) of *Molinia caerulea*. Lower scan rate numbers (A, B, C) exhibit higher  
800 scattering and noise, whereas the higher scan rates (D and E) are more detailed and smoother  
801 (Figure SM1.3). The broad -OH stretch ( $3300\text{cm}^{-1}$ ) has reduced noise exhibited in D and E,  
802 with the asymmetric  $\text{CH}_2$  ( $2923$  and  $2854\text{cm}^{-1}$ ) stretch exhibiting peak separation compared  
803 to the shouldering seen in A, B and C. The fingerprint region has a stronger level of  
804 absorbance in D and E, with C-O stretch ( $1163$  and  $1041\text{cm}^{-1}$ ) becoming more pronounced as  
805 the scan rate increases.



806

Figure SM1.2: 10 replicate scans of scan rate 16 (A), 32 (B), 64 (C), 128 (D) and 256 (E) of *Molinia caerulea*, averaged and plotted

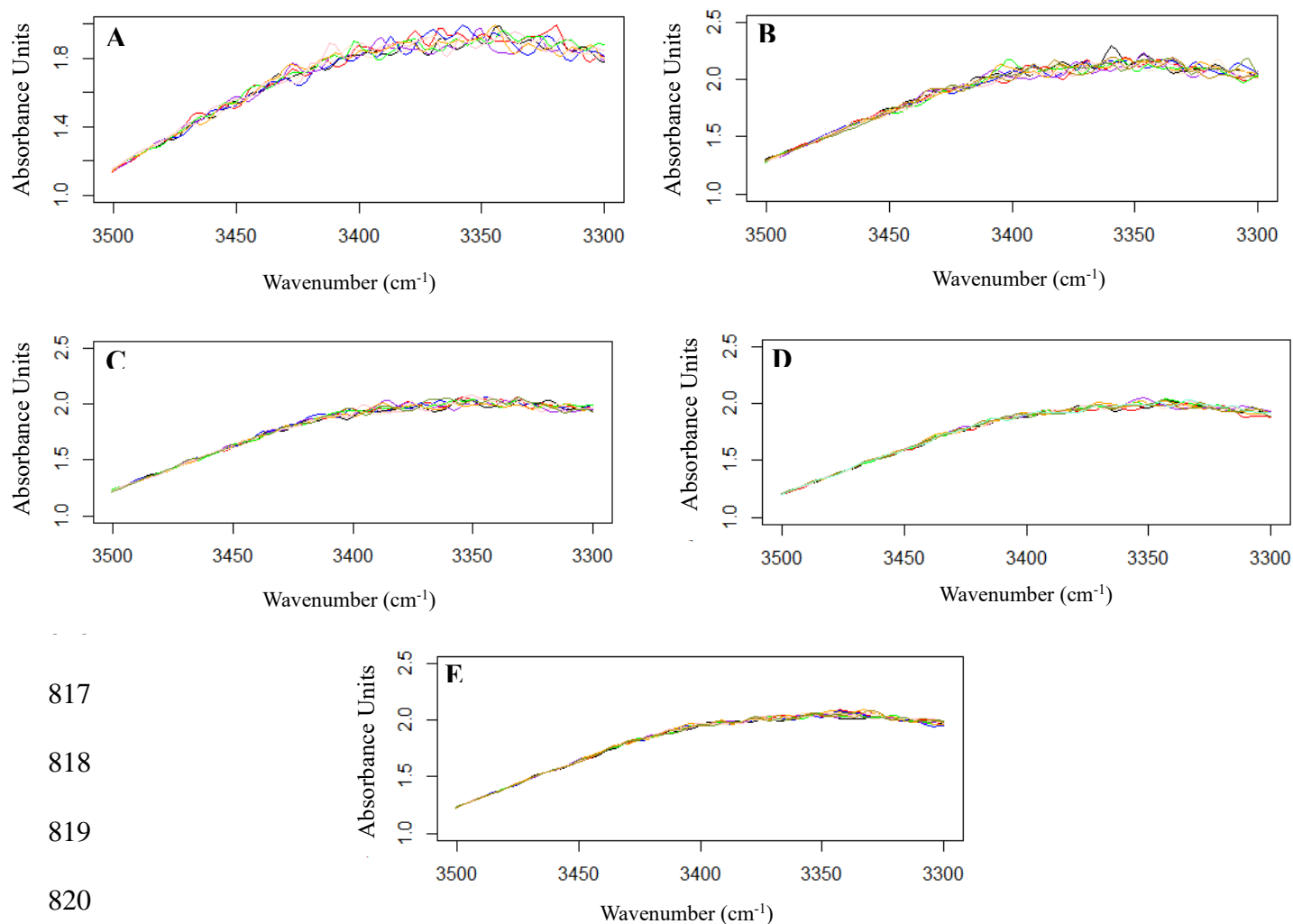
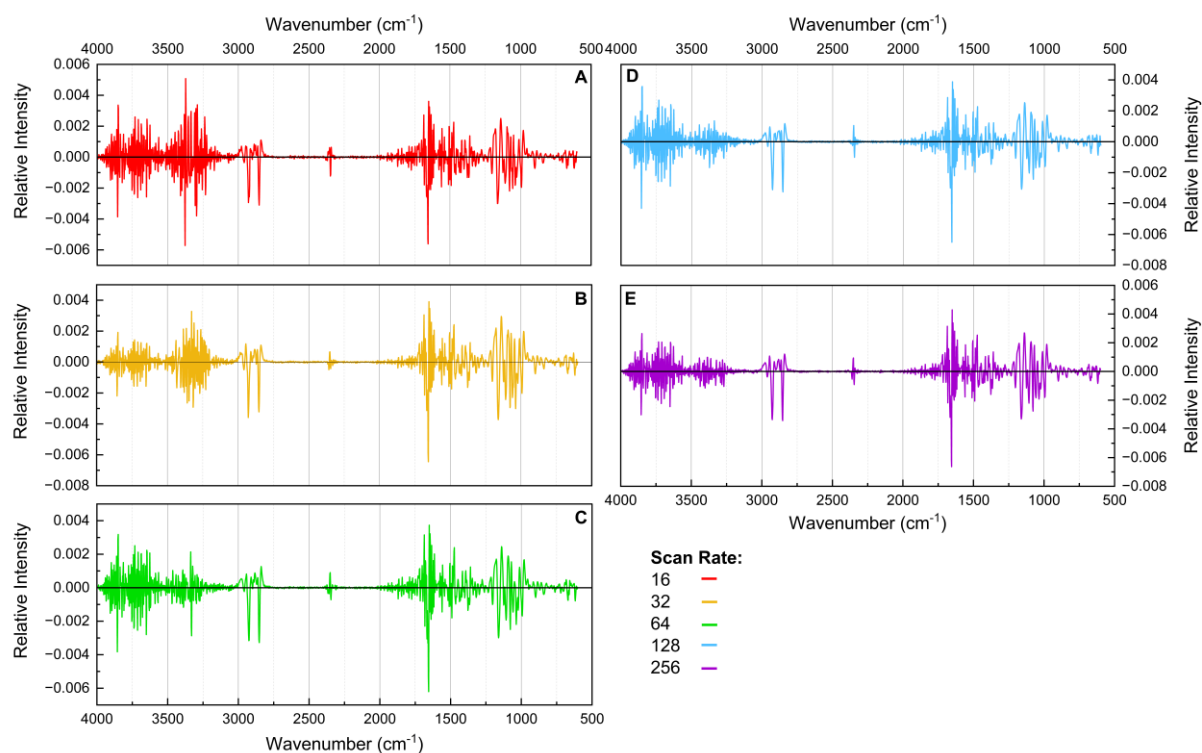


Figure SM1.3: Scattering of -OH peak at 3350  $\text{cm}^{-1}$  at different scan rates: 16 (A), 32 (B), 64 (C), 128 (D) and 256 (E) of *Molinia caerulea*.

### 3.1.2 Second Derivative

Figure SM1.4 presents second order derivatives of the scan rates, providing greater signal enhancement for chemical bands. There are similarities amongst all the scan rates, with key functional groups being present throughout and resolved peaks pointing downwards. As the scan rate increases the noise exhibited before at the -OH stretch ( $3300\text{cm}^{-1}$ ) reduces, with the broad band becoming nearly fully suppressed to baseline. Sharper upturned peaks between  $1700\text{-}1500\text{cm}^{-1}$  (C=O and C=C stretch) can be seen throughout Figure SM1.4 with D and E having more distinct separation between  $1250\text{-}1000\text{cm}^{-1}$  (C-O). The use of second

832 order derivatives has highlighted a new peak shown at roughly  $2400\text{ cm}^{-1}$ , indicative of a  
833 weak  $\text{C}\equiv\text{N}$  nitrile. Downturned peaks at roughly  $900$ ,  $800$  and  $700\text{ cm}^{-1}$  are more  
834 recognisable as aromatic rings and can be associated with sporopollenin bands.



835 *Figure SM1.4: Second derivatives of scan rates 16 (A), 32 (B), 64*  
836 *(C), 128 (D) and 256 (E) of Molinia caerulea.*

836 .

837

### 838 3.1.3 Data Analysis

839 Mean absorbance unit values for each scan rate were calculated by selecting a specific  
840 wavenumber from the fingerprint region ( $1654\text{ cm}^{-1}$ ) and tabulating the corresponding  
841 absorbance units for each replicate scan. The variable absorbance unit value depended on the  
842 variable scan rate; therefore, absorbance unit was treated as the dependant variable and the  
843 scan rate as the independent variable.

844

845

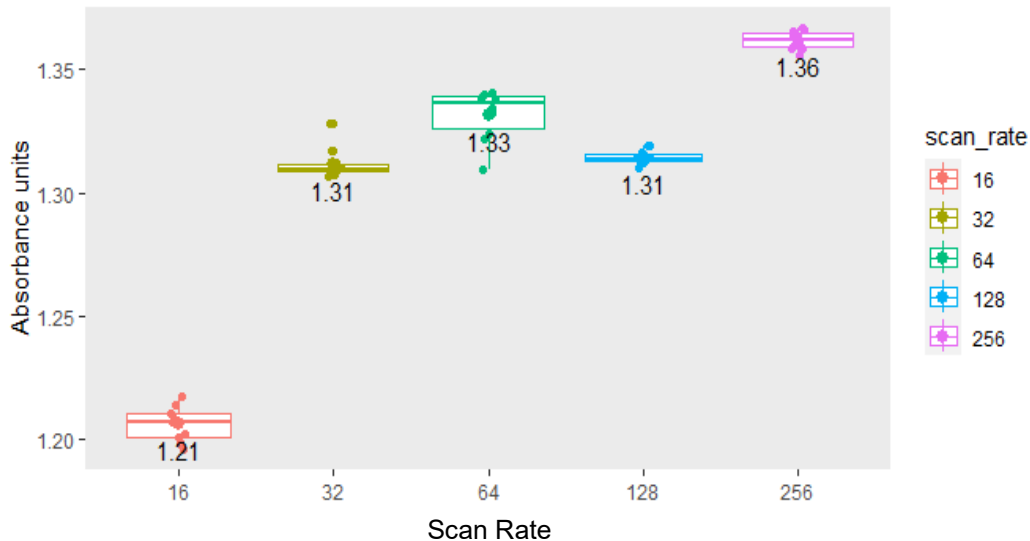


Figure SM1.5: Boxplot of scan rate mean absorbance unit values.

846

847

848 Figure SM1.5 shows box plots of the five scan rates and the mean absorbance unit

849 values. There is a noticeable increase in absorbance unit mean value as the scan rate is

850 increased to 64, then gradually decreases to 128 and then increases again at 256. 32 and 128

851 have the same mean value (1.31), suggesting there is no difference between both scan rates.

852 16 has a short boxplot with longer whiskers, indicating a wide distribution of data compared

853 to 32, 128 and 256. 32 exhibits a thin box plot and wider scattering, with two outliers at 1.320

854 and one at 1.316. 64 has the widest box plot with scattered data, indicating variance within

855 absorbance unit values for  $1654\text{cm}^{-1}$ . 128 exhibits tight clustering with a thin box plot,

856 indicating less variance within the data compared to between scan rate groups. 256 has the

857 highest mean value (1.36) and a thin box plot with tight clustering, most of the data points are

858 plotted around the mean.

859 An ANOVA test was run to determine whether the mean values were significantly

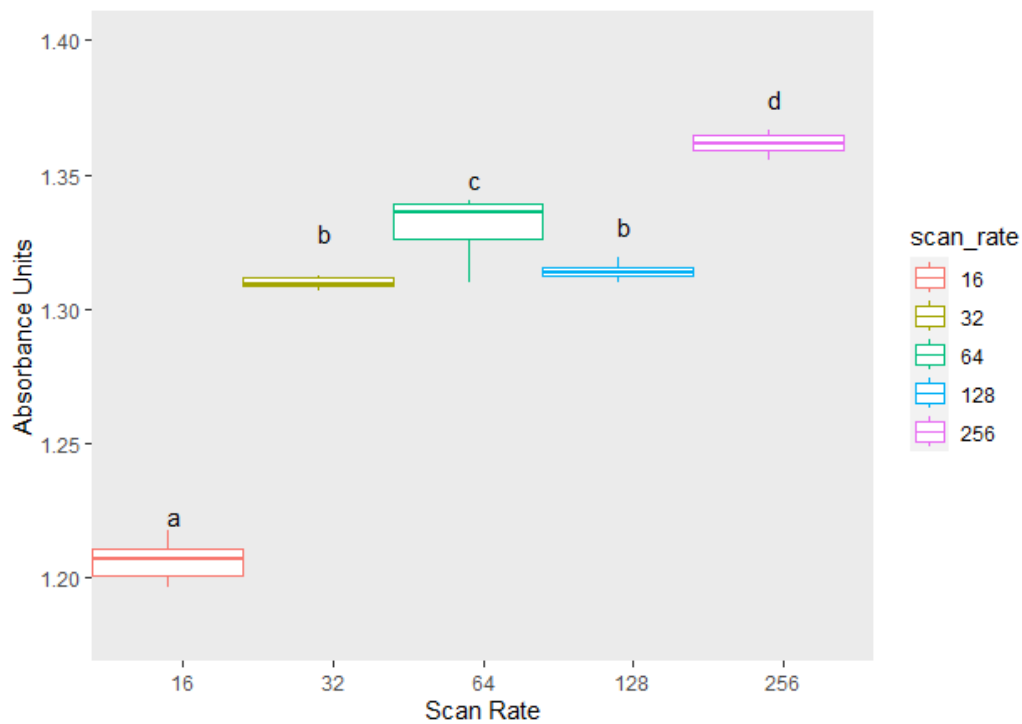
860 different from one another, working on the  $H_0$  hypothesis that all mean values were equal,

861 indicating there was no significant difference between scan rates. The  $p$ -value was

862 “ $1.503903\text{e-}40$ ” ( $1.503903 \times 10^{-40}$ ) which was  $<0.05$ , concluding that the mean values were

863 significantly different from one another. A Tukey Honestly Significant Difference

864 (TukeyHSD) test was used for pairwise comparisons. All pairs apart from 128-32 had a  $p$  adj  
865 value of “0.0e+00”, which was  $<0.05\%$ , indicating that there was a significant difference  
866 between each scan rate. 128-32 had a  $p$ -adj value of “0.8971323”, which was  $>0.05$ ,  
867 indicating no significant difference between 32 and 128. A Compact Letter Display (CLD)  
868 method was used to clarify the ANOVA and Tukey test output (Figure SM1.6).



