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

Scoble, L

Review of Palaeobotany and Palynology
Elsevier

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

Manuscript

Authors: Laura Scoble [a], Simon J. Ussher [a], Mark F. Fitzsimons [a], Lauren Ansell [b], Matthew Craven [b], Ralph M. Fyfe [a]

[a] *School of Geography, Earth and Environmental Sciences, University of Plymouth, Plymouth, PL4 8AA, UK.*

[b] *School of Engineering, Computing and Mathematics, University of Plymouth, Plymouth, PL4 8AA, UK.*

Abstract

A growing body of research is demonstrating the potential of Fourier-Transform Infrared spectroscopy (FT-IR) to identify and differentiate morphologically similar pollen taxa. The Poaceae (grass) family is a large and complex with morphologically similar pollen grains. It is not possible to use traditional light microscopy to differentiate Poaceae species, or genus, based on pollen morphological characteristics. This research presents a study of five species from the Poaceae family found across a wide variety of different moorland vegetation communities, to test the extent to which FT-IR microspectroscopy can be used to separate and identify these species and develop statistical approaches for the analyses of these data. Moorland grasses are of particular importance to assess conservation status and baselines in fragile and scarce vegetation communities, whose vegetation composition in the past remains cryptic owing to low taxonomic resolution. Non-differentiated and second derivative spectra were combined with Principal Component Analysis (PCA) and Hierarchical Cluster Analysis (HCA) to determine whether species had different chemical compositions and would cluster. Decision trees and random forest were used to classify each species and demonstrated 100 %

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

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

1. Introduction

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

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

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

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

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

Evidence of FT-IR's ability to identify and separate morphologically similar taxon can be seen in Julier et al. (*2016*), where 12 grass taxa from 8 subfamilies were identified across the grass phylogeny down to subfamily level, with an 80 % success rate. Jardine et al. (2019) classified eight domesticated and wild grasses based on the chemical signature of the pollen grains, achieving a 95% classification success rate when paired with k-nearest neighbour classification and leave-one-out cross validation. Zimmerman et al (*2016*) used FTIR microspectroscopy to classify singular pollen grains which included an optimising technique to prevent Mie-type scattering with a 95 % success rate, thus enabling better taxonomic resolution and classification. For wider context, Steemans et al (*2010*) used FT-IR microspectroscopy to demonstrate that cryptospores have similar spectra to that of trilete spores, which are composed of sporopollenin and characterised by *"absorption bands from aliphatic C–H in methylene (CH_2) and methyl (CH_3) groups, aromatic (C=C and C–H) groups and C=O groups of carboxylic acids*". Fraser et al (*2012*) analysed geologically unaltered sporopollenin from Pennsylvanian (310 million yr before present) cave deposits and demonstrated a strong chemical resemblance to extant relatives. Further comparisons indicated that the sporopollenin structure was similar across broader phylogenetic groups, with Fraser et al (*2012*) suggesting that "*land plant sporopollenin structure had remained stable since embryophytes invaded land*". Depciuch et al (*2018*) selected six *Betula* species to examine their chemical and morphological composition using FTIR. Their data showed that FTIR microspectroscopy could separate and manually characterise each individual chemical composition from most of the six *Betula* species, indicating that the technique can also identify morphologically similar tree taxa.

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

2. Methods

2.1 Sample collection and preparation

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

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

2.2 Chemical Analysis

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

2.3 Spectral pre-processing

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

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

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

2.4 Visual investigation and data analysis

The mean and standard deviation of the non-differentiated spectra was calculated for each species using R v. 4.2.2 (R Core Team, 2022), and plotted for visual investigation (Figure 1) following Jardine's (*2021*) R script. Key absorption bands were chosen from previous research on sporopollenin chemistry and FT-IR Poaceae classification (Table 2) (Julier, et al., 2016; Jardine, et al., 2019; Kendel & Zimmermann, 2020; Zimmerman & Kohler, 2014; Steemans, et al., 2010; Fraser, et al., 2013; Fraser, et al., 2012; Watson, et al., 2007; Zimmerman, et al., 2017) for comparison against the average spectra (Figure 2 and SM3.1). Some absorption bands (e.g., the -OH band at 3300 cm^{-1} and the CH_2 bands at 2925 and 2825 cm^{-1}) were omitted from data analysis as the bands offered no individual

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

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

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

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

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

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

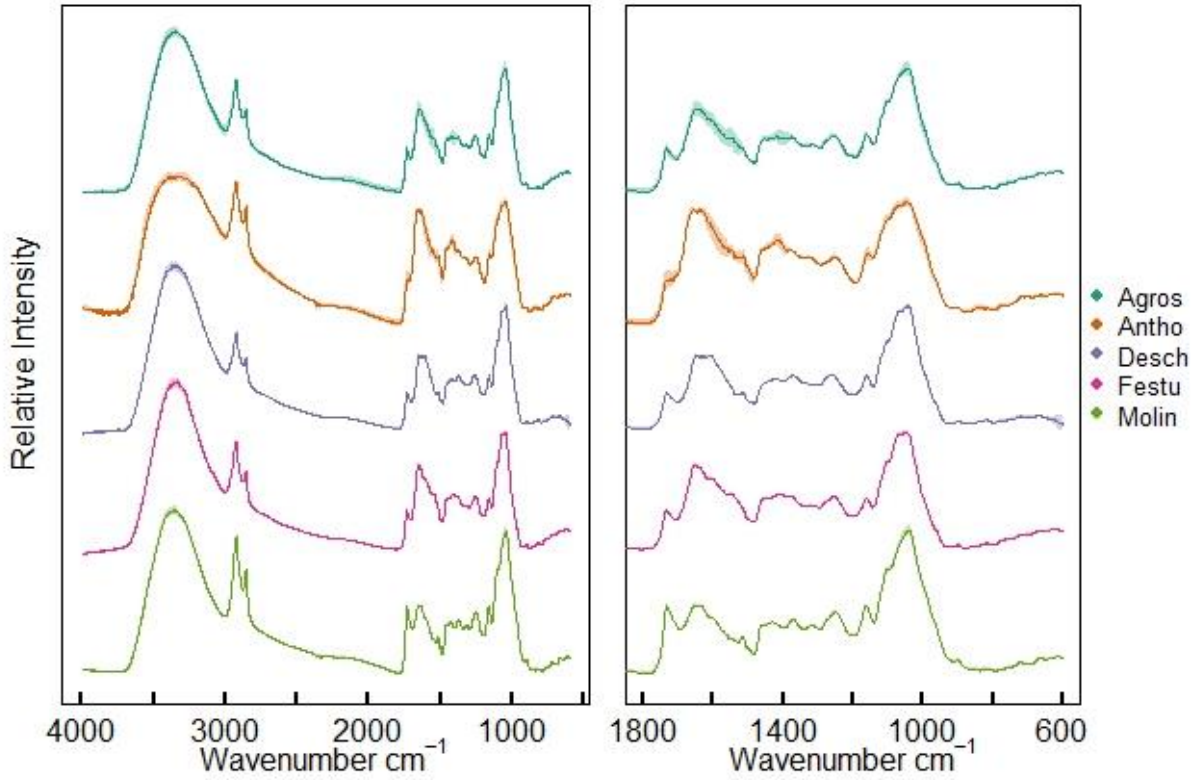
were selected and combined into one data frame. A total of 240 variables were rearranged in ascending order from most important to least important, with the top 24 being selected again and replotted onto a dotchart. The MDA data was transformed into a boxplot to investigate which wavenumbers had greater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

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

3. Results

FT-IR spectra of the five different Poaceae species (Figure 2 and SM3.1) exhibited the characteristic absorbance peaks that have been reported in pollen studies using FT-IR spectroscopy (Table 2) (Julier, et al., 2016; Jardine, et al., 2019; Kendel & Zimmermann, 2020; Zimmerman & Kohler, 2014; Steemans, et al., 2010; Fraser, et al., 2013; Fraser, et al., 2012; Watson, et al., 2007; Zimmerman, et al., 2017). They presented similar vibrational bands, with a broad -OH stretch at 3500-3000 cm^{-1}, asymmetric CH_2 stretches at 2925 and 2845 cm^{-1}, C=C stretch at ~1600 cm^{-1} and C-OH and C-O-C stretches at ~1040 cm^{-1} present in all spectra. Each spectrum signal represents vibrational modes of proteins, lipids, carbohydrates and sporopollenin (Zimmerman, et al., 2017; Zimmerman & Kohler, 2014; Jardine, et al., 2019; Bağcıoğlu, et al., 2015; Zimmerman, et al., 2015). Protein signals are represented at 1650 cm^{-1} (secondary amide I C=O stretch) and ~1550 cm^{-1} (amide II N-H deformation and C−N stretching) ; lipids at 2925 cm^{-1} (asymmetric CH_2 stretch), 2845 cm^{-1} (asymmetric CH_2 stretch), 1740-1710 cm^{-1} (C=O stretch), 1460-1450 cm^{-1} (CH_2 deformation) and 1400 cm^{-1} and carbohydrates between 1200-1000 cm^{-1} (C-O, C-OH and C-O-C stretches). Sporopollenin can be associated with bands at ~1600 cm^{-1} (aromatic C=C stretch), ~1515 cm^{-1} (aromatic C=C stretch), 1161 cm^{-1} (C-O stretch) and between ~900-800 cm^{-1} (C-H bend); and amino acids at 1375 cm^{-1} (symmetric CH_3 bend) and 1325 cm^{-1} (C-N bend). Variance across the spectra is obscure but can be seen with some shading in bands within the fingerprint region. For further investigation, spectral data between 1800-600 cm^{-1} was used to calculate the second derivatives for each specie (SM3.2).