869  
870  
871 Each scan rate had a specific lowercase letter and indicated that there is a statistically  
872 significant difference between all pairs of scan rates except 128-32. Therefore, null  
873 hypothesis ( $H_0$ ) is rejected and alternative ( $H_1$ ) is used concluding that changing the scan rate  
874 has an overall effect on the spectrum.

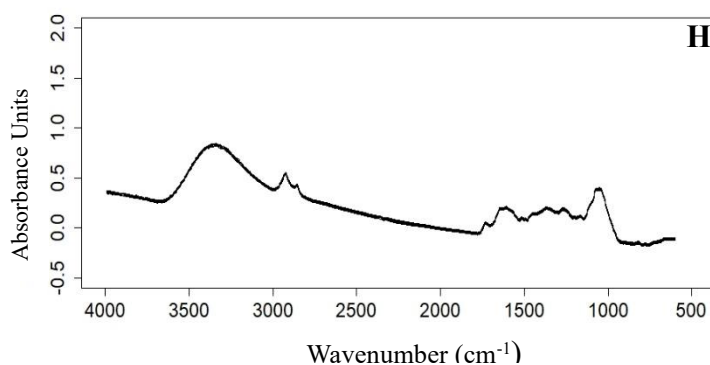
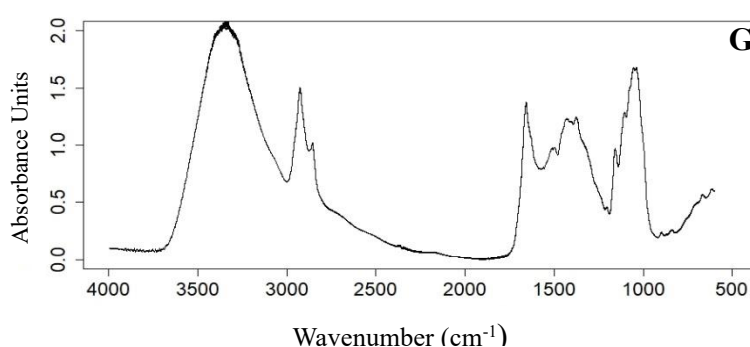
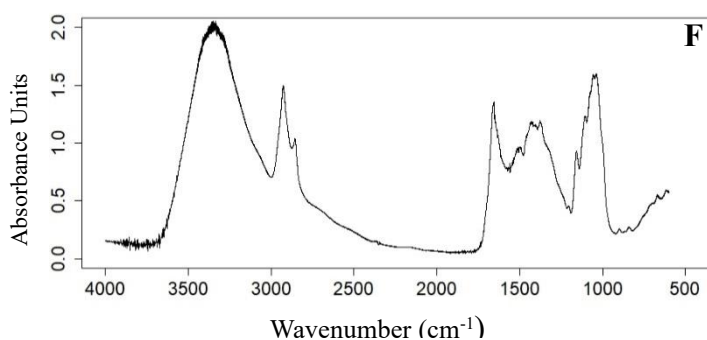
875

876

877

878 3.2 Resolution

879 Below are three different resolutions (Figure SM1.7):  $2\text{cm}^{-1}$  (F),  $4\text{cm}^{-1}$  (G) and  $8\text{cm}^{-1}$   
880 (H) using a scan rate of 256. F displays a noisy spectrum with a non-linear spectral line.  
881 Bands are well defined, but some appear to be sharp instead of broad because of the excess  
882 noise, e.g., -OH stretch ( $3300\text{cm}^{-1}$ ). The fingerprint region has some recognisable bands,  
883 however, the C=C shouldering at roughly  $1500\text{cm}^{-1}$  is challenging to identify. G has  
884 considerably less noise across the spectrum. Peaks and shoulders can be clearly differentiated  
885 as the spectral line looks more linear. The start of the spectrum is closer to the baseline and  
886 more distinguishable. H has a non-linear spectral line with weak absorbance. Bands are  
887 challenging to identify, especially within the fingerprint region, e.g. the anti-symmetric  $\text{CH}_2$   
888 bend ( $1433\text{cm}^{-1}$ ) and symmetric  $\text{CH}_3$  bend ( $1373\text{cm}^{-1}$ ). At the very end of the spectrum, the  
889 peaks dip below 0 absorbance.



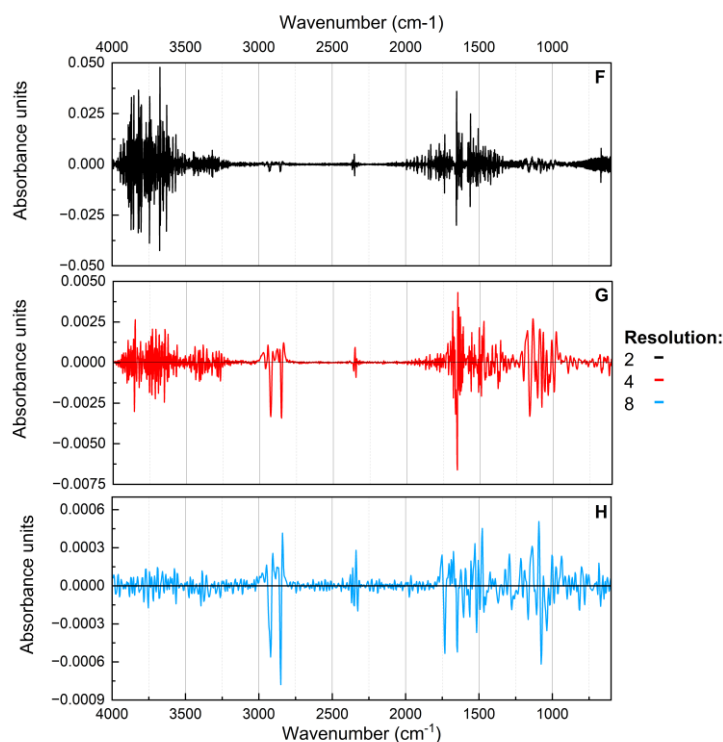
890

Figure SM1.7:10 replicate scans of resolution  $2$  (F),  $4$  (G) and  $8\text{cm}^{-1}$  (H) of *Molinia caerulea*, averaged and plotted using R.

891

892 3.2.1 Second derivatives

893 Figure SM1.8 presents second derivatives for the resolutions ( $\text{cm}^{-1}$ ). All three have  
894 very different absorbances, with F having the strongest and noisiest spectrum. F exhibits full  
895 suppression of the -OH stretch ( $3300 \text{ cm}^{-1}$ ), asymmetric  $\text{CH}_2$  stretch ( $2923$  and  $2854 \text{ cm}^{-1}$ ) and  
896 the C-O stretch ( $1163$  and  $1041 \text{ cm}^{-1}$ ). More obscure peaks cannot be identified as the  
897 spectrum is compacted. G exhibits a familiar spectrum with resolved peaks pointing  
898 downwards and a more defined fingerprint region. Strong C=C bands at roughly  $1600 \text{ cm}^{-1}$   
899 and C-O stretches between  $1100$ - $1000 \text{ cm}^{-1}$  are presented. Upward peaks can be seen between  
900  $1700$ - $1600 \text{ cm}^{-1}$  and  $1200$ - $1100 \text{ cm}^{-1}$ , indicative of a C=O and C-O stretch, respectively.  
901 Noise is still present at the beginning of the spectrum but not as strong. H has wider spacing  
902 between peaks, and very strong peaks. Resolved peaks have a strong negative absorbance  
903 with the asymmetric  $\text{CH}_2$  stretch ( $2854 \text{ cm}^{-1}$ ) measured at  $-0.0008$ . More pronounced  
904 upturned peaks are exhibited between  $1400$ - $1250 \text{ cm}^{-1}$ , which could be indicative of  
905 symmetric  $\text{CH}_3$  stretch.

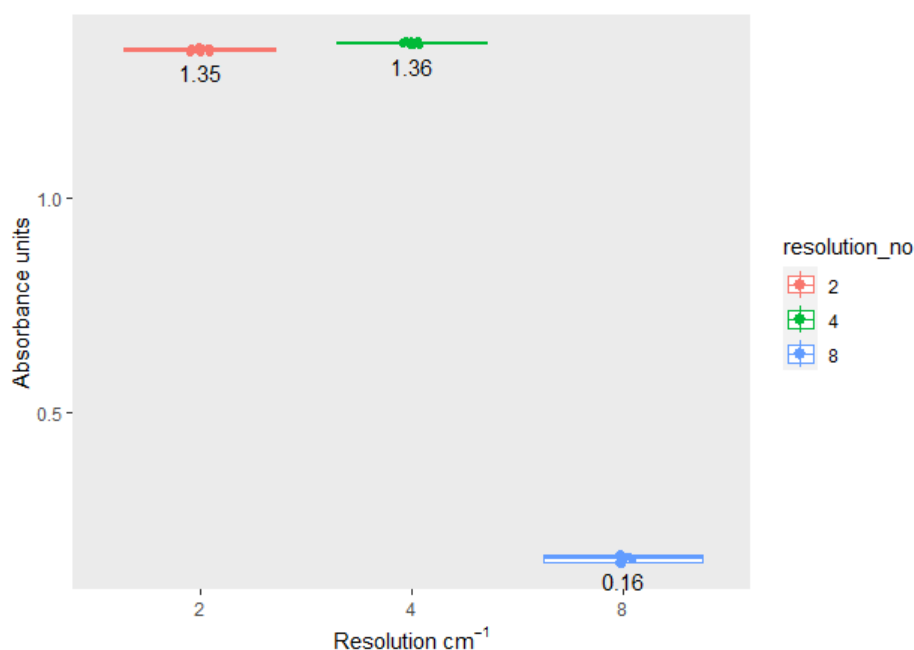


906 Figure SM1.8: second derivatives of resolution 2 (F), 4  
(G) and  $8 \text{ cm}^{-1}$  (H) of *Molinia caerulea*



907 3.1.3 Data analysis

908 Mean absorbance unit values for each resolution ( $\text{cm}^{-1}$ ) were calculated by selecting a  
909 specific wavenumber ( $1654 \text{ cm}^{-1}$ ) and tabulating the corresponding absorbance units for each  
910 replicate scan. The variable absorbance unit value depends on the variable resolution;  
911 therefore, absorbance unit is treated as the dependant variable and the resolution as the  
912 independent variable.

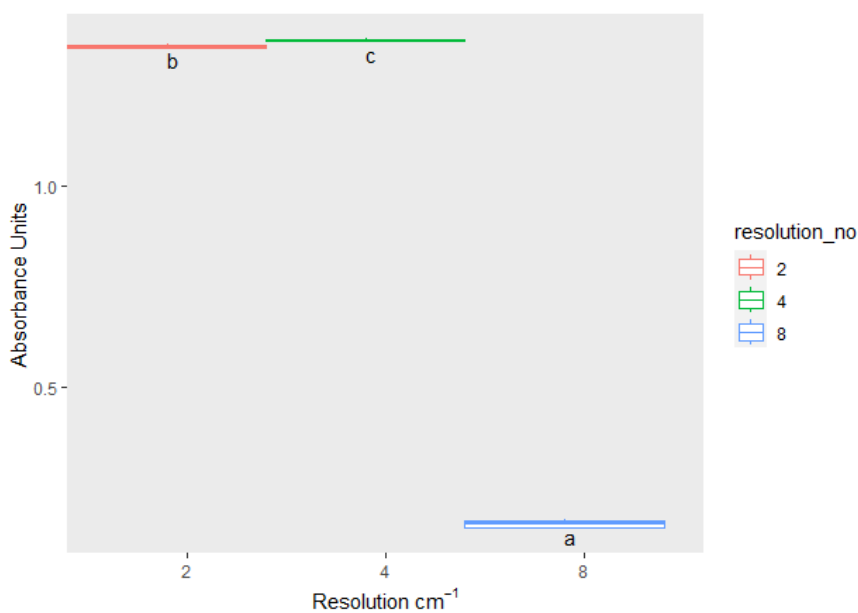


913 *Figure SM1.9: Boxplot of resolution mean*  
914 *absorbance unit value.*

915 Figure SM1.9 shows box plots of the three resolution numbers and the mean  
916 absorbance unit values. There is a subtle increase in absorbance unit value as the resolution is  
917 increased, until a rapid decrease between  $4\text{cm}^{-1}$  and  $8\text{cm}^{-1}$ .  $2\text{cm}^{-1}$  and  $4\text{cm}^{-1}$  exhibit tight  
918 clustering of data with thin box plots, indicating less variance within.  $8\text{cm}^{-1}$  has the lowest  
919 mean value of 0.16 and clustering dispersion is more spread out indicating more variance in  
920 the data.

921 An ANOVA test was run to determine whether the mean values were significantly  
922 different from one another. The  $p$ -value of the resolutions was “ $3.763198\text{e-}55$ ” ( $3.763198\times$   
923  $10^{-55}$ ) which is  $<0.05$ , concluding that the mean values are significantly different from one

924 another. A Tukey Honestly Significant Difference (TukeyHSD) test was used for pairwise  
925 comparisons All pairs apart from 4-2 ( $7.7e-06$ ) had a  $p$  adj value of “0.0e+00” which is  
926  $<0.05\%$ , indicating that there is a significant difference between each pair of resolutions ( $\text{cm}^{-1}$ )  
927 <sup>1</sup>). A Compact Letter Display (CLD) method was used to clarify the ANOVA and Tukey test  
928 output (Figure SM1.10).



929 .

930

931 Each resolution ( $\text{cm}^{-1}$ ) has a specific lowercase letter, and indicates that there is a  
932 statistically significant difference between all pairs. Therefore, null hypothesis ( $H_0$ ) is  
933 rejected and alternative ( $H_1$ ) is used concluding that changing the resolution ( $\text{cm}^{-1}$ ) has an  
934 overall effect on the spectrum.

935

#### 936 4. Discussion

937 Most analytical studies operationally define measurement parameters such as scan  
938 rate and resolution (Barra, et al., 2021), or base it on the suppliers' recommendations.  
939 Software such as Bruker OPUS spectroscopy provides spectrum acquisition for numerous  
940 analytical instruments, e.g Bruker Hyperion 1000 FT-IR Microscope. This includes scan rate  
941 and resolution measurement parameters but provides no in-depth explanation as to why these  
942 specific parameters have been chosen. Research surrounding FT-IR microspectroscopy pollen  
943 identification suggest the optimal parameters are 256 scan rate and  $4\text{cm}^{-1}$  resolution (Julier, et  
944 al., 2016) (Jardine, et al., 2019). However, as discussed in chapter 7's systematic review,  
945 there were inconsistencies in how the scan rate and resolution had been used to generate  
946 spectrum using FTIR microspectroscopy. To address the knowledge gap, experiments were  
947 undertaken on scan rate and resolution using ten replicate measurements from a bulk sample  
948 and from a single species (*Molinia caerulea*) to find the optimal parameters.

949 When analysing organic material, scan rates are crucial; the higher the scan rate, the  
950 more scans are performed. After 50 scans, the spectrum acquisition noticeably improves, with  
951 influential absorption bands becoming more prominent in the "fingerprint region." As shown  
952 in Figure SM1.2 and Figure SM1.3's comparison of the scan rates, 256 (E) has less noise and  
953 scattering. It exhibits a smooth spectrum with prominent peaks, essential for analysing  
954 functional groups for identification. While 64 (C) and 128 (D) exhibit a smooth spectrum in  
955 comparison to 16 (A) and 32 (B), 256 (E) offers additional detailing, such as pronounced  
956 shouldering and easier functional group recognition.

957 Figure SM1.4 presents the conversion of scan rate spectra into second derivatives.  
958 Second derivatives aid in chemical band interpretation, as it can resolve overlapping analyte  
959 signals by enhancing the signals within vibrational spectra (Kohler, et al., 2020). There were  
960 similarities across all five scan rates with resolved broad peaks pointing downwards. As scan

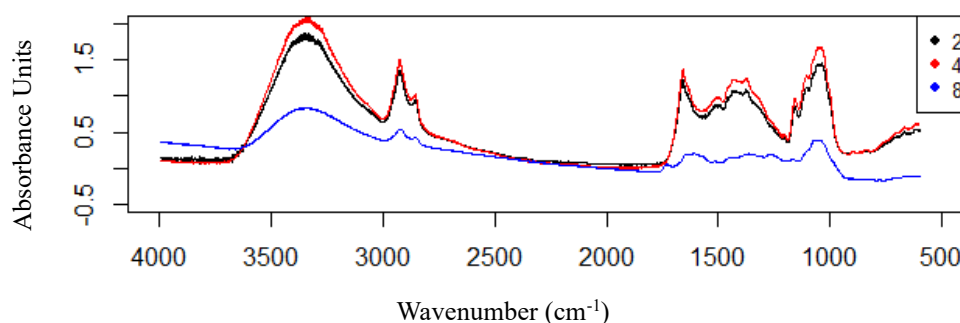
961 rate increases, there is a gradual reduction in noise and sharp peaks before the -OH stretch  
962 ( $3300\text{cm}^{-1}$ ). Upturned peaks separate out clearer in 128 (D) and 256 (E) compared to 16 (A),  
963 32 (B), and 64 (C) between  $1700\text{-}1500\text{cm}^{-1}$  and  $1250\text{-}1000\text{cm}^{-1}$ . Comparing this to Figure  
964 SM1.2, 128 (D) and 256 (E) have more pronounced peaks at roughly  $1650\text{cm}^{-1}$  (C=C) and  
965 shouldering between  $1150\text{-}1000\text{cm}^{-1}$  (C-O). A new peak at roughly  $2300\text{cm}^{-1}$  appeared within  
966 the second order derivatives, this could be indicative of a weak  $\text{C}\equiv\text{N}$  stretch as the peak was  
967 obscured across Figure SM1.2.

968 Resolution is considered as “*the ability to separate two spectral lines that are very*  
969 *close in wavelength or frequency*” (Schlindwein, 2020). If two IR absorption bands are  
970 similar, the resolving power must be increased to separate them. Typically, the type of  
971 material being analysed determines the resolution number. Since the absorption bands are  
972 narrow for gases, the vibration of the atoms is measured at a wavelength of  $0.2$  to  $0.5\text{ cm}^{-1}$ .  
973 As solids and liquids have wide absorption bands, choosing a value lower than  $2\text{cm}^{-1}$  would  
974 not provide any more information (Schlindwein, 2020).

975 Figure SM1.7 exhibits the resolution spectra and demonstrates that there is greater  
976 noise with a non-linear spectral line when the number is reduced to  $2\text{cm}^{-1}$  (F). Noise can be  
977 decreased by scanning the sample immediately after the background scan. Lowering the  
978 resolution lengthens the time between scans, increasing the likelihood of noise. A background  
979 scan would have to be conducted more frequently if  $2\text{cm}^{-1}$  resolution was used, making this  
980 less time efficient. Absorption bands are well defined, but some appear sharp instead of  
981 broad, e.g. -OH stretch ( $3300\text{cm}^{-1}$ ). Across the fingerprint region, bands are distinguished,  
982 however, the C=C shouldering at roughly  $1500\text{cm}^{-1}$  is challenging to identify. Comparing this  
983 to  $4\text{cm}^{-1}$  (G), the spectral line is deemed linear as there is a reduction in noise. Peaks and  
984 shoulders can be clearly differentiated, and the start of the spectrum is closer to baseline.  
985  $8\text{cm}^{-1}$  (H) has a non-linear spectral line with weak absorbance. Increasing the resolution

986 shortens the time between scans, decreasing the degree of fineness obtained (Ota, 2007).  
987 Chemical signals are difficult to identify especially within the fingerprint region, e.g the anti-  
988 symmetric  $\text{CH}_2$  bend ( $1433\text{ cm}^{-1}$ ) and symmetric  $\text{CH}_3$  bend ( $1373\text{ cm}^{-1}$ ). At the very end of  
989 the spectrum the peaks dip below 0 absorbance. If this was seen across the other resolution  
990 spectra, this could be indicative that the background scan was taken incorrectly or the ATR  
991 cell wasn't cleaned sufficiently beforehand. However, as the negative absorbance is only  
992 present in  $8\text{ cm}^{-1}$  (H) fingerprint region, it is highly probable that this is a result of the lack of  
993 resolving power and detail.

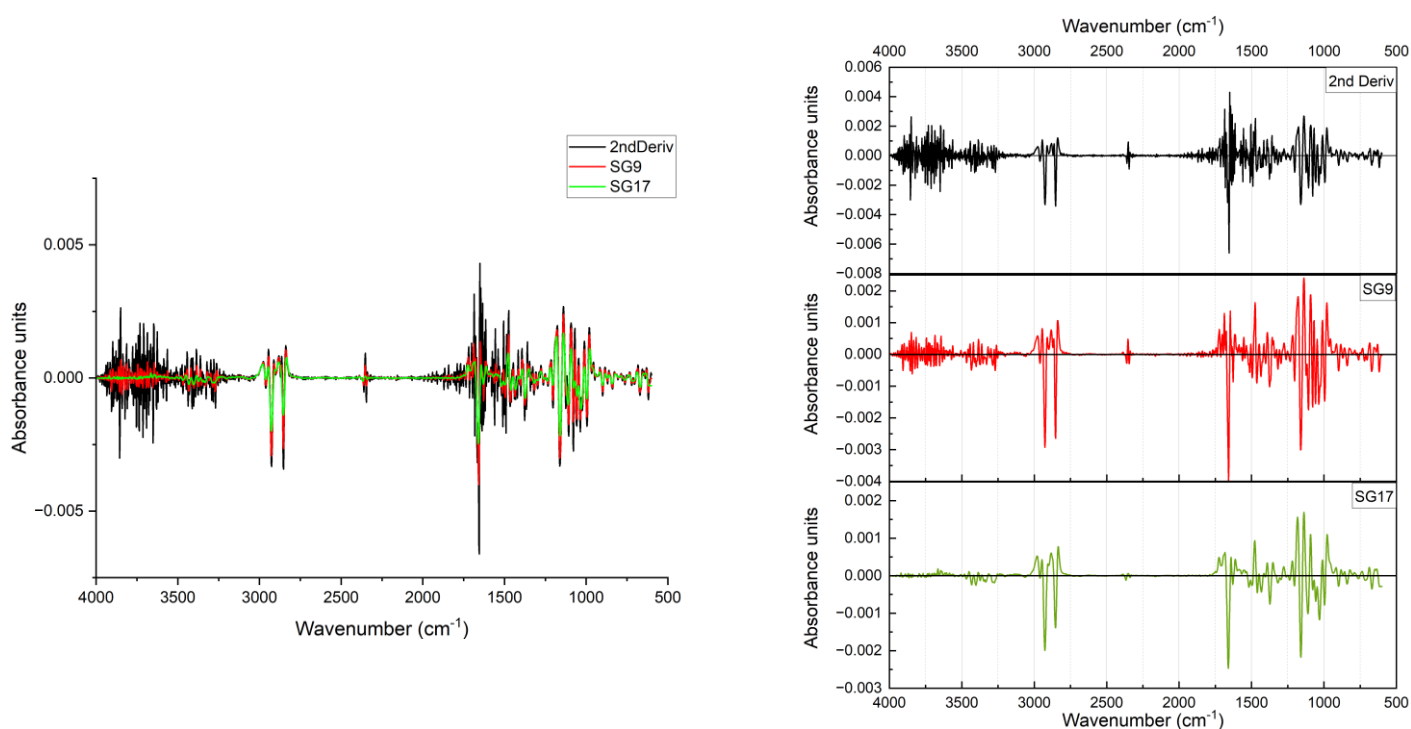
994 When stacking and comparing the spectra (Figure SM1.11), it is visually clear that 4  
995  $\text{cm}^{-1}$  resolution (red) compared to  $2\text{ cm}^{-1}$  provides a smooth spectrum with negligible noise  
996 between  $4000\text{-}3500\text{ cm}^{-1}$ , the  $\text{-OH}$  peak at  $3300\text{ cm}^{-1}$ , and the fingerprint region. When  
997 compared to  $8\text{ cm}^{-1}$  resolution (blue), the spectrum is barely distinguishable.



998 *Figure SM1.11: Average scans of resolution 2,4 and  $8\text{ cm}^{-1}$*   
999 *of Molinia caerulea.*

1000 Figure SM1.8 presents the conversion of resolution spectra into second order  
1001 derivatives. All three have very different strengths of absorbance, with  $2\text{ cm}^{-1}$  (F) having the  
1002 strongest and noisiest spectrum. As stated previously, increasing the resolution increases the  
1003 likelihood of noise. One way to decrease noise would be to use a smoothing algorithm such  
1004 as the Savitzky-Golay. This multifunctional pre-processing algorithm can be used for noise-

1005 reduction through the function of smoothing (Savitzky & Golay, 1964). It defines a moving  
1006 window which smooths out the spectrum, increasing the window size causes the smoothing  
1007 intensity to intensify. However, this can lead to loss of valuable chemical information and  
1008 analyte signals (Kohler, et al., 2020). Figure SM1.12 is an example of using the Savitzky-  
1009 Golay smoothing feature on  $4\text{cm}^{-1}$  (G) second order derivative (black). A window size of 9  
1010 (red) and 17 (green) was used with a polynomial of 2.



1011 *Figure SM1.12: Multiple line plot (a) and stacked line plot (b) of resolution  $4\text{cm}^{-1}$  (G) second order*  
1012 *derivative (black), Savitzky-Golay (SG) smoothing – window size 9 (red) and size 17 (green).*

1013 Across Figure SM1.8 and SM1.12,  $4\text{cm}^{-1}$  (G) exhibits a familiar spectrum with  
1014 resolved peaks pointing downwards and a more defined fingerprint region. It has a greater  
1015 level of detail compared to  $8\text{cm}^{-1}$  (H), but not an excess where the spectrum becomes noisy  
1016 and hard to interpret as seen in  $2\text{cm}^{-1}$  (F). Comparing line plots (Figure 8.13 (a/b)), the  
1017 second order derivative spectral line (black) is noisy whereas the two SG lines (red)(green)  
1018 display distinct peaks. SG9 (red) presents strong downturned resolved C=C bands at roughly  
1019  $1600\text{cm}^{-1}$  and C-O stretches between  $1100\text{-}1000\text{cm}^{-1}$ . Noise is still present at the beginning of

1020 the spectrum but not as strong. The -OH stretch ( $3300\text{cm}^{-1}$ ) has been slightly suppressed but is  
1021 still identifiable, whereas SG17 (green) has suppressed it more intensively. This suppression  
1022 is a clear example of loss of chemical information as a direct result of a larger window size.  
1023 While SG17 (green) has over-suppressed resolved peaks, the fingerprint region has excellent  
1024 separation between upturned peaks. Peaks can be identified between  $1700\text{-}1600\text{cm}^{-1}$  and  
1025  $1200\text{-}1100\text{cm}^{-1}$ , indicative of C=O and C-O stretches. The application of the Savitzky-Golay  
1026 algorithm across IR spectra can be beneficial if absorbance bands are difficult to distinguish.  
1027 However, users must be cautious when choosing a window size as this could lead to an over  
1028 suppression of analyte signal and ultimately loss of chemical information.

1029 Data analysis of both scan rate and resolution indicated that there is a statistically  
1030 significant difference when the numbers are changed. Increasing the scan rate saw an overall  
1031 increase in the mean absorbance values, clustering of data also became more compact with  
1032 less range (Figure SM1.5). The linear model ( $p$ -value:  $2.2\text{e-}16$ ) and ANOVA ( $p$ -value:  
1033  $1.503903\text{e-}40$ ) tests output indicated that there was a significant difference between the scan  
1034 rates mean values ( $p$ -value:  $<0.05$ ). Tukey test and CLD method (Figure SM1.6) was used to  
1035 clarify these outputs by comparing pairs, determining that all pairs of scan rates apart from  
1036 128-32 were significantly different as the  $p$ -adj was  $<0.05$ . 128-32 had a  $p$ -adj value of  
1037 “0.8971323”, concluding that this pair is not significantly different. Therefore, the null  
1038 hypothesis ( $H_0$ ) is rejected, and the alternative ( $H_1$ ) is used, signifying that increasing the scan  
1039 rate makes a significant difference in the overall spectrum. When comparing 16, 64, and 256  
1040 to determine which scan rate offers the best consistency and less variance, 256 has the  
1041 thinnest box plot with tighter clustering of data. Along with less range compared to 16 and  
1042 64, this indicates less variance within 256’s dataset compared to variance between groups.