Second derivative spectra (SM3.2) revealed that while there is similarity across all five spectra, most structural change happened between 1800-1400 $cm^{-1,}$ where signals associated with protein and lipids are identified, and 1200-1000 $cm^{-1,}$ where signals 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 cm^{-1}). Shaded areas show ± 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 1740cm^{-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 (1700-1600cm^{-1}) are present across the five species, and all exhibit a C=O stretch
283 at ~1650 cm^{-1} (amide I). *Agrostis*, *Deschampsia cespitosa* and *Festuca ovina* have an
284 aromatic C=C stretch at 1600 cm^{-1}, while *Anthoxanthum odoratum* and *Molinia caerulea*
285 display a C=O stretch at 1630cm^{-1}. *Agrostis*, *Anthoxanthum odoratum* and *Festuca ovina*
286 display downturned symmetric peak at 1550 cm^{-1} (amide II N-H deformation
287 and C–N stretching) and all exhibit peaks at ~1465cm^{-1} (CH_2 deformation). *Agrostis*,

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

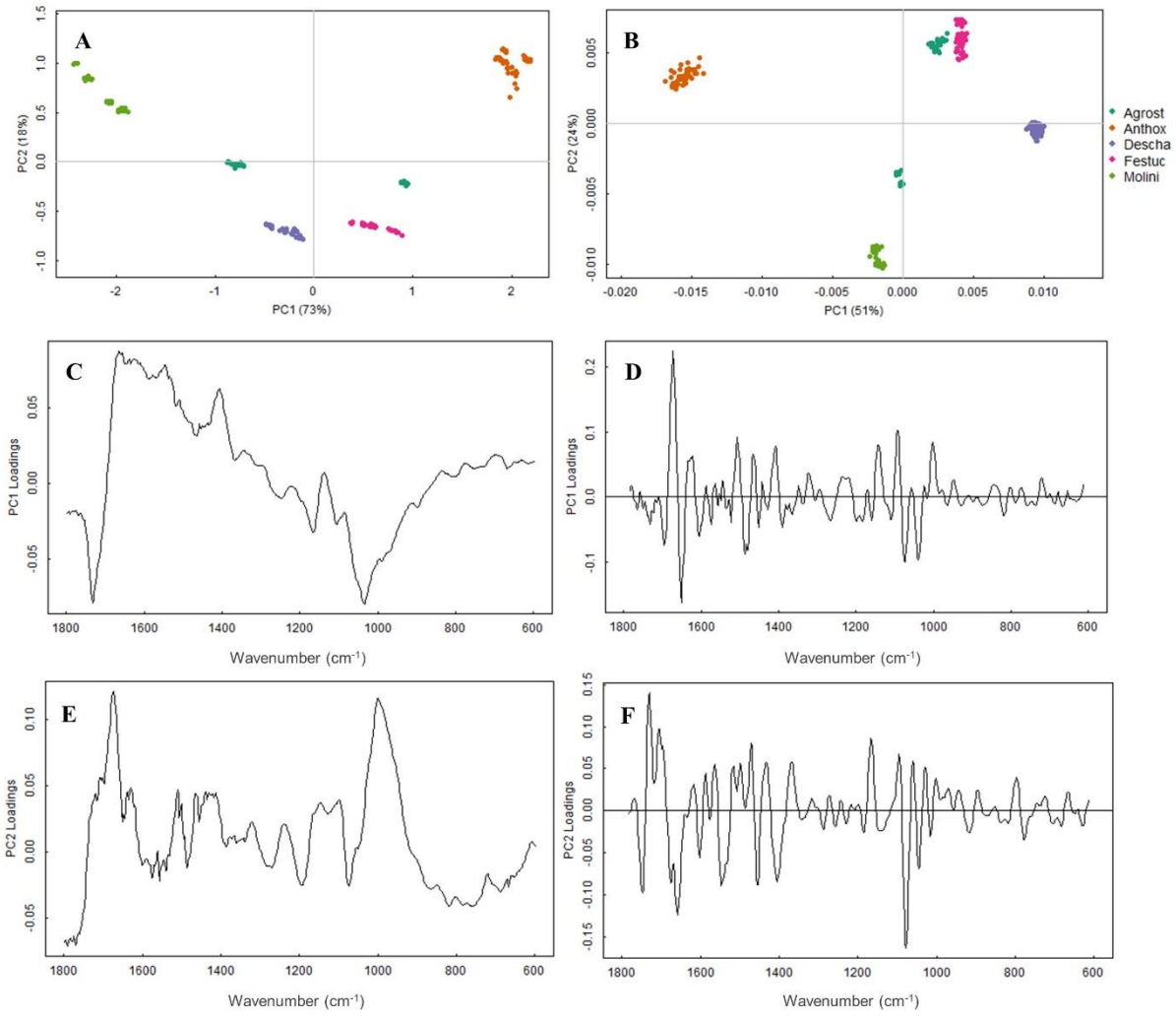


Figure 3: (A): Non differentiated PCA and loading plots (C & E); (B) Second derivative (Savitsky-Golay smoothed) and loading plots (D & F).

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

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

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

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

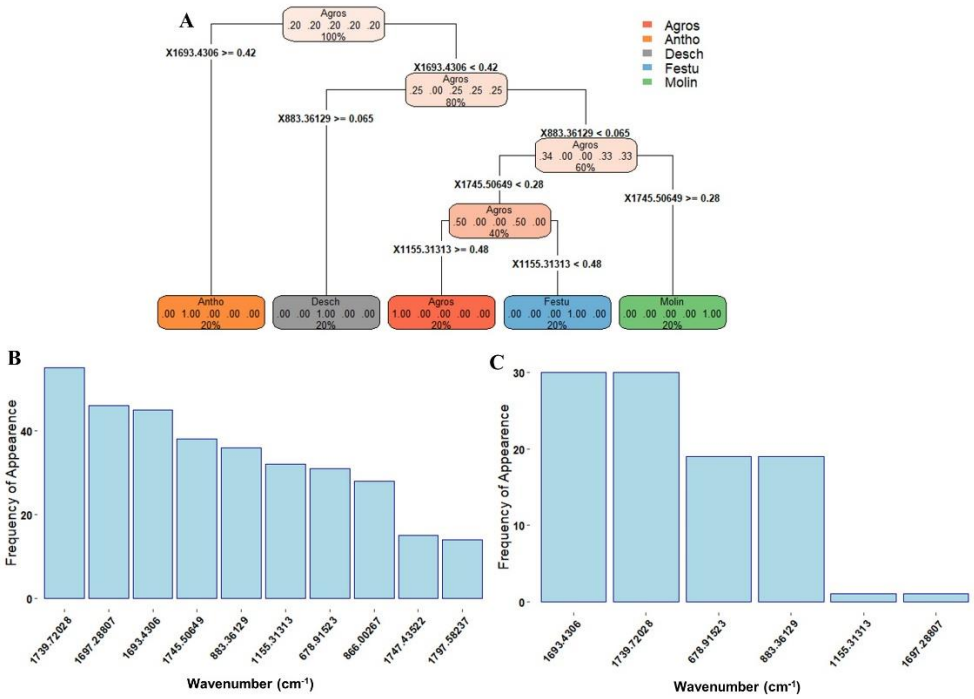
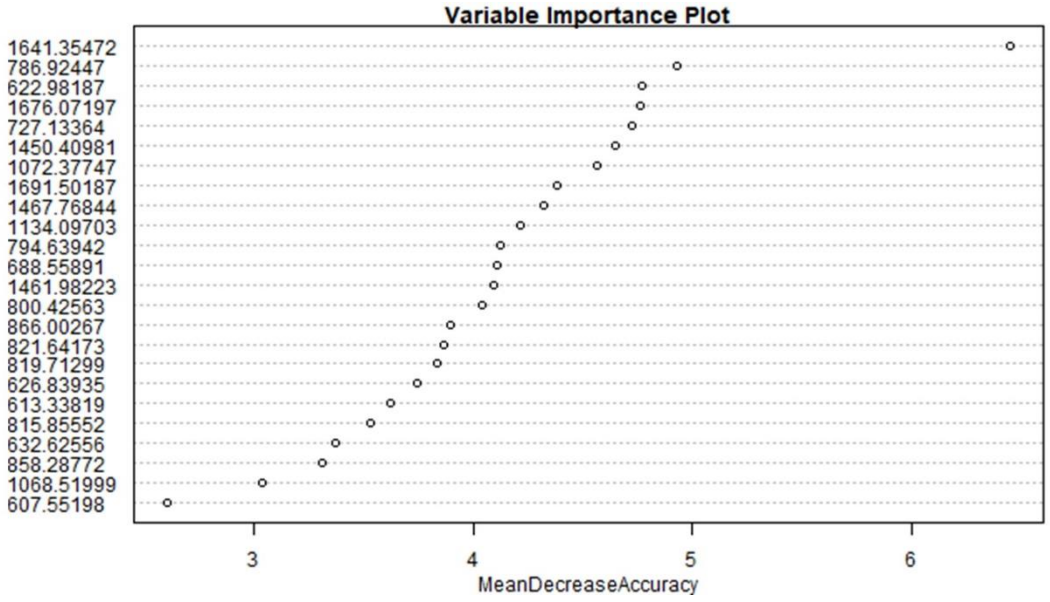


Figure 4: (A) Classification tree of pre-processed data, top decimals of terminal nodes represent successful classification, with bottom percentage indicating how many samples have been classified into that node. (B) Histogram of extracted wavenumbers used as decision rules and the frequency of appearance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

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

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

364 1461.98223 cm^{-1} and 1450.40981 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 cm^{-1} and 1467.76844 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

4. Discussion

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

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

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

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

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

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

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

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

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

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

Using random forest (RF) (Donges, 2023) achieved 100% successful classification and prediction accuracy for the pre-processed data respectively. While the decision tree (Figure 4A) had 100% successful classification, it is prone to overfitting, something of which RF overcomes by bootstrapping samples (Petkovic, et al., 2018). While RF has classed each species using a randomised algorithm, it has not separated the two potential *Agrostis* species exhibited in the PCAs (Figure 3B). Therefore, it would be beneficial to investigate whether RF can class different species from the same genus, instead of the same family. MeanDecreaseAccuracy (MDA) plot (Figure 5) highlighted wavenumbers between 1800-1600 cm^{-1}, 1500-1400 cm^{-1} and 1200-800 cm^{-1}, suggesting that the wavenumbers found within the lipid (1750-1730 cm^{-1}), secondary amide I (1700-1600 cm^{-1}), carbohydrate (1200-900cm^{-1}) and aromatic ring (800-600 cm^{-1}) regions have stronger influence on the RF classification than the rest of the dataset. By transforming the MDA data into a boxplot (SM3.6), it revealed *Agrostis* had greater within species variation within the protein (1691.35472cm^{-1}) and lipids (1461.98223 and 1450.981 cm^{-1}) region, this was also found between species. While the combined decision tree rules (Figure 4B) and RF's MDA plot (Figure 5) display similar but different wavenumbers, they are different machine learning techniques that are classifying the grass spectra successfully overall. By refining the data set used to train RF, classification accuracy was optimised and OOB error continued to stay at 0%. Furthermore, a 100% successful prediction (SM3.7) was also achieved when labels were removed from the test data set, suggesting that the specific 24 variables included in the overall model are driving the discrepancy between species. Optimising and using a machine learning technique that has managed to separate samples into their correct species successfully, demonstrates that FT-IR spectra can be used to separate and classify morphologically similar grass species if paired with multivariate data analysis.

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

above should focus more on the sporopollenin chemistry to aid in better classification at genus and species level.