1043 Changing the resolution ( $\text{cm}^{-1}$ ) exhibited similar results (Figure SM1.9). While the  
1044 difference between  $2\text{cm}^{-1}$  (1.35) and  $4\text{cm}^{-1}$  (1.36) mean absorbance values wasn’t visually

1045 significant,  $8\text{cm}^{-1}$  saw a rapid decline to 0.16. Clustering also became less compact as the  
1046 resolution was increased from  $4\text{cm}^{-1}$  to  $8\text{cm}^{-1}$ , suggesting data became more variable within  
1047 the group. The linear model ( $p$ -value:  $2.2\text{e-}16$ ) and ANOVA ( $p$ -value:  $3.763198\text{e-}55$ ) tests  
1048 output indicated there was significant difference between the resolution mean values ( $p$ -  
1049 value:  $<0.05$ ). Tukey test and CLD method was used for pairwise comparisons, determining  
1050 that all resolution pairs were significantly different from one another as the  $p$ -adj values were  
1051  $<0.05$  (Figure SM1.10). Therefore, the null hypothesis ( $H_0$ ) is rejected, and the alternative  
1052 ( $H_1$ ) is used, signifying that increasing the resolution significantly affects the overall  
1053 spectrum. However,  $8\text{cm}^{-1}$  sees wider variance (Figure SM1.9) and loss of chemical  
1054 information (Figure SM1.7), whereas  $2\text{cm}^{-1}$  has tight clustering (Figure SM1.9) but an  
1055 incredibly noisy spectrum making identification difficult (Figure SM1.8F).  $4\text{cm}^{-1}$  has tight  
1056 clustering, a thin boxplot (Figure SM1.9) and identifiable peaks (Figure SM1.7 and SM1.8G),  
1057 indicating less variance within the dataset and better consistency compared to the other two  
1058 resolutions.

1059

1060



1061 **5. Conclusion**

1062           When altered, scan rate and resolution can affect the generated average spectrum.  
1063 Understanding what the optimal parameters are for FT-IR pollen analysis will ultimately lead  
1064 to a more successful identification of functional groups and classification. Scan rate is crucial  
1065 as the more scans taken improves the spectrum acquisition, while increasing the resolution  
1066 aids in separating two similar absorption IR bands. The combined systematic review,  
1067 laboratory experiments and data analysis meant a comparison could be made between the  
1068 analytical methods and chosen parameters against the results above. Overall, 256 scan rate  
1069 and 4cm<sup>-1</sup> resolution are the best parameters for pollen identification. The only published  
1070 study that uses these parameters is Jardine et al (2019). 256 has reduced noise and scattering,  
1071 exhibiting a smooth spectrum with prominent peaks - essential for analysing functional  
1072 groups and identifying morphologically indistinct pollen families. 4cm<sup>-1</sup> provides enough  
1073 separation for IR absorption bands to be identifiable with minimal noise. To build reference  
1074 libraries of spectra that can be shared and used by other researchers, the scan rate and  
1075 resolution should be standardised using these parameters.

1076

1077 **5. Bibliography**

- 1078 Barra, I. et al., 2021. Optimizing setup of scan number in FTIR spectroscopy using the  
1079 moment distance index and PLS regression: application to soil spectroscopy. *Scientific*  
1080 *Reports*, Volume 11, p. 13558.
- 1081 Hothorn, T., Bretz, F. & Westfall, P., 2008. Simultaneous Inference in General Parametric  
1082 Models. *Biometrical Journal*, 50(3), pp. 346-363.
- 1083 Jardine, E. P. et al., 2019. Chemotaxonomy of domesticated grasses: a pathway to  
1084 understanding the origins of agriculture. *Micropalaeontol*, Volume 38, pp. 83-95.
- 1085 Julier, C. M. A. et al., 2016. Chemotaxonomy as a tool for interpreting the cryptic diversity of  
1086 Poaceae pollen. *Review of Palaeobotany and Palynology*, Volume 235, pp. 140-147.
- 1087 Kohler, A. et al., 2020. Model-Based Pre-Processing in Vibrational Spectroscopy. In:  
1088 *Comprehensive Chemometrics*. s.l.:Elsevier, pp. 83-100.
- 1089 Lenth, R., 2023. *emmeans: Estimated Marginal Means, aka Least-Squares Means*. [Online]  
1090 Available at: <https://CRAN.R-project.org/package=emmeans>  
1091 [Accessed 17 March 2023].
- 1092 Ota, H., 2007. Resolution and Aperture. *FTIR Talk Letter Vol.8*, October, pp. 02-07.
- 1093 Savitzky, A. & Golay, M. J. E., 1964. Smoothing and Differentiation of Data by Simplified  
1094 Least Squares Procedures. *Anal. Chem*, p. 36.
- 1095 Schlindwein, H. S., 2020. *About Spectral Resolution in FT-IR Spectroscopy*. [Online]  
1096 Available at: <https://www.opticsblog.bruker.com/guide-to-spectral-resolution-in-ft-ir/>  
1097 [Accessed 13 January 2023].
- 1098 Scoble, L., 2023. *Parameter flowchart*. Plymouth: s.n.
- 1099 Team, R. C., 2022. *R: a language and environment for statistical computing*. Vienna: R  
1100 Foundation for Statistical Computing.
- 1101 Wickham, H., 2016. *ggplot2: Elegant Graphics for Data Analysis*. New York: Springer-  
1102 Verlag.
- 1103 Wickham, H., 2022. *stringr: Simple, Consistent Wrappers for Common String Operations*.  
1104 [Online]  
1105 Available at: <https://CRAN.R-project.org/package=stringr>  
1106 [Accessed 17 March 2023].
- 1107 Wickham, H. et al., 2023. *dplyr: A Grammar of Data Manipulation. R Package 1.1.0*.  
1108 [Online]  
1109 Available at: <https://CRAN.R-project.org/package=dplyr>  
1110 [Accessed 15 March 2023].

1111 Wickham, H. & Girlich, M., 2022. *tidyr: Tidy Messy Data*. [Online]  
1112 Available at: <https://CRAN.R-project.org/package=tidyr>  
1113 [Accessed 13 January 2023].  
1114  
1115

1116 #####

1117 ## Scoble, L (2023) R script for ##

1118 ## pre-processing spectral data, data analysis ##

1119 ## and classification. ##

1120 #####

1121 **Table of Contents**

1122 Set up: 2

1123 Baseline, EMSC correction and 2nd Derivative: 4

1124 Plot non-differentiated spectra following parts of Jardine (2021) R script: 7

1125 PCA Analysis: 11

1126 HCA Plot: 14

1127 Decision Trees: 16

1128 randomForest: 23

1129 MDA wavenumber boxplot: 31

1130 References: 34

1131

1132

1133

1134

1135

```
1136 #####
1137 ## Set up working directory and load libraries      ##
1138 ## Scoble, L and Fyfe, R, (2023)                  ##
1139 #####
1140
1141 setwd("D:\\ALL DATA2\\baseline work")
1142
1143 library(corrplot)
1144 library(caret)
1145 library(tidyverse)
1146 library(class)
1147 library(prospectr)
1148 library(ggplot2)
1149 library(grid)
1150 library(baseline)
1151 library(EMSC)
1152 library(vegan)
1153 library(dendextend)
1154 library(circlize)
1155 library(ape)
1156 library(RColorBrewer)
1157 library(randomForest)
1158 library(dplyr)
1159 library(tree)
1160 library(caTools)
1161 library(tidyverse)
1162 library(readr)
1163 library(rpart)
1164 library(rpart.plot)
1165 library(reshape2)
1166
1167
1168
1169
```

```

1170  ## stick all scans together
1171  ## requires individual files with a .dpt extension
1172
1173
1174  #list all files (with .dpt extension)
1175  file.list <- list.files(pattern = "\\dpt$") #only lists dpt files
1176  #make empty dataframe
1177  df <- read.csv(file.list[1], header = F)
1178  df <- rbind(c("wavelength", "Agrostis"), df)
1179
1180  #loop across all files and stick together into single file
1181  count = 1
1182
1183  for(i in file.list){
1184    print(paste("count =", count, i)) #flag for progress
1185    #read in individual file
1186    dat <- read.csv(i, header = F)
1187    #extract sample code from filename
1188    sample <- gsub(".dpt", "", i)
1189    #append the data to the dataframe
1190    dat <- rbind(c("wavelength", sample), dat)
1191    df <- cbind(df, dat[,2])
1192
1193    count = count + 1
1194  }
1195
1196  #prepare the combined file for export
1197  df <- df[,-2] #drops col 2 (duplicate data)
1198  colnames(df) <- df[1,] #define column names as sample names
1199  df <- df[-1,] #drop top row (the un-needed names)
1200
1201  #export to csv format
1202  write.csv(df, "all.data.scans.final.csv", row.names = F)
1203

```

```

1204 #####
1205 ## Baseline, EMSC correction and 2nd Derivative      ##
1206 ## Scoble, L (2023)                                ##
1207 #####
1208
1209 ##### Read data in #####
1210 Species_data <- read.csv("all.data.scans.final.csv", check.names = F)
1211 Species_data <- data.frame(t(Species_data))
1212 colnames(Species_data) <- Species_data[1,]
1213 Species_data <- Species_data[-1,]
1214
1215 ##### Non-differentiated spectra #####
1216
1217 # Baseline correction
1218 species.baseline <- baseline(as.matrix(Species_data), method = "modpolyfit", deg = 2)
1219 species.corrected <- data.frame(species.baseline@corrected)
1220 colnames(species.corrected) <- colnames(Species_data)
1221 species.corrected <- as.data.frame(species.corrected,
1222                                   row.names = rownames(Species_data))
1223
1224 # EMSC correction of baseline corrected data
1225 Species.emsc1 <- EMSC(species.corrected, degree = 3,
1226                       reference = colMeans(species.corrected))
1227
1228
1229 emsc.corrected1 <- data.frame(t(Species.emsc1$corrected))
1230
1231 # Write file
1232 write.csv(emsc.corrected1, "all.data.emsc.baseline.final.csv", row.names = T)
1233
1234
1235
1236
1237

```

```

1238 ##### Second derivative of original data #####
1239 Species.derivtwo <- savitzkyGolay(Species_data, p = 2, w = 15, m = 2)
1240
1241 Species.derivtwo <- as.data.frame(Species.derivtwo, row.names = rownames(Species_data))
1242
1243 # EMSC correction of second derivative data
1244 Species.deriv.emsc <- EMSC(Species.derivtwo, degree = 1,
1245     reference = colMeans(Species.derivtwo))
1246
1247 Species.deriv.emsc <- Species.deriv.emsc$corrected
1248
1249 # Write file
1250 Species.deriv.emsc.pca <- data.frame(Species.deriv.emsc)
1251 names(Species.deriv.emsc.pca) <- sapply(str_remove_all(colnames(Species.deriv.emsc), "X"), "[")
1252
1253 write.csv(Species.deriv.emsc.pca, "Species.deriv.emsc.final.csv")
1254
1255
1256 ##### Prepare derivative file for OriginLabs #####
1257
1258 # Remove row names into column
1259 Species.deriv.emsc.o <- cbind(rownames(Species.deriv.emsc.pca),
1260     data.frame(Species.deriv.emsc.pca, row.names = NULL))
1261
1262 # Create new short names
1263 Species <- c(rep("Agros", 51),
1264     rep("Antho", 51),
1265     rep("Desch", 51),
1266     rep("Festu", 50),
1267     rep("Molin", 50))
1268
1269 #Factorise
1270 Species <- as.factor(Species)
1271

```



```

1272 #Bind the two together
1273 Species.deriv.emsco <- cbind(Species, Species.deriv.emsco)
1274
1275 #Remove the sample labels
1276 Species.deriv.emsco <- Species.deriv.emsco[,-2]
1277
1278 names(Species.deriv.emsco)<-
1279 sapply(str_remove_all(colnames(Species.deriv.emsco),"X"),"[")
1280
1281 # Average of each species
1282 Species.derivo.means <- aggregate(Species.deriv.emsco[,2:1749],
1283                                 by = list(Species),
1284                                 FUN = mean)
1285 rownames(Species.derivo.means) <- Species.derivo.means[,1]
1286 Species.derivo.means <- Species.derivo.means[,-1]
1287
1288 # Write file
1289 Species.derivo.means <- data.frame(t(Species.derivo.means))
1290 write.csv(Species.derivo.means, "Origin.means.emsc.all.final.csv")
1291
1292
1293

```

```

1294 #####
1295 ## Plot non-differentiated spectra      ##
1296 ## following parts of Jardine (2021)    ##
1297 ## R script.                            ##
1298 ## Scoble, L (2023)                     ##
1299 #####
1300
1301 # Read in file
1302 Ad1 <- read.csv("all.data.emsc.baseline.final.csv", check.names = F, row.names = 1)
1303 Ad1 <- data.frame(t(Ad1))
1304 names(Ad1)<-sapply(str_remove_all(colnames(Ad1),"X"),"[")
1305 Ad2 <- Ad1
1306 Ad1 <- Ad1[1:253,]
1307 str(Ad1)
1308
1309 # Remove row names into column
1310 Ad1 <- cbind(rownames(Ad1), data.frame(Ad1, row.names = NULL))
1311
1312 # Create new short names
1313 Species <- c(rep("Agros", 51),
1314             rep("Antho", 51),
1315             rep("Desch", 51),
1316             rep("Festu", 50),
1317             rep("Molin", 50))
1318
1319 #Factorise
1320 Species <- as.factor(Species)
1321
1322 #Bind the two together
1323 Ad1 <- cbind(Species, Ad1)
1324
1325 #Remove the sample labels
1326 Ad1 <- Ad1[,-2]
1327