5. Conclusions

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

6. Acknowledgement

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Therneau, T. & Akinson, B., 2022. *Recursive Partitioning and Regression Trees.* [Online]
Available at: https://CRAN.R-project.org/package=rpart
[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

Tables

Table 1: Sample name, location of collection and correlating legend

Sample name	Location	Legend
Agrostis	Holywell Pond	A
Anthoxanthum odoratum	Holywell Pond	A
Deschampsia cespitosa	Whitelee Moor	D
Festuca ovina	Whitelee Moor	C
Molinia caerulea	Milburn	B

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

Group	Wavenumber (cm-1)	*Agrostis*	*Anthoxanthum odoratum*	*Deschampsia cespitosa*	*Festuca ovina*	*Molinia caerulea*
-OH	3500-3000	+++	++	++	++	++
CH_2 (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_2(d)	1460-1450	sh	/	/	sh	sh
	~1400	+	+	/	sh	/
CH_3 (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	/	/	/	/	/

SM1: Parameter Experiment

Laura Scoble [a], Simon J. Ussher [a], Mark F. Fitzsimons [a], Lauren Ansell [b], Matthew Craven [b], Ralph M. Fyfe [a]

[a] *School of Geography, Earth and Environmental Sciences, University of Plymouth, Plymouth, PL4 8AA, UK.*

[b] *School of Engineering, Computing and Mathematics, University of Plymouth, Plymouth, PL4 8AA, UK.*

1. Introduction

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

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

2. Methods

2.1 Sample preparation

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

2.2 Chemical Analysis

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

2.3 Data Analysis

Data analysis was conducted on both scan rate and resolution data in R v.3.1.4 (Team, 2022), and focused on a specific wavenumber (scan rate and resolution: 1654 cm^{-1}) to compare and evaluate whether there was a significant statistical difference between each parameter. Mean absorbance units were calculated for each and plotted onto a boxplot for visual analysis. Null hypothesis stated (H_0): all mean values were equal; the alternative hypothesis stated (H_1): not all mean values were equal. If all mean values were equal then there was no significant difference between the scan rates/resolution and changing the

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

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

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

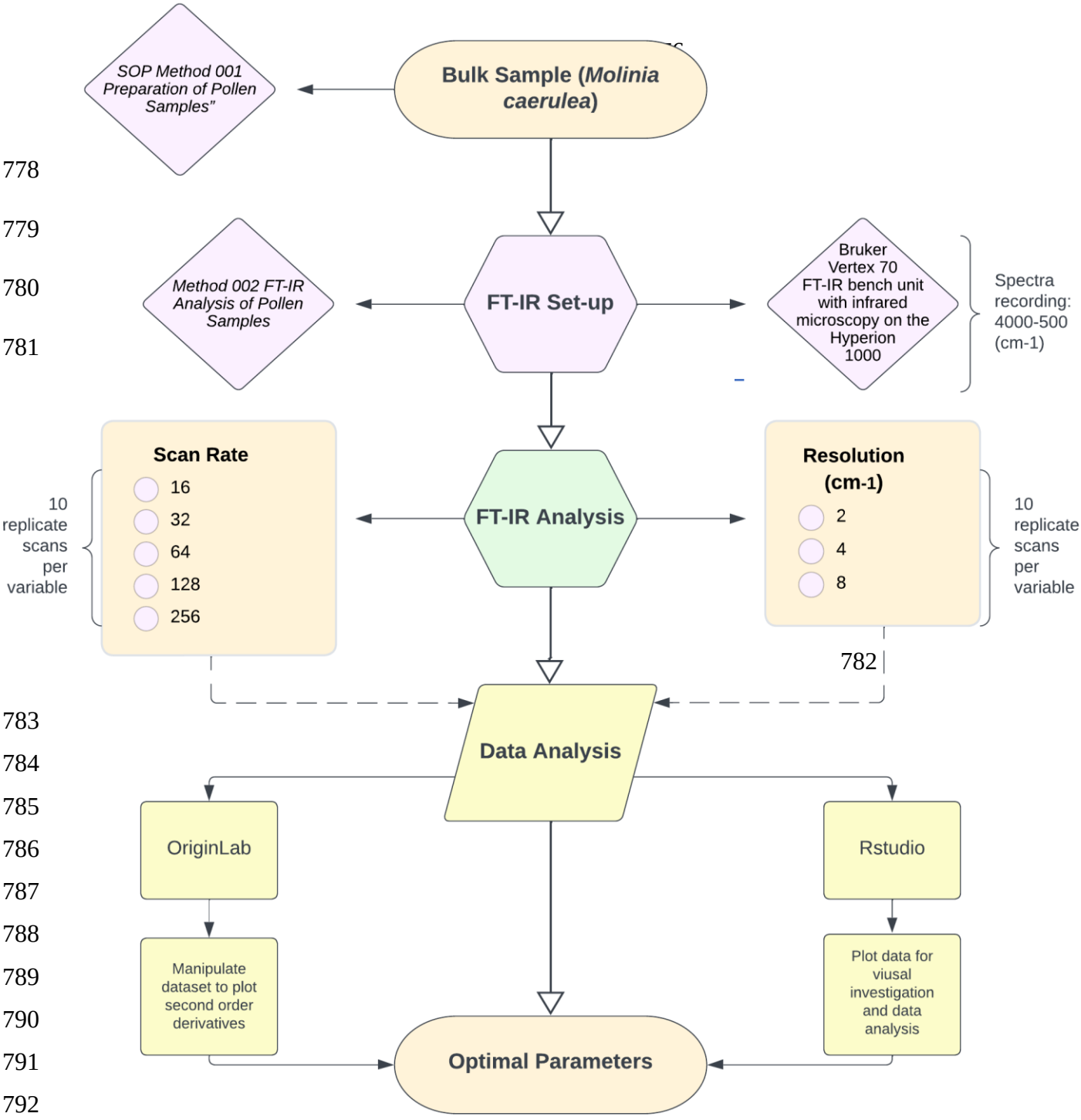


Figure SMI.1: Parameter methodology flowchart

(Scoble, 2023)

3. Results

3.1 Scan rate

Five different scan rates are presented in Figure SM1.2: 16 (A), 32 (B), 64 (C), 128 (D) and 256 cm^{-1} (E) of *Molinia caerulea*. Lower scan rate numbers (A, B, C) exhibit higher scattering and noise, whereas the higher scan rates (D and E) are more detailed and smoother (Figure SM1.3). The broad -OH stretch (3300cm^{-1}) has reduced noise exhibited in D and E, with the asymmetric CH_2 (2923 and 2854cm^{-1}) stretch exhibiting peak separation compared to the shouldering seen in A, B and C. The fingerprint region has a stronger level of absorbance in D and E, with C-O stretch (1163 and 1041cm^{-1}) becoming more pronounced as the scan rate increases.

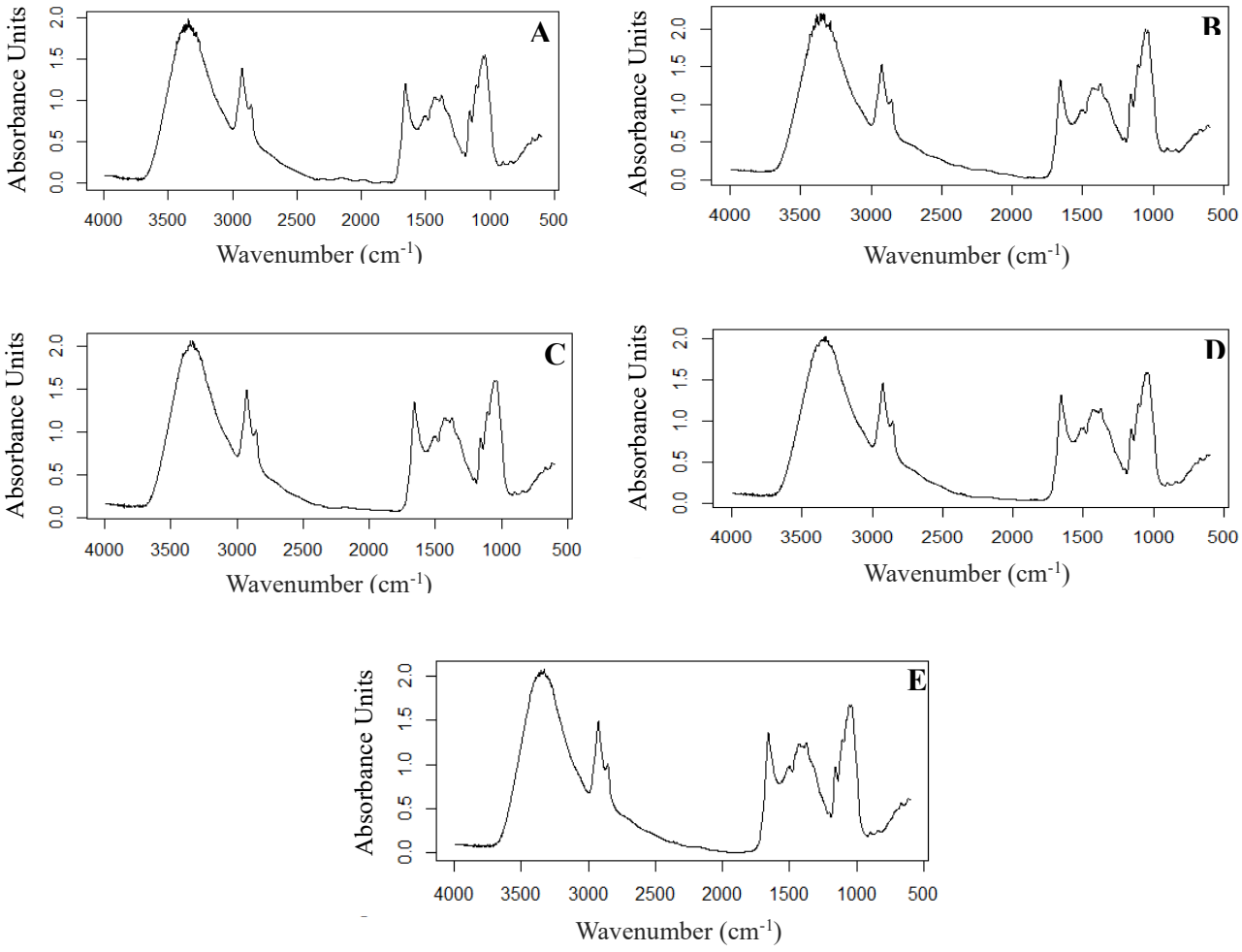


Figure SM1.2: 10 replicate scans of scan rate 16 (A),32 (B),64 (C),128 (D) and 256 (E) of Molinia caerulea, averaged and plotted