```

```

1328 ##### Mean and Standard Deviation #####
1329 grass.means <- aggregate(Ad2,
1330     by = list(Species),
1331     FUN = mean)
1332
1333 rownames(grass.means) <- grass.means[,1]
1334 grass.means <- grass.means[,-1]
1335
1336 grass.sd <- aggregate(Ad2,
1337     by = list(Species),
1338     FUN = sd)
1339 rownames(grass.sd) <- grass.sd[,1]
1340 grass.sd <- grass.sd[,-1]
1341
1342 ##### For Origin Plots #####
1343 grass.means <- data.frame(t(grass.means))
1344 write.csv(grass.means, "Grass.means.origin.csv")
1345
1346 ##### Plot Data #####
1347 grass.means <- data.frame(t(grass.means))
1348 # Full spectra (first plot only)
1349 par(mfrow = c(1,2), mar = c(3,2,1,0) + 0.01)
1350
1351 #fingerprint region (second plot only)
1352 par(mar = c(3, 1, 1, 3) + 0.01)
1353
1354 col <- brewer.pal(5, "Dark2")
1355
1356 # Select colours from RColorBrewer
1357 speciescol <- c("#1B9E77", "#D95F02", "#7570B3", "#E7298A", "#66A61E")
1358
1359 yvals <- seq(from = 4.5, to = 0.05, length.out = 5)
1360
1361 wavenumber <- (gsub("X", "", colnames(Ad1[,2:1763])))

```

```

1362 wavenumber <- as.numeric(wavenumber)
1363
1364 length(wavenumber)
1365 length(grass.means[1,])
1366 # Change xlim value for second plot (1800-600)
1367 plot(wavenumber, grass.means[1,], las = 1,
1368      type = "n", xlim = c(1800, 600), ylim = c(0, 6),
1369      xlab = "", ylab = "",
1370      yaxt = "n", xaxt = "n")
1371
1372 for(i in 5:1) {
1373   col.e <- col2rgb(speciescol[i])
1374   polygon(c(wavenumber, rev(wavenumber)),
1375          c(grass.means[i,]+yvals[i]+grass.sd[i,],
1376            rev(grass.means[i,]+yvals[i]-grass.sd[i,])),
1377          col = rgb(col.e[1], col.e[2], col.e[3], alpha = 80, maxColorValue = 255),
1378          border = NA)
1379
1380 for(i in 5:1) {
1381   lines(wavenumber, grass.means[i,]+yvals[i],
1382        col = speciescol[i])
1383
1384 }}
1385
1386 axis(1, lwd = 0, lwd.ticks = 2, tcl = 0.3,
1387      mgp = c(1.5, 0.2, 0),
1388      las = 1)
1389 mtext(expression("Wavenumber cm"-1), side = 1, line = 1)
1390
1391 # For first plot only
1392 mtext("Relative Intensity", side = 2, line = 0.6, las = 0)
1393 # For second plot only
1394 legend.names <- cbind.data.frame(row.names(grass.means), grass.means)
1395 #make and populate a new column with a short species name

```

```
1396 legend.names$spec <- substr(legend.names$`row.names(grass.means)` , 1,6)
1397
1398 legend.names$spec <- as.character(legend.names$spec)
1399 leg.txt <- unique(legend.names$spec)
1400
1401 legend("right", inset = c(-0.25, 0), leg.txt, pch = 19, cex = 0.65,
1402       col = col, xpd = TRUE, bty = "n")
1403
```

```

1404 #####
1405 ## Fyfe, R and Scoble, L (2023) ##
1406 ## PCA Analysis ##
1407 #####
1408
1409 # Non differentiated spectra
1410 dfbaseemsc <- read.csv("all.data.emsc.baseline.final.csv", check.names = F, row.names = 1)
1411 dfbaseemsc <- data.frame(t(dfbaseemsc))
1412 names(dfbaseemsc)<-sapply(str_remove_all(colnames(dfbaseemsc),"X"),"[")
1413
1414 # Truncate
1415 dfbaseemsc1 <- dfbaseemsc[1141:ncol(dfbaseemsc)]
1416
1417 # Remove anomalies
1418 dfbaseemsc2 <- dfbaseemsc1[-c(17, 18, 19, 90, 91, 92, 93, 95, 96, 97, 98, 99),]
1419
1420 #Rename to make easier for plot
1421 df.trunc <- dfbaseemsc2
1422
1423
1424 ##### Second Derivative Spectra (all data) #####
1425
1426 df.trunc1 <- read.csv("Species.deriv.emsc.final.csv", check.names = F, row.names = 1)
1427
1428 # Truncate (134 instead of 1141 so data starts at same wavenumber)
1429 df.trunc2 <- df.trunc1[1134:ncol(df.trunc1)]
1430 df.trunc2 <- df.trunc2[-c(17, 18, 19, 90, 91, 92, 93, 95, 96, 97, 98, 99),]
1431 names(df.trunc2)<-sapply(str_remove_all(colnames(df.trunc2),"X"),"[")
1432
1433 ##### Plot PCA, replace df.trunc with df.trunc2 for second plot #####
1434 dfs moo.pca <- prcomp(df.trunc)
1435 dfs moo.pca.scores <- as.data.frame(dfs moo.pca$x)
1436 dfs moo.pca.scores <- cbind.data.frame(row.names(df.trunc), dfs moo.pca.scores[,1:5])
1437 summary(dfs moo.pca)

```

```

1438 #make and populate a new column with a short species name
1439 dfs moo.pca.scores$spec <- substr(dfs moo.pca.scores$`row.names(df.trunc)` , 1,6)
1440
1441 #make a colour code for each species using short species names
1442 groups <- cbind.data.frame(unique(dfs moo.pca.scores$spec),
1443                             seq(1, length(unique(dfs moo.pca.scores$spec)), by = 1))
1444 colnames(groups) <- c("spec", "group")
1445 #join the colour codes to the PCA result file
1446 dfs moo.pca.scores <- merge(dfs moo.pca.scores, groups, by = "spec")
1447
1448 col <- brewer.pal(5, "Dark2")
1449
1450 dfs moo.pca.scores$group <- as.factor(dfs moo.pca.scores$group)
1451
1452 par(xpd = FALSE, mfrow = c(1,1), mar = c(5, 5, 5, 7), cex = 0.5, adj = 0.5, tck = 0.01)
1453
1454
1455 plot(dfs moo.pca.scores$PC1, dfs moo.pca.scores$PC2, group = dfs moo.pca.scores$groups,
1456      col = c("#1B9E77", "#D95F02", "#7570B3", "#E7298A",
1457             "#66A61E")[as.factor(dfs moo.pca.scores$group)],
1458      pch = 19, cex = 1.5, asp = 1, cex.axis = 1.5, xlab = "PC1 (74%)", ylab = "PC2 (18%)",
1459      cex.lab = 1.5)
1460 abline(h = 0, col = "grey")
1461 abline(v = 0, col = "grey")
1462
1463
1464 # Second derivative plot only
1465 dfs moo.pca.scores$spec <- as.character(dfs moo.pca.scores$spec)
1466 leg.txt <- unique(dfs moo.pca.scores$spec)
1467
1468 legend("right", inset = c(-0.15, 0), leg.txt, pch = 19, cex = 1.5,
1469      col = col, xpd = TRUE, bty = "n")
1470
1471

```

```

1472 ##### Loading Plots - Run for each PCA plot #####
1473 loadings <- as.data.frame(dfsmoo.pca$rotation)[1:2]
1474
1475 scale <- min(max(abs(dfsmoo.pca.scores$PC1))/max(abs(loadings$PC1)),
1476             max(abs(dfsmoo.pca.scores$PC2))/max(abs(loadings$PC2))) * 0.8
1477
1478
1479 #extract the wavenumbers as numbers from rotation
1480 wavenumbers <- as.numeric(rownames(dfsmoo.pca$rotation))
1481
1482 #extracts the first column (PCA1). Change [,1] to [,2] for PCA2 etc.
1483 PC1loading <- as.data.frame(loadings[,1])
1484 PC2loading <- as.data.frame(loadings[,2])
1485
1486 #writes the wavenumbers to the PCA1loadings object
1487 PC1loading$wavenumber <- wavenumbers
1488 PC2loading$wavenumber <- wavenumbers
1489
1490 colnames(PC1loading) <- c("loading", "wavenumber")
1491 colnames(PC2loading) <- c("loading", "wavenumber")
1492
1493
1494 #switch PC1loadings to PC2 for other plot
1495 plot(PC1loading$loading ~ PC1loading$wavenumber, type = "l",
1496      xlim = c(1800,600), xlab = "Wavenumber", ylab = "PC1 Loadings", cex.axis = 1.5,
1497      cex.lab = 1.5)
1498 #2nd deriv line
1499 abline(h = 0, col = "black")
1500
1501 # Write files for loadings
1502
1503 write.csv(PC1loading, "PC1Loading.csv")
1504 write.csv(PC2loading, "PC2Loading.csv")
1505

```



```

1506 #####
1507 ## Fyfe, R and Scoble, L (2023)      ##
1508 ## HCA Plot - Repeat For Each Set   ##
1509 ## of Data (df.trunc/df.trunc2)     ##
1510 #####
1511
1512
1513 diss <- dist(df.trunc2, method = "euclidean")
1514
1515 # Cluster analysis
1516 cluster <- as.dendrogram(hclust(diss))
1517
1518 # Set plotting margins and font size for the general plots
1519 par(cex=0.5, mar=c(5, 8, 4, 1))
1520
1521 # c=Choose number of clusters, 5 separates the main species
1522
1523 k = 5
1524
1525 # Set up plotting colours
1526 cluster <- cluster %>%
1527   color_branches(k = k) %>%
1528   color_labels(k = k)
1529
1530 # Plot circular dendrogram
1531 circlize_dendrogram(cluster)
1532
1533 # Export the cluster numbers assigned to samples
1534 cuts <- cbind.data.frame(rownames(df.trunc), cutree(cluster, k = k))
1535 colnames(cuts) <- c("sample", "cluster_number")
1536 write.csv(cuts, "cluster.groups.by.sample.diff.final.csv", row.names = F)
1537
1538
1539

```

```
1540 # ITOL file
1541 my_tree <- as.phylo(cluster)
1542
1543 write.tree(phy = my_tree, file = "Treefinal.diff.newick")
1544
1545
1546
```

```

1547 #####
1548 ## Scoble, L (2023) ##
1549 ## Decision Trees and randomForest ##
1550 #####
1551
1552 ##### PART 1 - Decision trees: Extracting rpart rules which show which wavenumbers
1553 # are causing discrepancies between species - then compare to PCA loading plots #####
1554
1555 # Read in file
1556 d <- read.table("all.data.emsc.baseline.final.csv", sep = ",", header = T, row.names = 1)
1557 d <- data.frame(t(d))
1558 names(d)<-sapply(str_remove_all(colnames(d),"X"),"[")
1559
1560 # Truncate spectra
1561 d <- d[,1141:ncol(d)]
1562
1563 d <- d[1:253,]
1564 str(d)
1565 d <- cbind(rownames(d), data.frame(d, row.names = NULL))
1566
1567
1568 # Make column with short specie names
1569 Species <- c(rep("Agros", 51),
1570             rep("Antho", 51),
1571             rep("Desch", 51),
1572             rep("Festu", 50),
1573             rep("Molin", 50))
1574
1575 # Factorise
1576 Species <- as.factor(Species)
1577
1578 # Bind the two together
1579 d <- cbind(Species, d)
1580

```

```

1581 # Remove the sample labels
1582 d <- d[,-2]
1583
1584
1585 summary(d$Species)
1586 set.seed(2)
1587
1588 ##### First decision (classification) tree using all data #####
1589 fit <- rpart(Species ~., data = d, method = "class")
1590 par(mar = c(2, 4, 4, 4))
1591 par(mfrow = c(1,1))
1592
1593 # Plot classification tree
1594 plot(fit)
1595 text(fit, cex = 0.9, xpd = TRUE)
1596
1597 # Use rplot for more better visuals (legend position may need to be changed)
1598 rplot <- rpart.plot(fit, type = 4, extra = "auto", clip.right.labs = FALSE,
1599                   legend.x = 0.85, legend.y = 1, legend.cex = 1.3,
1600                   cex = 0.8)
1601
1602
1603 # Extract the rules that the algorithm uses to build tree and splits
1604 # This is to look at what wavenumbers are driving the discrepancy between
1605 # species
1606
1607 # Digits = 3 to get an extra decimal place (easier to refer to the data)
1608 rpart.rules(fit)
1609 rules <- rpart.rules(fit, digit = 3)
1610
1611 # Remove columns that aren't relevant
1612 rules <- rules[,-2]
1613 rules <- rules[,-2]
1614 rules <- rules[,-5]

```

```

1615 rules <- rules[,-8]
1616 rules <- rules[,-11]
1617
1618 # Change colnames (Less than, Equal to, Greater than (L/E/G), Absorbance units (Au))
1619 colnames(rules) <- c("Species", "Wavenumber1", "L/E/G", "Au", "Wavenumber2",
1620                    "L/E/G", "Au", "wavenumber3", "L/E/G", "Au",
1621                    "wavenumber4", "L/E/G", "Au")
1622
1623 # Write csv
1624 write.csv(rules, "rpart.wavenumber.rules.final.csv")
1625
1626 # Find wavenumbers in original dataset to cross check rules
1627 WN1 <- d %>% dplyr::select(X1693.4306)
1628 WN2 <- d %>% dplyr::select(X883.36129)
1629 WN3 <- d %>% dplyr::select(X1745.50649)
1630 WN4 <- d %>% dplyr::select(X1151.45566)
1631
1632 cross_check <- cbind(Species, WN1, WN2, WN3, WN4)
1633
1634 colnames(cross_check) <- gsub("X","",colnames(cross_check[,1:5]))
1635
1636 write.csv(cross_check, "Cross_check_wavenumbers.final.csv")
1637
1638 ##### Looped Variance #####
1639
1640 ##### Split the data and run decision tree 100 times in a loop #####
1641 # Will the same four wavenumbers still be prominent or will splitting the data
1642 # create more variance.
1643 set.seed(2)
1644 tree_lengths <- data.frame()
1645
1646 for(i in 1:100) {
1647   train <- sample(nrow(d), 0.8*nrow(d))
1648   training_data <- d[train,]

```

```

1649   dim(training_data)
1650   summary(training_data$Species)
1651
1652   testing_data <- d[-train, ]
1653   dim(testing_data)
1654   summary(testing_data$Species)
1655
1656   tree_i <- rpart(Species ~ ., data = training_data, method = "class")
1657   wavesum <- tree_i$frame$var
1658   tree_lengths <- rbind(tree_lengths, wavesum)
1659   names(tree_lengths) <- NULL
1660 }
1661
1662 par(mfrow = c(1,1))
1663 par(mar = c(2, 4, 4, 2))
1664 rpart.plot(tree_i, type = 4, extra = 104, clip.right.labs = FALSE, digits = 2,
1665           round = 0, legend.x = 0.85, legend.y = 1, legend.cex = 1,
1666           cex = 0.7)
1667
1668
1669 # Pull one tree from loop to look at rules
1670 # digits = 3 to get an extra decimal place (easier to refer to the data)
1671 rpart.rules(tree_i)
1672 rules_one <- rpart.rules(tree_i, digit = 3)
1673
1674 # Remove columns that aren't relevant
1675 rules_one <- rules_one[,-2]
1676 rules_one <- rules_one[,-2]
1677 rules_one <- rules_one[,-5]
1678 rules_one <- rules_one[,-8]
1679 rules_one <- rules_one[,-11]
1680
1681 #change colnames (Less than, Equal to, Greater than (L/E/G), Absorbance units (Au))
1682 colnames(rules_one) <- c("Species", "Wavenumber1", "L/E/G", "Au", "Wavenumber2",

```

```

1683         "L/E/G", "Au", "wavenumber3", "L/E/G", "Au",
1684         "wavenumber4", "L/E/G", "Au" )
1685 # What does the new set of rules for split data show compared to the previous?
1686 write.csv(rules_one, "rpart_wavenumbers_rules_loop.final.csv")
1687
1688 WN5 <- d %>% dplyr::select(X1693.4306)
1689 WN6 <- d %>% dplyr::select(X883.36129)
1690 WN7 <- d %>% dplyr::select(X1745.50649)
1691 WN8 <- d %>% dplyr::select(X1155.31313)
1692
1693 cross_check1 <- cbind(Species, WN5, WN6, WN7, WN8)
1694
1695
1696 colnames(cross_check1) <- gsub("X","",colnames(cross_check1[,1:5]))
1697
1698 write.csv(cross_check, "Cross_check_wavenumbers_loop.csv")
1699
1700 # Clean up the tree_lengths data frame to only have wave numbers present
1701
1702 tree_lengths <- tree_lengths[,-9]
1703 tree_lengths <- tree_lengths[,-8]
1704
1705 tree_lengths <- as.data.frame(apply(tree_lengths, 2, function(x) {
1706   x <- gsub("X", "", x)
1707   }))
1708 tree_lengths <- as.data.frame(apply(tree_lengths, 2, function(x) {
1709   x <- gsub("<leaf>", "0", x)
1710   }))
1711
1712 # Convert to num
1713 tree_lengths <- type.convert(tree_lengths, as.is = TRUE)
1714
1715
1716 tree_lengths2 <- melt(tree_lengths, id.vars = c("V1", "V2", "V3", "V4", "V5", "V6", "V7"))

```

```

1717
1718
1719 # Place all wavenumbers into one column
1720 tree_lengths2 <- reshape(tree_lengths, direction = "long", sep = "", varying = 1:7)
1721
1722 # Remove time column
1723 tree_lengths2 <- tree_lengths2[,-1]
1724 table <- table(tree_lengths2$V)
1725 table <- as.data.frame(table)
1726
1727 # Remove zero (first row) as not relevant
1728 table <- table[-1,]
1729
1730 # Arrange table so Freq is descending from largest to smallest
1731 table2 <- table %>%
1732   arrange(desc(Freq))
1733
1734 # What is table showing and how does that compare to fit and also the PCA loadings
1735
1736 # Plot histogram
1737 table2 <- table2[1:10,]
1738 table3 <- as.data.frame(table2)
1739
1740 par(mfrow = c(1,1))
1741 par(mar = c(2, 4, 4, 4))
1742 ggplot(table3, aes(x = reorder(Var1, -Freq), y = Freq, fill = rules)) +
1743   geom_histogram(stat = "Identity", colour = "darkblue", fill = "lightblue") +
1744   labs(x = "Wavenumber", y = "Frequency of Appearance") +
1745   theme(panel.grid = element_blank(), strip.text.y = element_blank(),
1746         axis.text.x = element_text(angle = 50, vjust = 1, hjust = 1, size = 11, face = "bold",
1747         colour = "black"), axis.title.x = element_text(size = 15), axis.title.y = element_text(size
1748 = 14),
1749         axis.text.y = element_text(size = 11, face = "bold", colour = "black"),
1750         panel.background = element_blank())

```



```

1751
1752 # Write csv for table
1753 write.csv(table3, "Final.table.loop.freq.csv")
1754
1755 ##### Repeat for first wavenumber split #####
1756 set.seed(2)
1757 tree_lengths <- data.frame()
1758
1759 for(i in 1:100) {
1760   train <- sample(nrow(d), 0.8*nrow(d))
1761   training_data <- d[train,]
1762   dim(training_data)
1763   summary(training_data$Species)
1764
1765   testing_data <- d[-train, ]
1766   dim(testing_data)
1767   summary(testing_data$Species)
1768
1769   tree_i <- rpart(Species ~ ., data = training_data, method = "class")
1770   wavesum <- tree_i$frame$var[1]
1771   tree_lengths <- rbind(tree_lengths, wavesum)
1772   names(tree_lengths) <- NULL
1773 }
1774
1775 tree_lengths <- as.data.frame(apply(tree_lengths, 1, function(x) {
1776   x <- gsub("X", "", x)
1777 })))
1778 colnames(tree_lengths) <- "wavenumber"
1779
1780 tree_lengths2 <- tree_lengths %>% group_by(tree_lengths$wavenumber) %>%
1781   count(sort = TRUE)
1782 tree_lengths2 <- tree_lengths2[1:6,]
1783
1784

```

```

1785
1786 ggplot(tree_lengths2, aes(x = reorder(`tree_lengths$wavenumber`, -n), y = n, fill = rules)) +
1787   geom_histogram(stat = "Identity", colour = "darkblue", fill = "lightblue") +
1788   labs(x = "Wavenumber", y = "Frequency of Appearance") +
1789   theme(panel.grid = element_blank(), strip.text.y = element_blank(),
1790         axis.text.x = element_text(angle = 50, vjust = 1, hjust = 1, size = 11, face = "bold",
1791                                   colour = "black"), axis.title.x = element_text(size = 15), axis.title.y =
1792   element_text(size = 14),
1793   axis.text.y = element_text(size = 11, face = "bold", colour = "black"),
1794   panel.background = element_blank())
1795
1796
1797 # Write csv for table
1798 write.csv(tree_lengths2, "first.wavenumber.rule.split.csv")
1799
1800 ##### Part 2 - RandomForest #####
1801 ## Classification using RandomForest          ##
1802 ## Build model using randomForest and training data ##
1803 #####
1804
1805 # Bagged trees
1806 set.seed(2)
1807 train <- sample(nrow(d), 0.8*nrow(d))
1808 training_data <- d[train,]
1809 dim(training_data)
1810 summary(training_data$Species)
1811
1812 testing_data <- d[-train, ]
1813 dim(testing_data)
1814 summary(testing_data$Species)
1815
1816 set.seed(2)
1817 bag.RF <- randomForest(Species ~ ., data = training_data, mtry = 622, ntree = 100,
1818                       importance = TRUE, proximity = TRUE, do.trace = TRUE)

```

```

1819
1820 bag.RF
1821 #Look at error matrix
1822 plot(bag.RF)
1823 print(bag.RF)
1824
1825 #Predict to see if trained forest will accurately predict test data
1826 bag.tree <- predict(bag.RF, testing_data, type = "class")
1827 tab <- table(bag.tree, testing_data$Species)
1828 tab
1829
1830 write.csv(tab, "prediction.RF.final.csv")
1831 (tab[1,5] + tab[5,1] / sum(tab))
1832
1833 #Plot the Variable importance
1834 par(mfrow = c(1,1), mar = c(2,2,1,2))
1835 varImpPlot(bag.RF,
1836           n.var = 24,
1837           type = 1,
1838           sort = TRUE,
1839           main = "Variable Importance Plot")
1840
1841 ##### Looped randomForest for MDA investigations #####
1842
1843 set.seed(2)
1844 # Make an empty list of 10
1845 ls <- list()
1846 n = 10
1847 datalist = list()
1848 # Pre-allocate for slightly more efficiency
1849 datalist = vector("list", length = n)
1850
1851
1852

```

```

1853 # Run loop
1854 for(i in 1:10) {
1855
1856   importance.tree <- randomForest(Species ~ ., d, ntree = 150, mtry = 24, importance =
1857   TRUE)
1858   plot(importance.tree)
1859   wavesum <- importance.tree$importance[,6, drop = FALSE]
1860   datalist[[i]] <- cbind(rownames(wavesum), data.frame(wavesum, row.names = NULL))
1861   colnames(datalist[[i]]) <- c("Wavenumber", "MeanDecreaseAccuracy")
1862
1863
1864   for (i in 1:length(datalist)) {
1865     assign(paste0("datalist", i), as.data.frame(datalist[[i]]))}
1866 }
1867
1868 #repeat for each datalist
1869 datalist1 <- datalist1 %>%
1870   arrange(desc(MeanDecreaseAccuracy))
1871
1872
1873 # Cbind all dataframes together
1874 dataframeall <- cbind.data.frame(datalist1, datalist2, datalist3, datalist4,
1875   datalist5, datalist6, datalist7, datalist8,
1876   datalist9, datalist10)
1877 # Convert to numeric
1878 dataframeall <- type.convert(dataframeall, as.is = TRUE)
1879
1880 # Trim rows to only have the top 24 (24 is the square root of 622)
1881 dataframeall <- dataframeall[1:24,]
1882
1883 # Rename column names
1884 colnames(dataframeall) <- c("V1", "V2", "V3", "V4", "V5", "V6", "V7", "V8", "V9", "V10",
1885   "V11", "V12", "V13", "V14", "V15", "V16", "V17", "V18", "V19", "V20")
1886

```

```

1887 # Split into Wavenumber and MDA
1888 dataframewavenumber <- data.frame(dataframeall[, c(1, 3, 5, 7, 9, 11, 13, 15, 17, 19)])
1889 dataframeMDA <- data.frame(dataframeall[, c(2, 4, 6, 8, 10, 12, 14, 16, 18, 20)])
1890
1891 # Rename column names
1892 colnames(dataframewavenumber) <- c("V1", "V2", "V3", "V4", "V5", "V6", "V7", "V8",
1893 "V9", "V10")
1894
1895 # Place all wavenumbers into one column
1896 dataframewavenumber1 <- melt(dataframewavenumber, id.vars = c("V1", "V2", "V3", "V4",
1897 "V5",
1898 "V6", "V7", "V8", "V9", "V10"))
1899
1900 dataframewavenumber1 <- reshape(dataframewavenumber1 , direction = "long",
1901 sep = "", varying = 1:10)
1902
1903 # Rename column names
1904 colnames(dataframeMDA) <- c("V1", "V2", "V3", "V4", "V5", "V6", "V7", "V8",
1905 "V9", "V10")
1906
1907 # Place all MDA into one column
1908 dataframeMDA1 <- melt(dataframeMDA, id.vars = c("V1", "V2", "V3", "V4",
1909 "V5", "V6", "V7", "V8", "V9", "V10"))
1910
1911 dataframewaveMDA1 <- reshape(dataframeMDA1 , direction = "long",
1912 sep = "", varying = 1:10)
1913
1914 # Combine the Wavenumber and MDA column from each dataframe
1915 combinedWNMDA <- cbind.data.frame(dataframewavenumber1$V,
1916 dataframewaveMDA1$V)
1917
1918 # Rename
1919 colnames(combinedWNMDA) <- c("Wavenumber", "MDA")
1920

```

```

1921 # Remove "X" character
1922 combinedWNMDA <- as.data.frame(apply(combinedWNMDA, 2, function(x) {
1923   x <- gsub("X", "", x) })))
1924
1925 # Convert to numeric
1926 combinedWNMDA <- type.convert(combinedWNMDA, as.is = TRUE)
1927
1928
1929 # Arrange in descending order of MDA numbers
1930 combinedWNMDA <- combinedWNMDA %>%
1931   arrange(desc(MDA))
1932
1933 # Select only top 24 of all dataframes combined
1934 table <- combinedWNMDA[1:24,]
1935 table2 <- as.data.frame(table)
1936 table2 <- table2 %>% arrange(MDA)
1937
1938
1939 par(mar = c(3, 1, 0, 1))
1940
1941 # Plot dotcharts
1942 dotchart(table2$MDA, table2$Wavenumber, xlim = range(table2$MDA),
1943   xlab = "MeanDecreaseAccuracy", mgp=c(2,1,.5), las=1, cex = 0.9)
1944
1945
1946 ##### Run Rf using isolated variables #####
1947 table2$Wavenumber
1948 set.seed(2)
1949 isolated <- d %>% dplyr::select(Species, X1691.50187, X1676.07197,
1950   X1641.35472, X1467.76844, X1461.98223, X1450.40981,
1951   X1134.09703, X1072.37747, X1068.51999, X866.00267, X858.28772,
1952   X821.64173, X819.71299, X815.85552, X800.42563, X794.63942,
1953   X786.92447, X727.13364, X688.55891, X632.62556, X626.83935,
1954   X622.98187, X613.33819, X607.55198)

```

```

1955
1956 train <- sample(nrow(isolated), 0.8*nrow(isolated))
1957 training_data1 <- isolated[train,]
1958 dim(training_data1)
1959 summary(training_data1$Species)
1960
1961 testing_data1 <- isolated[-train, ]
1962 dim(testing_data1)
1963 summary(testing_data1$Species)
1964
1965
1966 set.seed(2)
1967 isolated.rf <- randomForest(Species ~ ., data = training_data1, ntree = 100,
1968                             importance = TRUE, proximity = TRUE, do.trace = TRUE)
1969
1970 isolated.rf
1971 #Look at error matrix
1972 plot(isolated.rf)
1973 print(isolated.rf)
1974
1975 #Predict test data
1976 bag.tree <- predict(isolated.rf, testing_data, type = "class")
1977 tab <- table(bag.tree, testing_data$Species)
1978 tab
1979
1980 write.csv(tab, "prediction.RF.isolated.csv")
1981 (tab[1,5] + tab[5,1] / sum(tab))
1982
1983
1984
1985
1986
1987
1988

```

```

1989 #Plot the Variable importance
1990 par(mar = c(3, 1 , 1 ,1))
1991 varImpPlot(isolated.rf,
1992     type = 1,
1993     sort = TRUE,
1994     main = "Variable Importance Plot",
1995     cex = 0.75)
1996
1997
1998 # Run with all isolated data
1999 set.seed(2)
2000 isolated.rf1 <- randomForest(Species ~ ., data = isolated, ntree = 100,
2001     importance = TRUE, proximity = TRUE, do.trace = TRUE)
2002
2003 isolated.rf1
2004 #Look at error matrix
2005 plot(isolated.rf1)
2006 print(isolated.rf1)
2007
2008 #Plot the Variable importance
2009 par(mar = c(4, 1 , 1 ,1))
2010 varImpPlot(isolated.rf1,
2011     type = 1,
2012     sort = TRUE,
2013     main = "Variable Importance Plot")
2014
2015 par(mar = c(3, 3, 1, 3))
2016
2017
2018
2019
2020
2021
2022