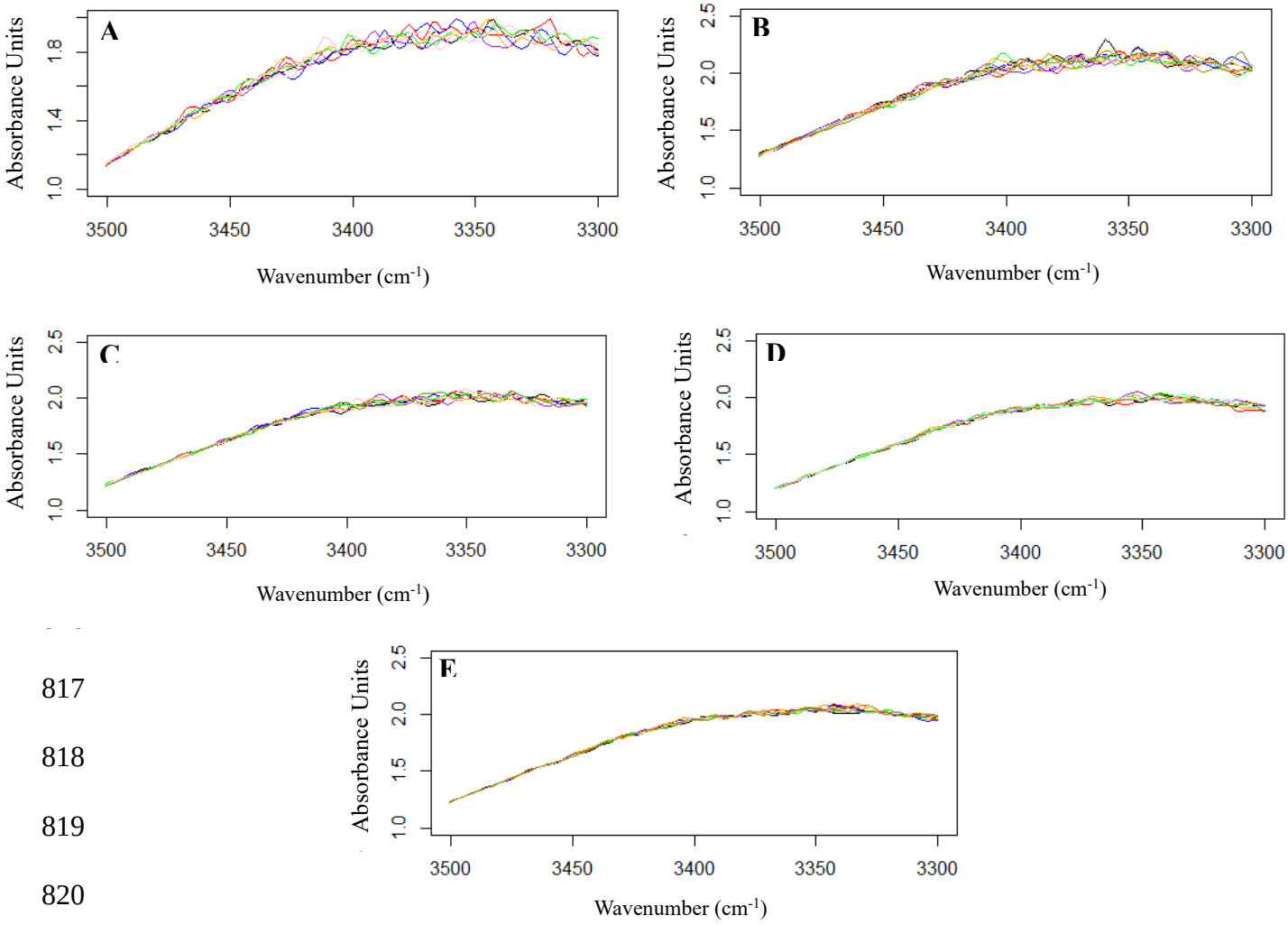


Figure SM1.3: Scattering of -OH peak at 3350 cm^{-1} at different scan rates: 16 (A),32 (B),64 (C),128 (D) and 256 (E) of Molinia caerulea.

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 (3300cm^{-1}) reduces, with the broad band becoming nearly fully suppressed to baseline. Sharper upturned peaks between 1700-1500cm^{-1} (C=O and C=C stretch) can be seen throughout Figure SM1.4 with D and E having more distinct separation between 1250-1000cm^{-1} (C-O). The use of second

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

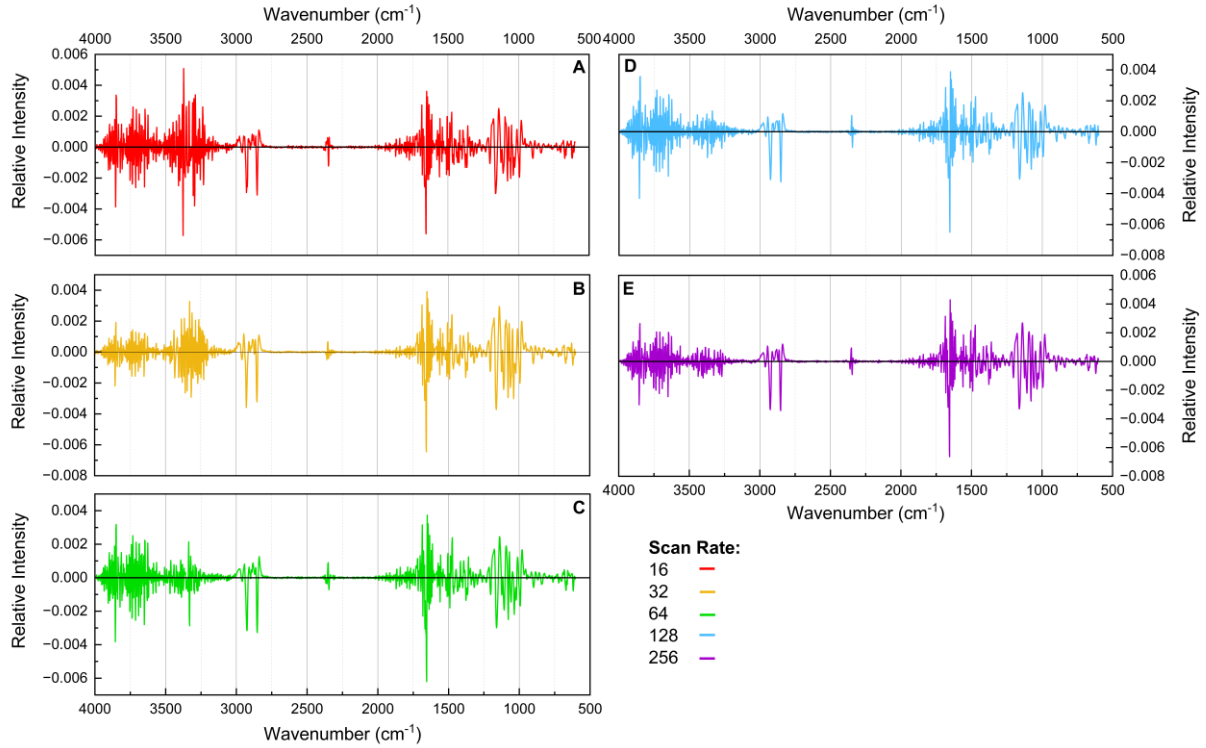


Figure SM1.4: Second derivates of scan rates 16 (A),32 (B),64 (C),128 (D) and 256 (E) of Molinia caerulea.

.

3.1.3 Data Analysis

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

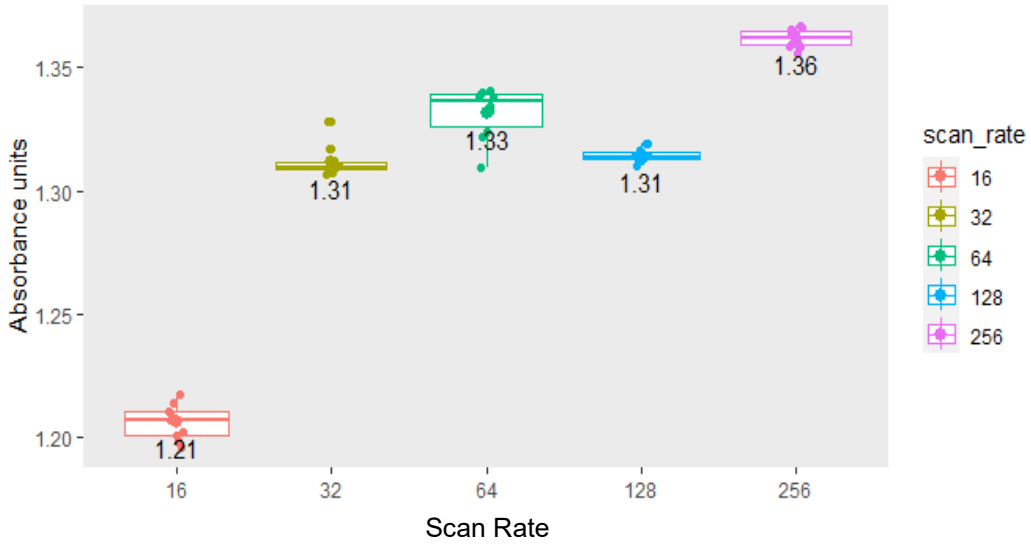
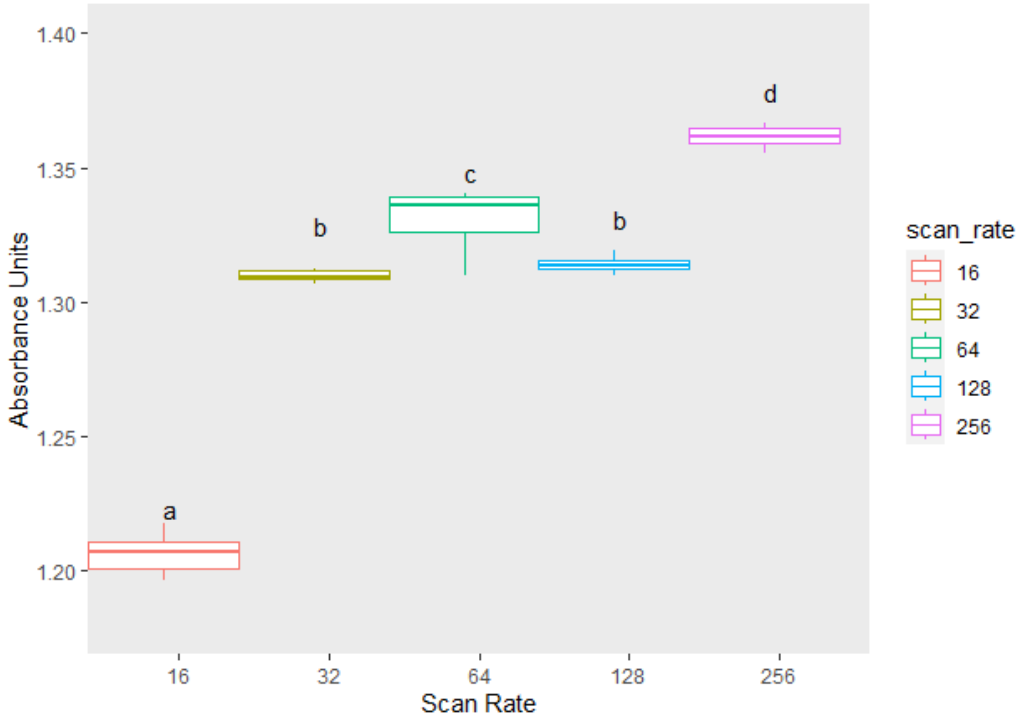


Figure SM1.5: Boxplot of scan rate mean absorbance unit values.

Figure SM1.5 shows box plots of the five scan rates and the mean absorbance unit values. There is a noticeable increase in absorbance unit mean value as the scan rate is increased to 64, then gradually decreases to 128 and then increases again at 256. 32 and 128 have the same mean value (1.31), suggesting there is no difference between both scan rates. 16 has a short boxplot with longer whiskers, indicating a wide distribution of data compared to 32, 128 and 256. 32 exhibits a thin box plot and wider scattering, with two outliers at 1.320 and one at 1.316. 64 has the widest box plot with scattered data, indicating variance within absorbance unit values for 1654cm^{-1}. 128 exhibits tight clustering with a thin box plot, indicating less variance within the data compared to between scan rate groups. 256 has the highest mean value (1.36) and a thin box plot with tight clustering, most of the data points are plotted around the mean.