```



```

2023 # Plot final dotchart
2024 # All data
2025 varImpPlot(isolated.rf1,
2026     type = 1,
2027     sort = TRUE,
2028     main = "Variable Importance Plot",
2029     cex = 0.75,
2030     mgp=c(2,1,.5))
2031
2032 #See if trained data can predict unlabelled test data
2033 #make copy of testing_data
2034 testing_data1 <- testing_data
2035
2036 # Actual Species names
2037 Species_1 <- testing_data1[1]
2038
2039 #Remove the sample labels
2040 testing_data1 <- testing_data1[,-1]
2041
2042 #Unlabel data
2043 new_data <- data.frame(testing_data1[,-1])
2044
2045 #Predict for accuracy
2046 new_data$predictedlabel <- predict(isolated.rf, new_data)
2047 new_data$predictedlabel
2048 Predicted <- as.character(new_data$predictedlabel)
2049 Actual <- as.character(testing_data1$Species)
2050
2051 #Cbind the predicted labels with the known species labels from test data
2052 new_data1 <- as.data.frame(cbind(Predicted, Species_1))
2053
2054 #View results as csv
2055 write.csv(new_data1, "prediciton_name_data_isolated_final.csv")
2056

```

```

2057 #####
2058 ## Plot MDA wavenumbers with relative intensity  ##
2059 ## as a boxplot                                     ##
2060 ## Scoble, L (2023)                               ##
2061 #####
2062
2063
2064 impdf <- data.frame(importance(isolated.rf1))
2065
2066 #Remove X from dataframe
2067 rownames(impdf) <- (gsub("X","",rownames(impdf[1:7])))
2068
2069 impdf <- cbind(rownames(impdf), data.frame(impdf, row.names = NULL))
2070
2071 # Convert to numeric
2072 impdf <- type.convert(impdf, as.is = TRUE)
2073 impdf <- impdf[,-8]
2074
2075 # Arrange data in desc of MDA
2076 impdf <- impdf %>%
2077   arrange(desc(MeanDecreaseAccuracy))
2078 impdf <- impdf[,-7]
2079
2080 #Prepare dataframe for boxplot
2081 impdf <- data.frame(t(impdf))
2082 colnames(impdf) <- impdf[1,]
2083 impdf <- impdf[-1,]
2084 impdf <- cbind(rownames(impdf), data.frame(impdf, row.names = NULL))
2085
2086
2087
2088
2089
2090

```

```

2091 # Create new Species labels
2092 Species <- c(rep("Agros", 1),
2093             rep("Antho", 1),
2094             rep("Desch", 1),
2095             rep("Festu", 1),
2096             rep("Molin", 1))
2097
2098 # Factorise
2099 Species <- as.factor(Species)
2100
2101 # Bind the two together
2102 impdf <- cbind(Species, impdf)
2103
2104 # Remove the sample labels
2105 impdf <- impdf[,-2]
2106
2107 colnames(impdf) <- (gsub("X", "", colnames(impdf)))
2108
2109 #Melt all data together
2110 melt <- melt(impdf)
2111
2112 #Plot boxplot of MDA as x axis
2113 p <- ggplot(melt, aes(factor(variable), value, fill = Species))
2114 p + geom_boxplot() + facet_wrap(~variable, scale="free") +
2115   theme(axis.text.x = element_blank())
2116
2117
2118
2119
2120
2121
2122
2123
2124

```

```

2125 #boxplot of wavenumbers in order of MDA dotchart
2126 colnames(impdf)
2127 varimporder <- d %>% dplyr::select(Species, X1641.35472, X786.92447, X622.98187,
2128           X1676.07197, X727.13364, X1450.40981,
2129           X1072.37747, X1691.50187, X1467.76844,
2130           X1134.09703, X794.63942, X688.55891, X1461.98223,
2131           X800.42563, X866.00267, X821.64173, X819.71299,
2132           X626.83935, X613.33819, X815.85552, X632.62556,
2133           X858.28772, X1068.51999, X607.55198)
2134
2135 colnames(varimporder) <- (gsub("X","",colnames(varimporder)))
2136
2137 #melt all the data together
2138 melt <- melt(varimporder)
2139 boxplot(melt, value ~ variable)
2140
2141 p <- ggplot(melt, aes(factor(variable), value, fill = Species))
2142 p + geom_boxplot() + facet_wrap(~variable, scale="free") +
2143   labs(x = "Wavenumber", y = "Relative Intensity") +
2144   theme(axis.text.x = element_blank())
2145

```

2146 **References**

2147 Jardine, E. P., 2021. *Data and code for "Sporopollenin chemistry and its durability in the*  
2148 *geological record: an integration of extant and fossil chemical data across the seed plants"*.

2149 [Online] Available at: <https://doi.org/10.6084/m9.figshare.11382102.v1>

2150 [Accessed 28 September 2023].

2151

2152

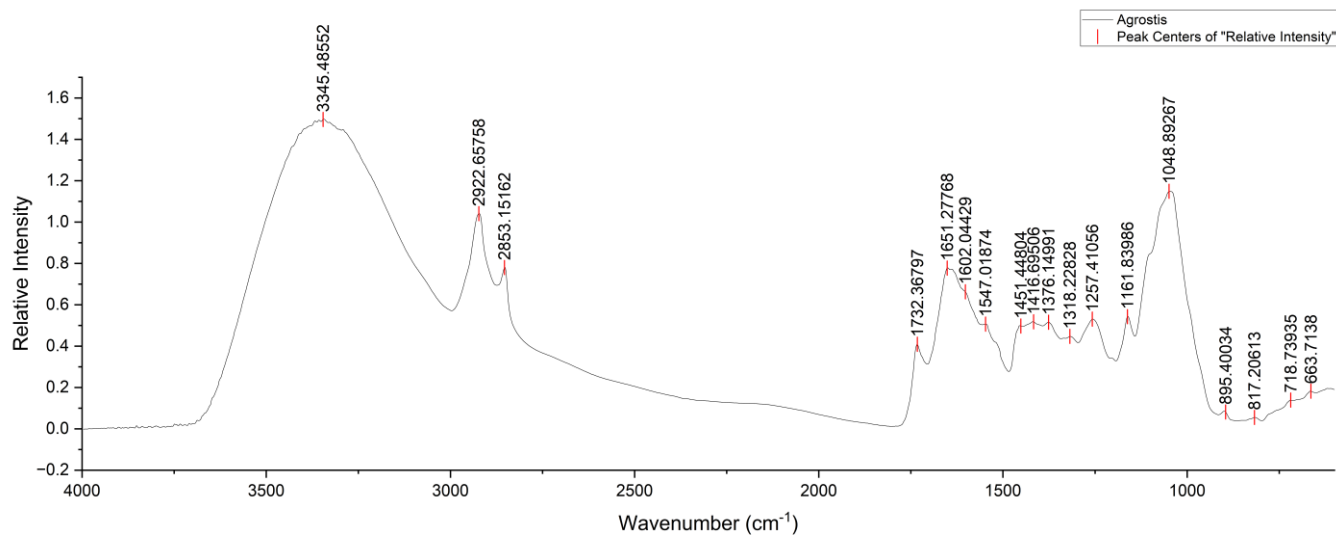
## Supplementary Material SM3

2153

### SM3.1 Non-Differentiated Grass Species Averaged Spectra With Peak Numbers

2155

#### SM3.1.1 Agrostis



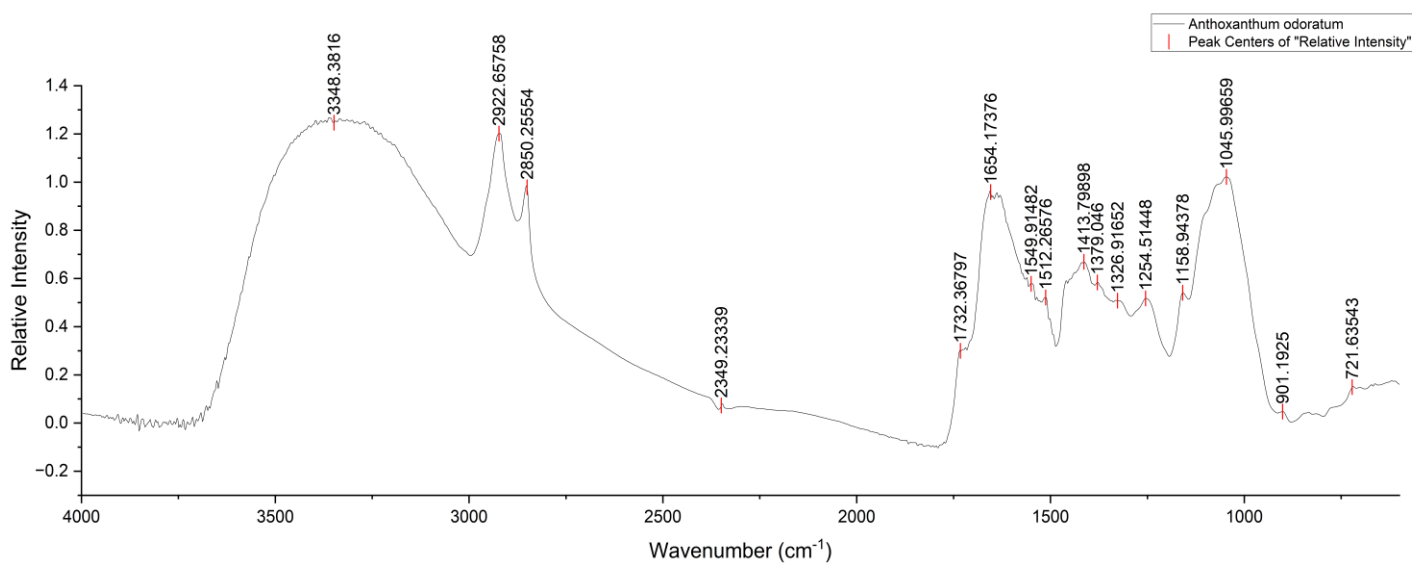
2157 *Figure SM3.1.1: Averaged FT-IR spectra of Agrostis with peak numbers included.*

2158

2159

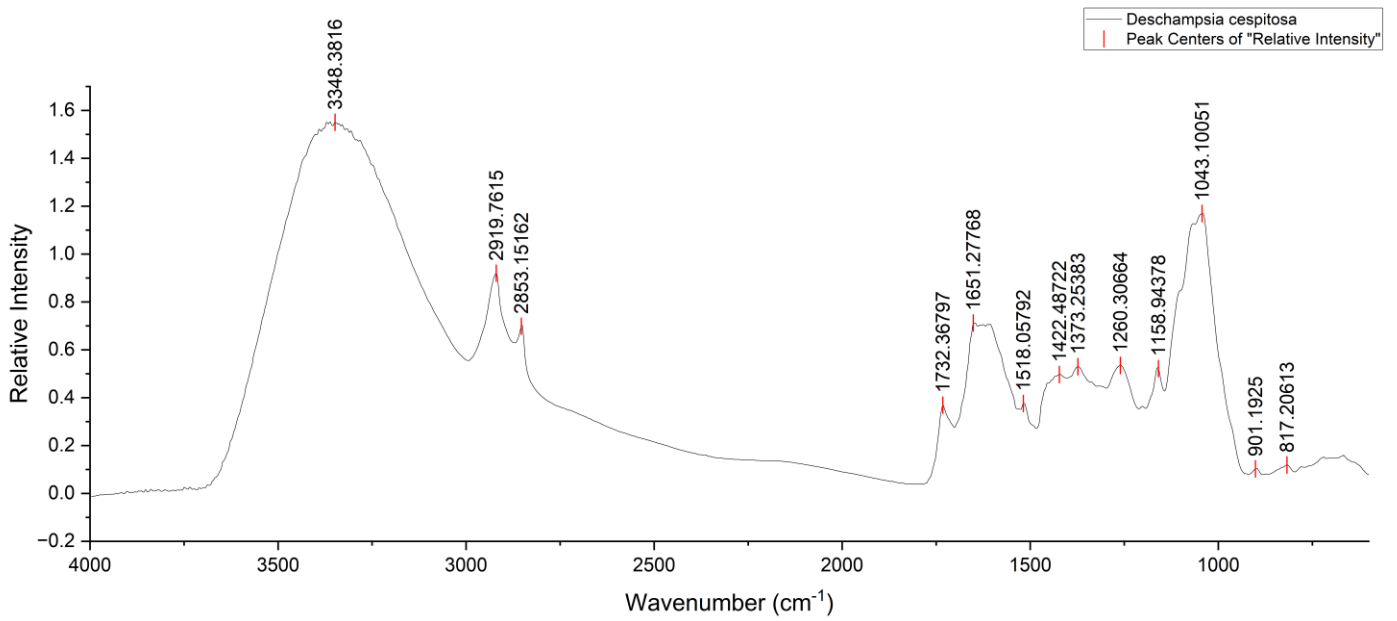
2160

#### 3.1.2 Anthoxanthum odoratum



2162 *Figure SM3.1.2: Averaged FT-IR spectra of Anthoxanthum odoratum with peak numbers included.*