An ANOVA test was run to determine whether the mean values were significantly different from one another, working on the H_0 hypothesis that all mean values were equal, indicating there was no significant difference between scan rates. The *p*-value was "1.503903e-40" (1.503903×10^{-40}) which was <0.05, concluding that the mean values were significantly different from one another. A Tukey Honestly Significant Difference

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



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

3.2 Resolution

Below are three different resolutions (Figure SM1.7): $2cm^{-1}$ (F), $4cm^{-1}$ (G) and $8cm^{-1}$ (H) using a scan rate of 256. F displays a noisy spectrum with a non-linear spectral line. Bands are well defined, but some appear to be sharp instead of broad because of the excess noise, e.g., -OH stretch ($3300(cm^{-1})$. The fingerprint region has some recognisable bands, however, the C=C shouldering at roughly $1500cm^{-1}$ is challenging to identify. G has considerably less noise across the spectrum. Peaks and shoulders can be clearly differentiated as the spectral line looks more linear. The start of the spectrum is closer to the baseline and more distinguishable. H has a non-linear spectral line with weak absorbance. Bands are challenging to identify, especially within the fingerprint region, e.g. the anti-symmetric CH_2 bend ($1433cm^{-1}$) and symmetric CH_3 bend ($1373cm^{-1}$). At the very end of the spectrum, the peaks dip below 0 absorbance.

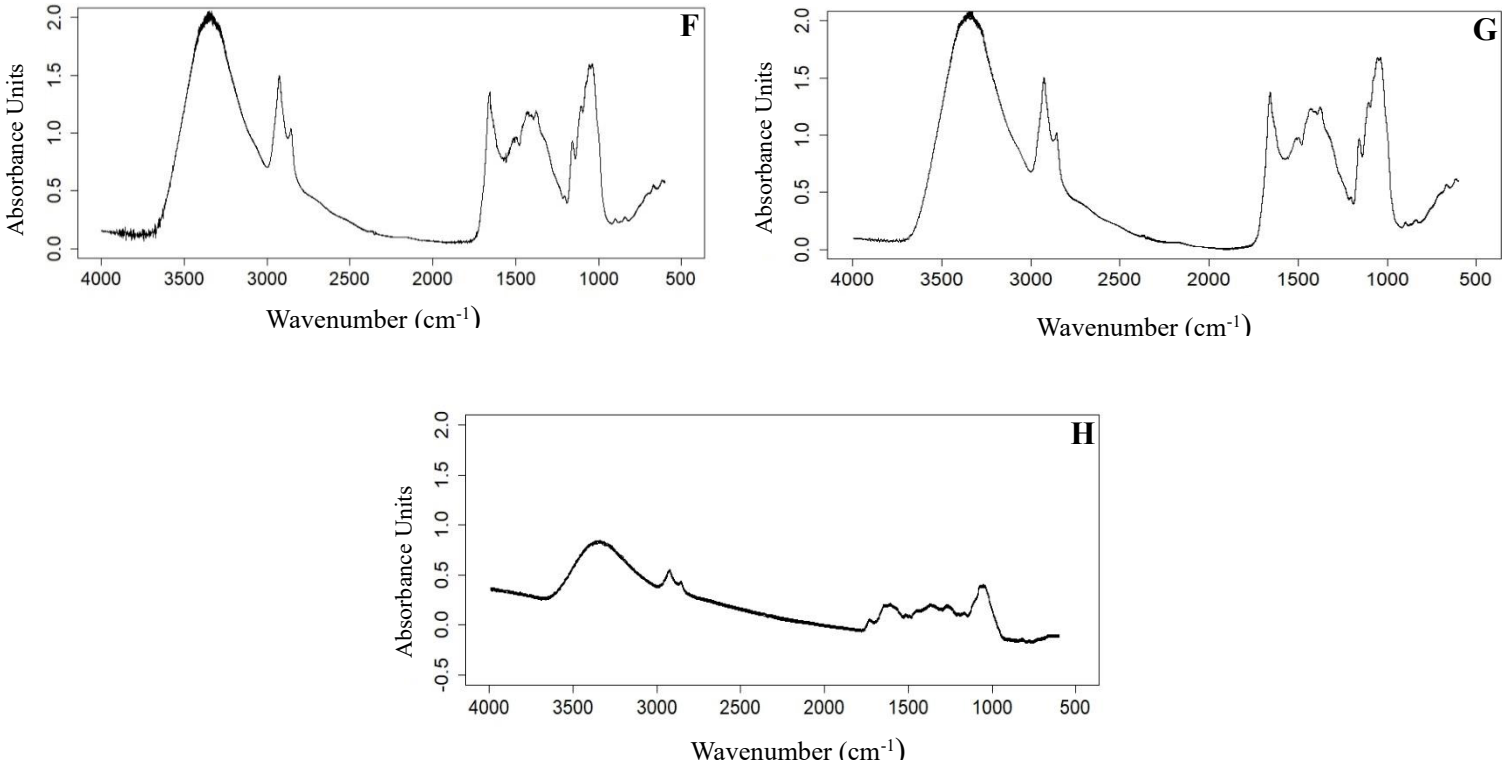


Figure SM1.7:10 replicate scans of resolution 2 (F), 4 (G) and $8cm^{-1}$(H) of Molinia caerulea, averaged and plotted using R.

3.2.1 Second derivatives

Figure SM1.8 presents second derivatives for the resolutions (cm^{-1}). All three have very different absorbances, with F having the strongest and noisiest spectrum. F exhibits full supression of the -OH stretch (3300 cm^{-1}), asymmetric CH_2 stretch (2923 and 2854 cm^{-1}) and the C-O stretch (1163 and 1041 cm^{-1}). More obscure peaks cannot be identified as the spectrum is compacted. G exhibits a familiar spectrum with resolved peaks pointing downwards and a more defined fingerprint region. Strong C=C bands at roughly 1600cm^{-1} and C-O stretches between 1100-1000cm^{-1} are presented. Upward peaks can be seen between 1700-1600 cm^{-1} and 1200-1100 cm^{-1}, indicative of a C=O and C-O stretch, respectively. Noise is still present at the beginning of the spectrum but not as strong. H has wider spacing between peaks, and very strong peaks. Resolved peaks have a strong negative absorbance with the asymmetric CH_2 stretch (2854 cm^{-1}) measured at -0.0008. More pronounced upturned peaks are exhibited between 1400-1250cm^{-1}, which could be indicative of symmetric CH_3 stretch.

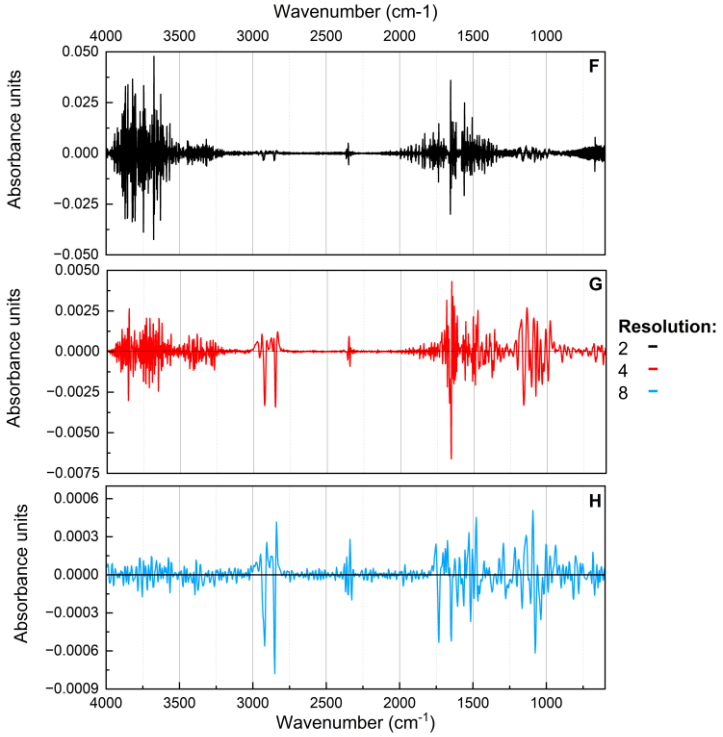
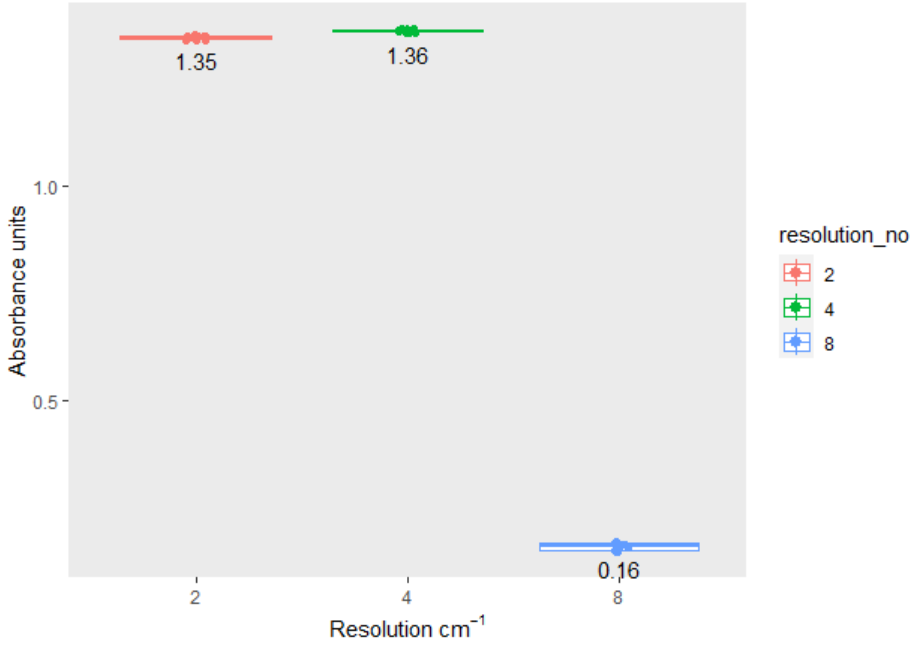


Figure SM1.8: second derivatives of resolution 2 (F), 4 (G) and 8cm^{-1}(H) of Molinia caerulea

907 *3.1.3 Data analysis*

908 Mean absorbance unit values for each resolution (cm^{-1}) were calculated by selecting a
909 specific wavenumber (1654 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.



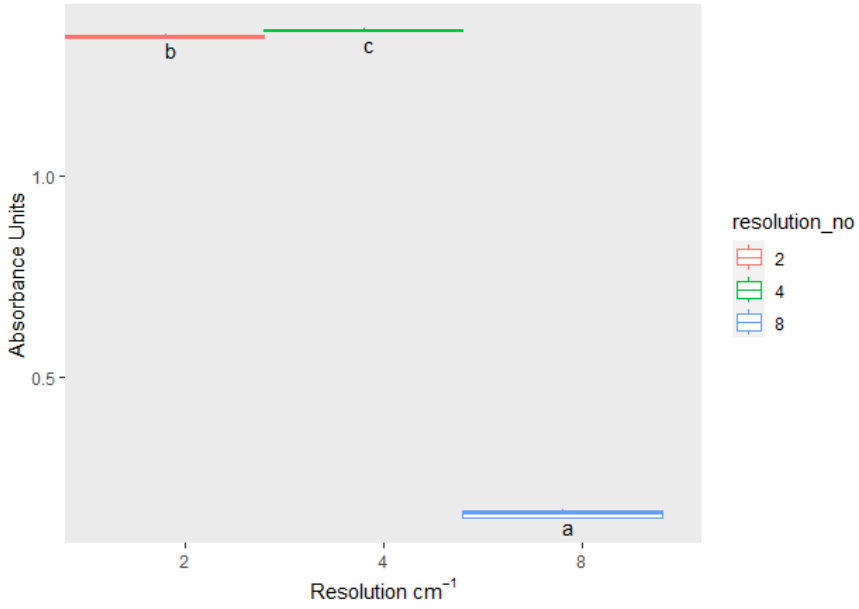
913 *Figure SM1.9: Boxplot of resolution mean absorbance unit value.*

914

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 $4cm^{-1}$ and $8cm^{-1}$. $2cm^{-1}$ and $4cm^{-1}$ exhibit tight
918 clustering of data with thin box plots, indicating less variance within. $8cm^{-1}$ 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 "3.763198e-55" (3.763198×10^{-55})
923 which is <0.05, concluding that the mean values are significantly different from one

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



.

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

4. Discussion

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

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

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

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

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

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

shortens the time between scans, decreasing the degree of fineness obtained (Ota, 2007). Chemical signals are difficult to identify especially within the fingerprint region, e.g the anti-symmetric CH_2 bend (1433 cm^{-1}) and symmetric CH_3 bend (1373 cm^{-1}). At the very end of the spectrum the peaks dip below 0 absorbance. If this was seen across the other resolution spectra, this could be indicative that the background scan was taken incorrectly or the ATR cell wasn't cleaned sufficiently beforehand. However, as the negative absorbance is only present in 8cm^{-1} (H) fingerprint region, it is highly probable that this is a result of the lack of resolving power and detail.

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

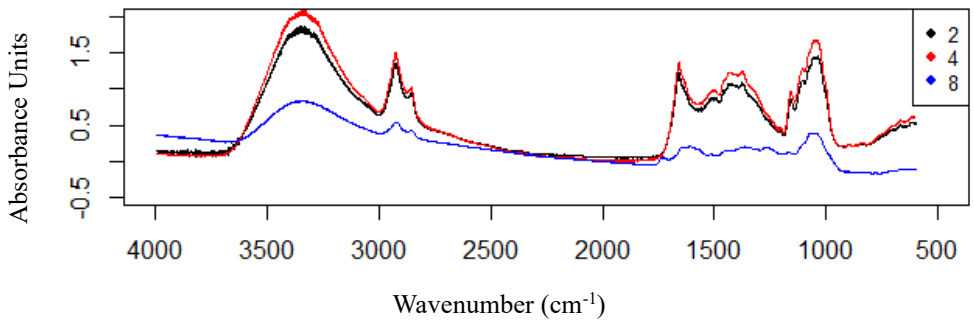


Figure SM1.11: Average scans of resolution 2,4 and 8cm^{-1} of Molinia caerulea.

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

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

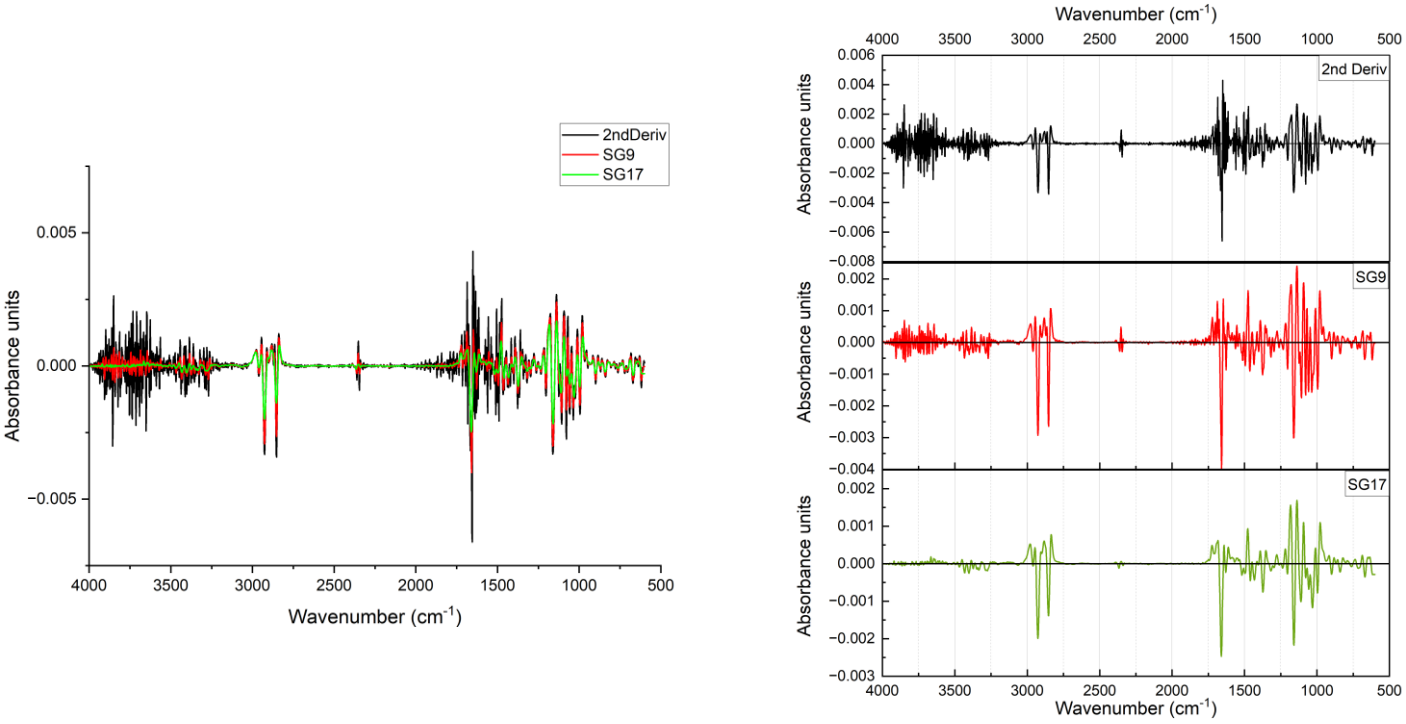


Figure SM1.12: Multiple line plot (a) and stacked line plot (b) of resolution 4cm^{-1} (G) second order derivative (black), Savitzky-Golay (SG) smoothing – window size 9 (red) and size 17 (green).

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

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

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

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

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

5. Conclusion

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

5. Bibliography

Barra, I. et al., 2021. Optimizing setup of scan number in FTIR spectroscopy using the moment distance index and PLS regression: application to soil spectroscopy. *Scientific Reports,* Volume 11, p. 13558.

Hothorn, T., Bretz, F. & Westfall, P., 2008. Simultaneous Inference in General Parametric Models. *Biometrical Journal,* 50(3), pp. 346-363.

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

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

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

Lenth, R., 2023. *emmeans: Estimated Marginal Means, aka Least-Squares Means.* [Online] Available at: https://CRAN.R-project.org/package=emmeans [Accessed 17 March 2023].

Ota, H., 2007. Resolution and Aperture. *FTIR Talk Letter Vol.8*, October, pp. 02-07.

Savitzky, A. & Golay, M. J. E., 1964. Smoothing and Differentiation of Data by Simplified Least Squares Procedures. *Anal. Chem,* p. 36.

Schlindwein, H. S., 2020. *About Spectral Resolution in FT-IR Spectroscopy.* [Online] Available at: https://www.opticsblog.bruker.com/guide-to-spectral-resolution-in-ft-ir/ [Accessed 13 January 2023].

Scoble, L., 2023. *Parameter flowchart.* Plymouth: s.n.

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

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

Wickham, H., 2022. *stringr: Simple, Consistent Wrappers for Common String Operations.* [Online] Available at: https://CRAN.R-project.org/package=stringr [Accessed 17 March 2023].

Wickham, H. et al., 2023. *dplyr: A Grammar of Data Manipulation. R Package 1.1.0.* [Online] Available at: https://CRAN.R-project.org/package=dplyr [Accessed 15 March 2023].

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

```
##############################################
##  Scoble, L (2023) R script for           ##
##  pre-processing spectral data, data analysis  ##
##  and classification.                     ##
##############################################
```

Table of Contents

```
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
```

```
## stick all scans together
## requires individual files with a .dtp extension


#list all files (with .dpt extension)
file.list <- list.files(pattern = "\\.dpt$") #only lists dpt files
#make empty dataframe
df <- read.csv(file.list[1], header = F)
df <- rbind(c("wavelength","Agrostis"), df)

#loop across all files and stick together into single file
count = 1

for(i in file.list){
  print(paste("count =", count, i)) #flag for progress
  #read in individual file
  dat <- read.csv(i, header = F)
  #extract sample code from filename
  sample <- gsub(".dpt", "", i)
  #append the data to the dataframe
  dat <- rbind(c("wavelength",  sample), dat)
  df <- cbind(df, dat[,2])

  count = count + 1
}

#prepare the combined file for export
df <- df[,-2] #drops col 2 (duplicate data)
colnames(df) <- df[1,] #define column names as sample names
df <- df[-1,] #drop top row (the un-needed names)

#export to csv format
write.csv(df, "all.data.scans.final.csv", row.names = F)

```

```
##############################################################
## Baseline, EMSC correction and 2nd Derivative              ##
## Scoble, L (2023)                                          ##
##############################################################

##### Read data in #####
Species_data <- read.csv("all.data.scans.final.csv", check.names = F)
Species_data <- data.frame(t(Species_data))
colnames(Species_data) <- Species_data[1,]
Species_data <- Species_data[-1,]

##### Non-differentiated spectra #####

# Baseline correction
species.baseline <- baseline(as.matrix(Species_data), method = "modpolyfit", deg = 2)
species.corrected <- data.frame(species.baseline@corrected)
colnames(species.corrected) <- colnames(Species_data)
species.corrected <- as.data.frame(species.corrected,
                                   row.names = rownames(Species_data))

# EMSC correction of baseline corrected data
Species.emsc1 <- EMSC(species.corrected, degree = 3,
                      reference = colMeans(species.corrected))


emsc.corrected1 <- data.frame(t(Species.emsc1$corrected))

# Write file
write.csv(emsc.corrected1, "all.data.emsc.baseline.final.csv", row.names = T)
```

```
##### Second derivative of original data #####
Species.derivtwo <- savitzkyGolay(Species_data, p = 2, w = 15, m = 2)

Species.derivtwo <- as.data.frame(Species.derivtwo, row.names = rownames(Species_data))

# EMSC correction of second derivative data
Species.deriv.emsc <- EMSC(Species.derivtwo, degree = 1,
                           reference = colMeans(Species.derivtwo))

Species.deriv.emsc <- Species.deriv.emsc$corrected

# Write file
Species.deriv.emsc.pca <- data.frame(Species.deriv.emsc)
names(Species.deriv.emsc)<-sapply(str_remove_all(colnames(Species.deriv.emsc),"X"),"[")

write.csv(Species.deriv.emsc.pca, "Species.deriv.emsc.final.csv")


##### Prepare derivative file for OriginLabs #####

# Remove row names into column
Species.deriv.emsco  <- cbind(rownames(Species.deriv.emsc.pca),
data.frame(Species.deriv.emsc.pca, row.names = NULL))

# Create new short names
Species <- c(rep("Agros", 51),
             rep("Antho", 51),
             rep("Desch", 51),
             rep("Festu", 50),
             rep("Molin", 50))

#Factorise
Species <- as.factor(Species)