2163 **3.1.3 Deschampsia cespitosa**

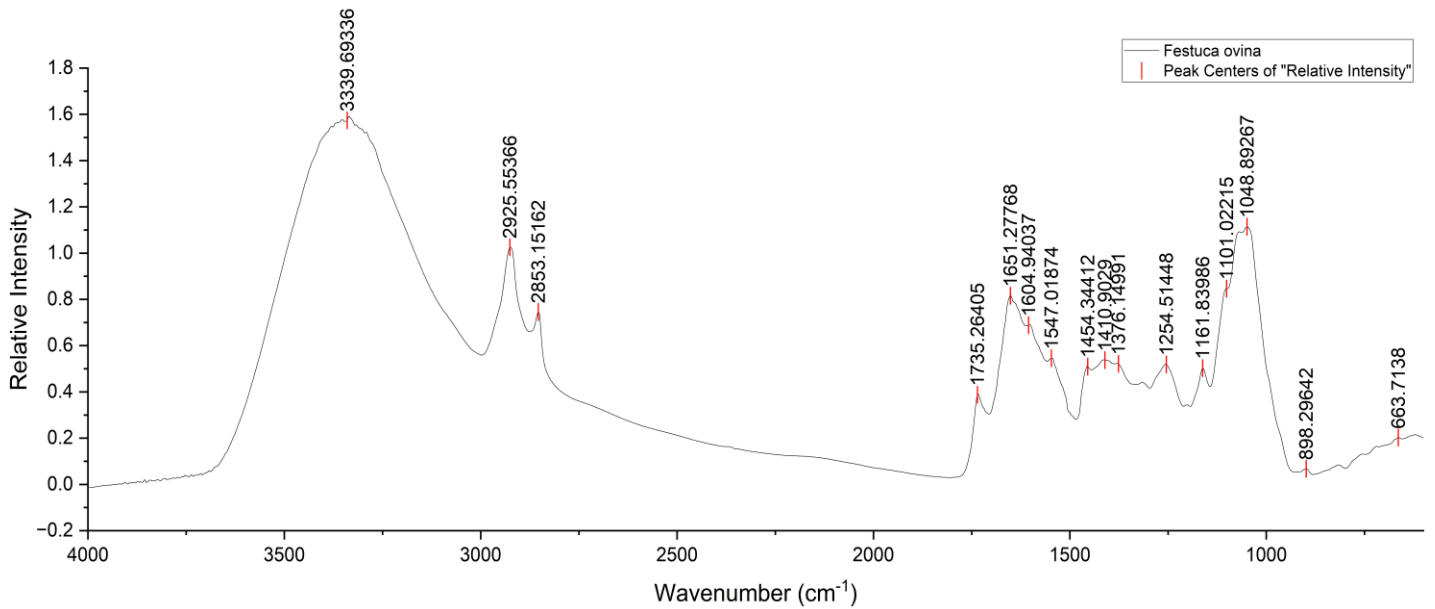


2164 *Figure SM3.1.3: Averaged FT-IR spectra of Deschampsia cespitosa with peak numbers included.*

2165

2166

2167 **3.1.4 Festuca ovina**

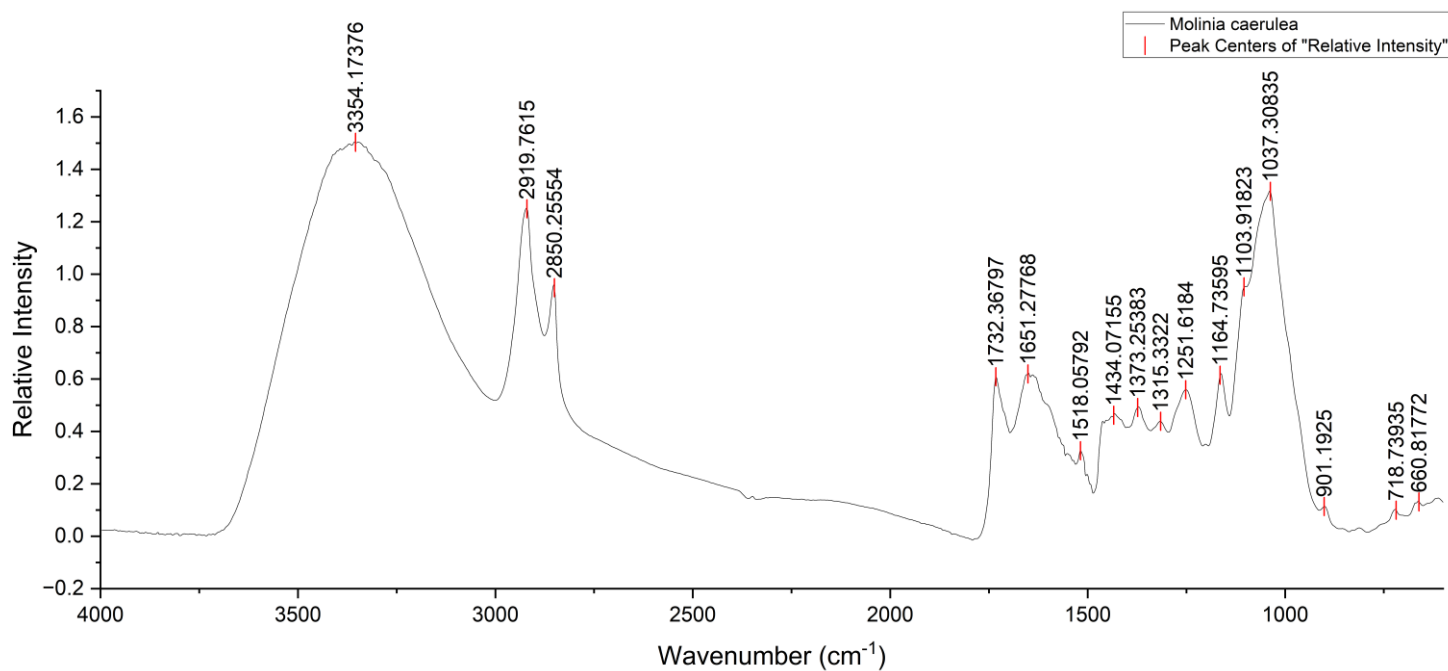


2168 *Figure SM3.1.4: Averaged FT-IR spectra of Festuca ovina with peak numbers included.*

2169

2170

2171 **3.1.5 Molinia caerulea**



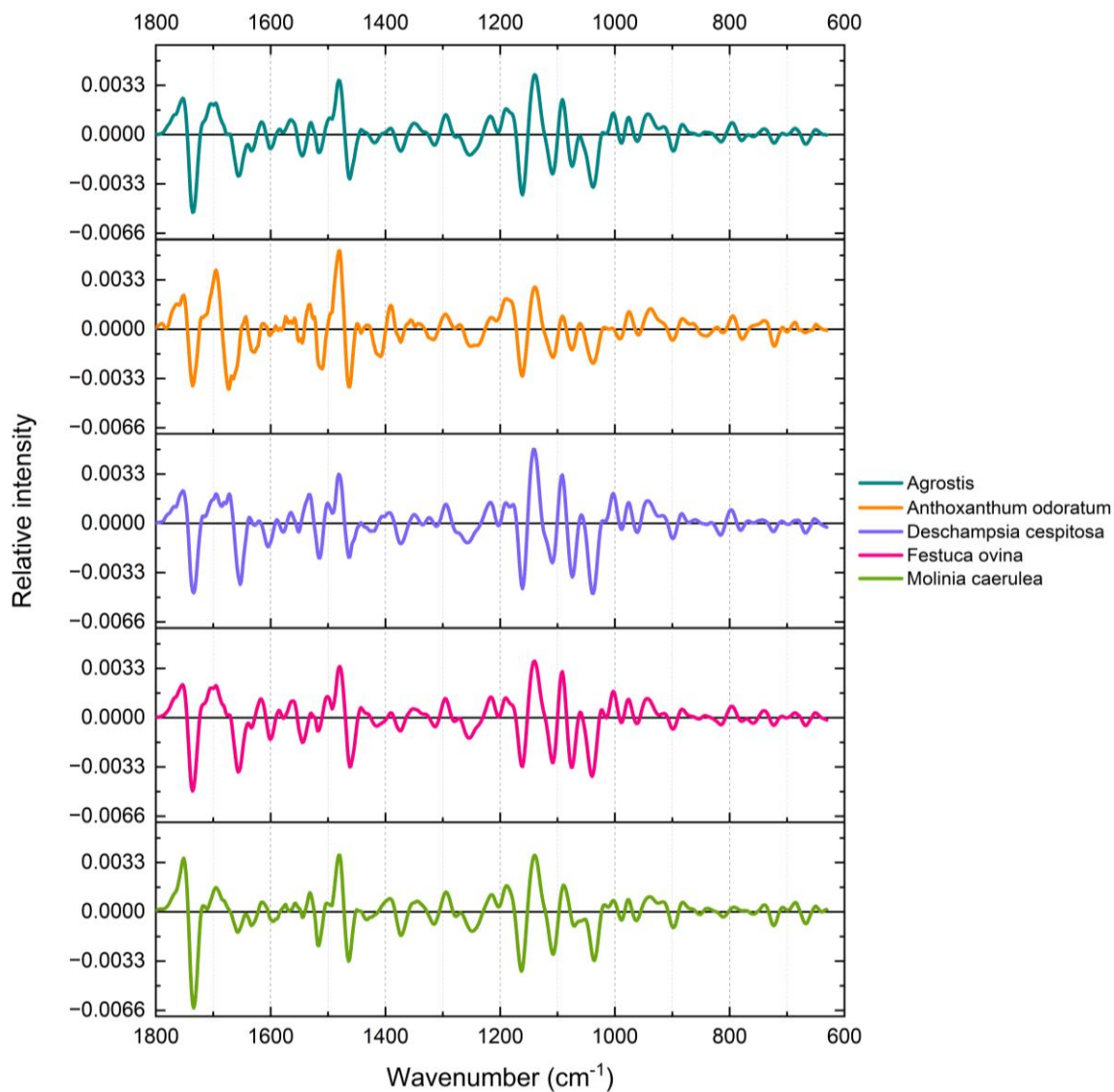
2172 *Figure SM3.1.5: Averaged FT-IR spectra of Molinia caerulea with peak numbers included.*

2173  
2174  
2175  
2176  
2177  
2178  
2179  
2180  
2181  
2182  
2183  
2184  
2185  
2186  
2187  
2188  
2189  
2190  
2191



2192 **SM3.2 Second derivatives of averaged grass spectra**

2193



2194 *Figure SM3.2: Averaged Savitzky-Golay smoothed, second derivative FT-IR spectra of the five*

2195 *moorland grass species (Agrostis is genus).*

2196

2197 **SM3.3: HCA of non-differentiated spectra**



2198

2199 *Figure SM3.3: Hierarchical cluster analysis (HCA) of the five moorland grass species using non-*  
2200 *differentiated FT-IR spectral data. Colours represent each cluster (five clusters).*

2201

2202



2204 *Figure SM3.4: Hierarchical cluster analysis (HCA) of the five moorland grass species using FT-IR*  
 2205 *Savitzky Golay smoothed, second derivative spectral data. Colours represent each cluster (five*  
 2206 *clusters).*

2207 **SM3.5: Confusion matrix of bagged randomForest model**

2208 *Table SM3.5: Confusion matrix of bagged randomForest model, rows are actual values, and columns*  
 2209 *are predicted.*

	Agros	Antho	Desch	Festu	Molin
Agros	8	0	0	0	0
Antho	0	8	0	0	0
Desch	0	0	16	0	0
Festu	0	0	0	12	0
Molin	0	0	0	0	7

2210

2211

2212

2213

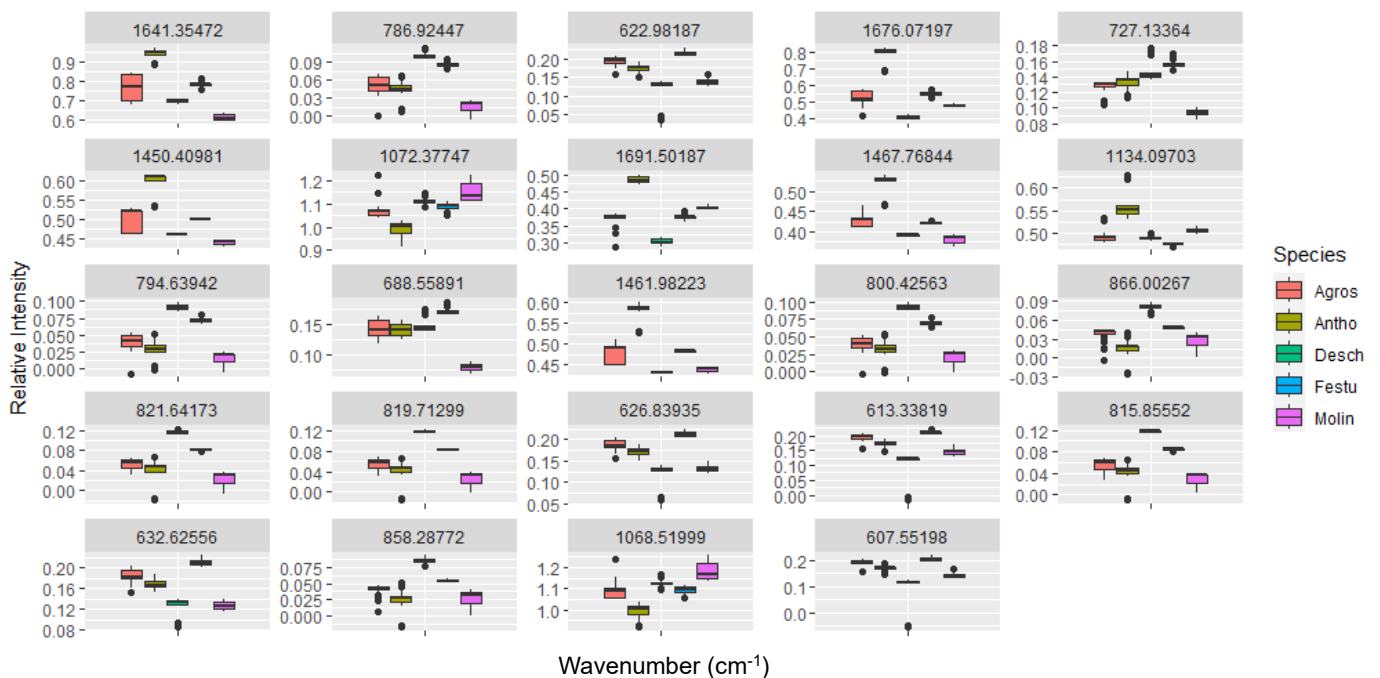
2214 **SM3.6: Boxplots of top 24 wavenumbers identified by randomForest variable**

2215 **importance measure (MeanDecreaseAccuracy)**

2216 *Figure SM3.6: Boxplots of the top 24 important wavenumbers identified by looped*

2217 *randomForest model. Y- axis is relative intensity, and x – axis is species boxplots*

2218 *(colourcoded), plot visually describes within and between species variation.*



2219

2220

2221

2222

2223

2224 **SM3.7: Confusion matrix of refined randomForest model**

2225 *Figure SM3.7: Confusion matrix of refined randomForest model, rows are actual values, and columns*  
2226 *are predicted*

2227

	<b>Agros</b>	<b>Antho</b>	<b>Desch</b>	<b>Festu</b>	<b>Molin</b>
<b>Agros</b>	8	0	0	0	0
<b>Antho</b>	0	8	0	0	0
<b>Desch</b>	0	0	16	0	0
<b>Festu</b>	0	0	0	12	0
<b>Molin</b>	0	0	0	0	7

2228