```

```
#Bind the two together
Species.deriv.emsco <- cbind(Species, Species.deriv.emsco)

#Remove the sample labels
Species.deriv.emsco <- Species.deriv.emsco[,-2]

names(Species.deriv.emsco)<-
sapply(str_remove_all(colnames(Species.deriv.emsco),"X"),"[")

# Average of each species
Species.derivo.means <- aggregate(Species.deriv.emsco[,2:1749],
                                  by = list(Species),
                                  FUN = mean)
rownames(Species.derivo.means) <- Species.derivo.means[,1]
Species.derivo.means <- Species.derivo.means[,-1]

# Write file
Species.derivo.means <- data.frame(t(Species.derivo.means))
write.csv(Species.derivo.means, "Origin.means.emsc.all.final.csv")
```

```
##########################################
##  Plot non-differentiated spectra     ##
##  following parts of Jardine (2021)   ##
##  R script.                           ##
## Scoble, L (2023)                     ##
##########################################

# Read in file
Ad1 <- read.csv("all.data.emsc.baseline.final.csv", check.names = F, row.names = 1)
Ad1 <- data.frame(t(Ad1))
names(Ad1)<-sapply(str_remove_all(colnames(Ad1),"X"),"[")
Ad2 <- Ad1
Ad1 <- Ad1[1:253,]
str(Ad1)

# Remove row names into column
Ad1 <- cbind(rownames(Ad1), data.frame(Ad1, row.names = NULL))

# Create new short names
Species <- c(rep("Agros", 51),
             rep("Antho", 51),
             rep("Desch", 51),
             rep("Festu", 50),
             rep("Molin", 50))

#Factorise
Species <- as.factor(Species)

#Bind the two together
Ad1 <- cbind(Species, Ad1)

#Remove the sample labels
Ad1 <- Ad1[,-2]

```

```
##### Mean and Standard Deviation #####
grass.means <- aggregate(Ad2,
                         by = list(Species),
                        FUN = mean)

rownames(grass.means) <- grass.means[,1]
grass.means <- grass.means[,-1]

grass.sd <- aggregate(Ad2,
                      by = list(Species),
                     FUN = sd)
rownames(grass.sd) <- grass.sd[,1]
grass.sd <- grass.sd[,-1]

#### For Origin Plots #####
grass.means <- data.frame(t(grass.means))
write.csv(grass.means, "Grass.means.origin.csv")

##### Plot Data #####
grass.means <- data.frame(t(grass.means))
# Full spectra (first plot only)
par(mfrow = c(1,2), mar = c(3,2,1,0) + 0.01)

#fingerprint region (second plot only)
par(mar = c(3, 1, 1, 3) + 0.01)

col <- brewer.pal(5, "Dark2")

# Select colours from RColorBrewer
speciescol  <- c("#1B9E77","#D95F02", "#7570B3", "#E7298A", "#66A61E")

yvals <- seq(from = 4.5, to = 0.05, length.out = 5)

wavenumber <- (gsub("X","",colnames(Ad1[,2:1763])))
```

```
wavenumber <- as.numeric(wavenumber)

length(wavenumber)
length(grass.means[1,])
# Change xlim value for second plot (1800-600)
plot(wavenumber, grass.means[1,], las = 1,
     type = "n", xlim = c(1800, 600), ylim = c(0, 6),
     xlab = "", ylab = "",
     yaxt = "n", xaxt = "n")

for(i in 5:1) {
  col.e <- col2rgb(speciescol[i])
  polygon(c(wavenumber, rev(wavenumber)),
          c(grass.means[i,]+yvals[i]+grass.sd[i,],
            rev(grass.means[i,]+yvals[i]-grass.sd[i,])),
          col = rgb(col.e[1], col.e[2], col.e[3], alpha = 80, maxColorValue = 255),
          border = NA)

for(i in 5:1) {
  lines(wavenumber, grass.means[i,]+yvals[i],
        col = speciescol[i])

}}

axis(1, lwd = 0, lwd.ticks = 2, tcl = 0.3,
     mgp = c(1.5, 0.2, 0),
     las = 1)
mtext(expression("Wavenumber cm"^-1), side = 1, line = 1)

# For first plot only
mtext("Relative Intensity", side = 2, line = 0.6, las = 0)
# For second plot only
legend.names <- cbind.data.frame(row.names(grass.means), grass.means)
#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
```

```
##### ##############################
## Fyfe, R and Scoble, L (2023)        ##
## PCA Analysis                        ##
##################################

# Non differentiated spectra
dfbaseemsc <- read.csv("all.data.emsc.baseline.final.csv", check.names = F, row.names = 1)
dfbaseemsc <- data.frame(t(dfbaseemsc))
names(dfbaseemsc)<-sapply(str_remove_all(colnames(dfbaseemsc),"X"),"[")

# Truncate
dfbaseemsc1 <- dfbaseemsc[1141:ncol(dfbaseemsc)]

# Remove anomalies
dfbaseemsc2 <- dfbaseemsc1[-c(17, 18, 19, 90, 91, 92, 93, 95, 96, 97, 98, 99),]

#Rename to make easier for plot
df.trunc <- dfbaseemsc2


##### Second Derivative Spectra (all data) #####

df.trunc1 <- read.csv("Species.deriv.emsc.final.csv", check.names = F, row.names = 1)

# Truncate (134 instead of 1141 so data starts at same wavenumber)
df.trunc2 <- df.trunc1[1134:ncol(df.trunc1)]
df.trunc2 <- df.trunc2[-c(17, 18, 19, 90, 91, 92, 93, 95, 96, 97, 98, 99),]
names(df.trunc2)<-sapply(str_remove_all(colnames(df.trunc2),"X"),"[")

##### Plot PCA, replace df.trunc with df.trunc2 for second plot #####
dfsmoo.pca <- prcomp(df.trunc)
dfsmoo.pca.scores <- as.data.frame(dfsmoo.pca$x)
dfsmoo.pca.scores <- cbind.data.frame(row.names(df.trunc), dfsmoo.pca.scores[,1:5])
summary(dfsmoo.pca)
```

```
#make and populate a new column with a short species name
dfsmoo.pca.scores$spec <- substr(dfsmoo.pca.scores$`row.names(df.trunc)`, 1,6)

#make a colour code for each species using short species names
groups <- cbind.data.frame(unique(dfsmoo.pca.scores$spec),
                           seq(1, length(unique(dfsmoo.pca.scores$spec)), by = 1))
colnames(groups) <- c("spec", "group")
#join the colour codes to the PCA result file
dfsmoo.pca.scores <- merge(dfsmoo.pca.scores, groups, by = "spec")

col <- brewer.pal(5, "Dark2")

dfsmoo.pca.scores$group <- as.factor(dfsmoo.pca.scores$group)

par(xpd = FALSE, mfrow = c(1,1), mar = c(5, 5, 5, 7), cex = 0.5, adj = 0.5, tck = 0.01)


plot(dfsmoo.pca.scores$PC1, dfsmoo.pca.scores$PC2, group = dfsmoo.pca.scores$groups,
    col = c("#1B9E77","#D95F02", "#7570B3", "#E7298A",
"#66A61E")[as.factor(dfsmoo.pca.scores$group)],
    pch = 19, cex = 1.5, asp = 1, cex.axis = 1.5, xlab = "PC1 (74%)", ylab = "PC2 (18%)",
cex.lab = 1.5)
abline(h = 0, col = "grey")
abline(v = 0, col = "grey")


# Second derivative plot only
dfsmoo.pca.scores$spec <- as.character(dfsmoo.pca.scores$spec)
leg.txt <- unique(dfsmoo.pca.scores$spec)

legend("right", inset = c(-0.15, 0), leg.txt, pch = 19, cex = 1.5,
       col = col, xpd = TRUE, bty = "n")


```

```
##### Loading Plots - Run for each PCA plot #####
loadings <- as.data.frame(dfsmoo.pca$rotation)[1:2]

scale <- min(max(abs(dfsmoo.pca.scores$PC1))/max(abs(loadings$PC1)),
        max(abs(dfsmoo.pca.scores$PC2))/max(abs(loadings$PC2))) * 0.8


#extract the wavenumbers as numbers from rotation
wavenumbers <- as.numeric(rownames(dfsmoo.pca$rotation))

#extracts the first column (PCA1). Change [,1] to [,2] for PCA2 etc.
PC1loading <- as.data.frame(loadings[,1])
PC2loading <- as.data.frame(loadings[,2])

#writes the wavenumbers to the PCA1loadings object
PC1loading$wavenumber <- wavenumbers
PC2loading$wavenumber <- wavenumbers

colnames(PC1loading) <- c("loading", "wavenumber")
colnames(PC2loading) <- c("loading", "wavenumber")


#switch PC1loadings to PC2 for other plot
plot(PC1loading$loading ~ PC1loading$wavenumber, type = "l",
    xlim = c(1800,600), xlab = "Wavenumber", ylab = "PC1 Loadings", cex.axis = 1.5,
    cex.lab = 1.5)
#2nd deriv line
abline(h = 0, col = "black")

# Write files for loadings

write.csv(PC1loading, "PC1Loading.csv")
write.csv(PC2loading, "PC2Loading.csv")

```

```
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
```

```
# ITOL file
my_tree <- as.phylo(cluster)

write.tree(phy = my_tree, file = "Treefinal.diff.newick")
```

```
#######################################
## Scoble, L (2023)                   ##
## Decision Trees and randomForest    ##
#######################################

##### PART 1 - Decision trees: Extracting rpart rules which show which wavenumbers
# are causing discrepancies between species - then compare to PCA loading plots #####

# Read in file
d <- read.table("all.data.emsc.baseline.final.csv", sep = ",", header = T, row.names = 1)
d <- data.frame(t(d))
names(d)<-sapply(str_remove_all(colnames(d),"X"),"[")

# Truncate spectra
d <- d[,1141:ncol(d)]

d <- d[1:253,]
str(d)
d <- cbind(rownames(d), data.frame(d, row.names = NULL))


# Make column with short specie names
Species <- c(rep("Agros", 51),
             rep("Antho", 51),
             rep("Desch", 51),
             rep("Festu", 50),
             rep("Molin", 50))

# Factorise
Species <- as.factor(Species)

# Bind the two together
d <- cbind(Species, d)

```

```
# Remove the sample labels
d <- d[,-2]


summary(d$Species)
set.seed(2)

##### First decision (classification) tree using all data #####
fit <- rpart(Species ~., data = d, method = "class")
par(mar = c(2, 4, 4, 4))
par(mfrow = c(1,1))

# Plot classification tree
plot(fit)
text(fit, cex = 0.9, xpd = TRUE)

# Use rplot for more better visuals (legend position may need to be changed)
rplot <- rpart.plot(fit, type = 4, extra = "auto", clip.right.labs = FALSE,
                    legend.x = 0.85, legend.y = 1, legend.cex = 1.3,
                    cex = 0.8)


# Extract the rules that the algorithm uses to build tree and splits
# This is to look at what wavenumbers are driving the discrepancy between
# species

# Digits = 3 to get an extra decimal place (easier to refer to the data)
rpart.rules(fit)
rules <- rpart.rules(fit, digit = 3)

# Remove columns that aren't relevant
rules <- rules[,-2]
rules <- rules[,-2]
rules <- rules[,-5]
```

```
rules <- rules[,-8]
rules <- rules[,-11]

# Change colnames (Less than, Equal to, Greater than (L/E/G), Absorbance units (Au))
colnames(rules) <- c("Species", "Wavenumber1", "L/E/G", "Au", "Wavenumber2",
                     "L/E/G", "Au", "wavenumber3", "L/E/G", "Au",
                     "wavenumber4", "L/E/G", "Au")

# Write csv
write.csv(rules, "rpart.wavenumber.rules.final.csv")

# Find wavenumbers in original dataset to cross check rules
WN1 <- d %>% dplyr::select(X1693.4306)
WN2 <- d %>% dplyr::select(X883.36129)
WN3 <- d %>% dplyr::select(X1745.50649)
WN4 <- d %>% dplyr::select(X1151.45566)

cross_check <- cbind(Species, WN1, WN2, WN3, WN4)

colnames(cross_check) <- gsub("X","",colnames(cross_check[,1:5]))

write.csv(cross_check, "Cross_check_wavenumbers.final.csv")

##### Looped Variance #####

##### Split the data and run decision tree 100 times in a loop #####
# Will the same four wavenumbers still be prominent or will splitting the data
# create more variance.
set.seed(2)
tree_lengths <- data.frame()

for(i in 1:100) {
  train <- sample(nrow(d), 0.8*nrow(d))
  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",
```

```
                    "L/E/G", "Au", "wavenumber3", "L/E/G", "Au",
                    "wavenumber4", "L/E/G", "Au" )
# What does the new set of rules for split data show compared to the previous?
write.csv(rules_one, "rpart_wavenumbers_rules_loop.final.csv")

WN5 <- d %>% dplyr::select(X1693.4306)
WN6 <- d %>% dplyr::select(X883.36129)
WN7 <- d %>% dplyr::select(X1745.50649)
WN8 <- d %>% dplyr::select(X1155.31313)

cross_check1 <- cbind(Species, WN5, WN6, WN7, WN8)


colnames(cross_check1) <- gsub("X","",colnames(cross_check1[,1:5]))

write.csv(cross_check, "Cross_check_wavenumbers_loop.csv")

# Clean up the tree_lengths data frame to only have wave numbers present

tree_lengths <- tree_lengths[,-9]
tree_lengths <- tree_lengths[,-8]

tree_lengths <- as.data.frame(apply(tree_lengths, 2, function(x) {
  x <- gsub("X", "", x)
}))
tree_lengths <- as.data.frame(apply(tree_lengths, 2, function(x) {
  x <- gsub("<leaf>", "0", x)
}))

# Convert to num
tree_lengths <- type.convert(tree_lengths, as.is = TRUE)


tree_lengths2 <- melt(tree_lengths, id.vars = c("V1", "V2", "V3", "V4", "V5", "V6", "V7"))
```

```
# Place all wavenumbers into one column
tree_lengths2 <- reshape(tree_lengths, direction = "long", sep = "", varying = 1:7)

# Remove time column
tree_lengths2 <- tree_lengths2[,-1]
table <- table(tree_lengths2$V)
table <- as.data.frame(table)

# Remove zero (first row) as not relevant
table <- table[-1,]

# Arrange table so Freq is descending from largest to smallest
table2 <- table %>%
  arrange(desc(Freq))

# What is table showing and how does that compare to fit and also the PCA loadings

# Plot histogram
table2 <- table2[1:10,]
table3 <- as.data.frame(table2)

par(mfrow = c(1,1))
par(mar = c(2, 4, 4, 4))
ggplot(table3, aes(x = reorder(Var1, -Freq), y = Freq, fill = rules)) +
  geom_histogram(stat = "Identity", colour = "darkblue", fill = "lightblue") +
  labs(x = "Wavenumber", y = "Frequency of Appearence") +
  theme(panel.grid = element_blank(), strip.text.y = element_blank(),
        axis.text.x = element_text(angle = 50, vjust = 1, hjust = 1, size = 11, face = "bold",
        colour = "black"), axis.title.x = element_text(size = 15), axis.title.y = element_text(size
= 14),
        axis.text.y = element_text(size = 11, face = "bold", colour = "black"),
        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$wavenumber`, -n), y = n, fill = rules)) +
1787    geom_histogram(stat = "Identity", colour = "darkblue", fill = "lightblue") +
1788    labs(x = "Wavenumber", y = "Frequency of Appearence") +
1789    theme(panel.grid = element_blank(), strip.text.y = element_blank(),
1790          axis.text.x = element_text(angle = 50, vjust = 1, hjust = 1, size = 11, face = "bold",
1791                                     colour = "black"), axis.title.x = element_text(size = 15), axis.title.y =
1792  element_text(size = 14),
1793          axis.text.y = element_text(size = 11, face = "bold", colour = "black"),
1794          panel.background = element_blank())
1795
1796
1797  #  Write csv for table
1798  write.csv(tree_lengths2,"first.wavenumber.rule.split.csv")
1799
1800  ##### Part 2 - RandomForest ######################
1801  ## Classification using RandomForest                    ##
1802  ## Build model using randomForest and training data ##
1803  ################################################
1804
1805  # Bagged trees
1806  set.seed(2)
1807  train <- sample(nrow(d), 0.8*nrow(d))
1808  training_data <- d[train,]
1809  dim(training_data)
1810  summary(training_data$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)
```

```
bag.RF
#Look at error matrix
plot(bag.RF)
print(bag.RF)

#Predict to see if trained forest will accurately predict test data
bag.tree <- predict(bag.RF, testing_data, type = "class")
tab <- table(bag.tree, testing_data$Species)
tab

write.csv(tab, "prediction.RF.final.csv")
(tab[1,5] + tab[5,1] / sum(tab))

#Plot the Variable importance
par(mfrow = c(1,1), mar = c(2,2,1,2))
varImpPlot(bag.RF,
        n.var = 24,
        type = 1,
        sort = TRUE,
        main = "Variable Importance Plot")

##### Looped randomForest for MDA investigations #####

set.seed(2)
# Make an empty list of 10
ls <- list()
n = 10
datalist = list()
# Pre-allocate for slightly more efficiency
datalist = vector("list", length = n)
```

```
# Run loop
for(i in 1:10) {

  importance.tree <- randomForest(Species ~ ., d, ntree = 150, mtry = 24, importance =
TRUE)
  plot(importance.tree)
  wavesum <- importance.tree$importance[,6, drop = FALSE]
  datalist[[i]] <- cbind(rownames(wavesum), data.frame(wavesum, row.names = NULL))
  colnames(datalist[[i]]) <- c("Wavenumber", "MeanDecreaseAccuracy")


    for (i in 1:length(datalist)) {
    assign(paste0("datalist", i), as.data.frame(datalist[[i]]))}
}

#repeat for each datalist
datalist1 <- datalist1 %>%
  arrange(desc(MeanDecreaseAccuracy))


# Cbind all dataframes together
dataframeall <- cbind.data.frame(datalist1, datalist2, datalist3, datalist4,
                                 datalist5, datalist6, datalist7, datalist8,
                                 datalist9, datalist10)
# Convert to numeric
dataframeall <- type.convert(dataframeall, as.is = TRUE)

# Trim rows to only have the top 24 (24 is the square root of 622)
dataframeall <- dataframeall[1:24,]

# Rename column names
colnames(dataframeall) <- c("V1", "V2", "V3", "V4", "V5", "V6", "V7", "V8", "V9", "V10",
                            "V11", "V12", "V13", "V14", "V15", "V16", "V17", "V18", "V19", "V20")

```

```
# Split into Wavenumber and MDA
dataframewavenumber <- data.frame(dataframeall[, c(1, 3, 5, 7, 9, 11, 13, 15, 17, 19)])
dataframeMDA <- data.frame(dataframeall[, c(2, 4, 6, 8, 10, 12, 14, 16, 18, 20)])

# Rename column names
colnames(dataframewavenumber) <- c("V1", "V2", "V3", "V4", "V5", "V6", "V7", "V8",
"V9", "V10")

# Place all wavenumbers into one column
dataframewavenumber1 <- melt(dataframewavenumber, id.vars = c("V1", "V2", "V3", "V4",
"V5",
                                                    "V6", "V7", "V8", "V9", "V10"))

dataframewavenumber1  <- reshape(dataframewavenumber1 , direction = "long",
                            sep = "", varying = 1:10)

# Rename column names
colnames(dataframeMDA) <- c("V1", "V2", "V3", "V4", "V5", "V6", "V7", "V8",
                        "V9", "V10")

# Place all MDA into one column
dataframeMDA1 <- melt(dataframeMDA, id.vars = c("V1", "V2", "V3", "V4",
                                        "V5","V6", "V7", "V8", "V9", "V10"))

dataframewaveMDA1  <- reshape(dataframeMDA1 , direction = "long",
                          sep = "", varying = 1:10)

# Combine the Wavenumber and MDA column from each dataframe
combinedWNMDA <- cbind.data.frame(dataframewavenumber1$V,
dataframewaveMDA1$V)

# Rename
colnames(combinedWNMDA) <- c("Wavenumber", "MDA")

```

```
# Remove "X" character
combinedWNMDA <- as.data.frame(apply(combinedWNMDA, 2, function(x) {
  x <- gsub("X", "", x) }))

# Convert to numeric
combinedWNMDA <- type.convert(combinedWNMDA, as.is = TRUE)


# Arrange in descending order of MDA numbers
combinedWNMDA <- combinedWNMDA %>%
  arrange(desc(MDA))

# Select only top 24 of all dataframes combined
table <- combinedWNMDA[1:24,]
table2 <- as.data.frame(table)
table2 <- table2 %>% arrange(MDA)


par(mar = c(3, 1 , 0 ,1))

# Plot dotcharts
dotchart(table2$MDA, table2$Wavenumber, xlim = range(table2$MDA),
        xlab = "MeanDecreaseAccuracy", mgp=c(2,1,.5), las=1, cex = 0.9)


#### Run Rf using isolated variables #####
table2$Wavenumber
set.seed(2)
isolated <- d %>% dplyr::select(Species, X1691.50187, X1676.07197,
                          X1641.35472, X1467.76844, X1461.98223, X1450.40981,
                          X1134.09703, X1072.37747, X1068.51999,  X866.00267,  X858.28772,
                          X821.64173, X819.71299, X815.85552, X800.42563, X794.63942,
                          X786.92447, X727.13364, X688.55891,  X632.62556, X626.83935,
                          X622.98187,  X613.33819,  X607.55198)
```

```
train <- sample(nrow(isolated), 0.8*nrow(isolated))
training_data1 <- isolated[train,]
dim(training_data1)
summary(training_data1$Species)

testing_data1 <- isolated[-train, ]
dim(testing_data1)
summary(testing_data1$Species)


set.seed(2)
isolated.rf <- randomForest(Species ~ ., data = training_data1, ntree = 100,
                            importance = TRUE, proximity = TRUE, do.trace = TRUE)

isolated.rf
#Look at error matrix
plot(isolated.rf)
print(isolated.rf)

#Predict test data
bag.tree <- predict(isolated.rf, testing_data, type = "class")
tab <- table(bag.tree, testing_data$Species)
tab

write.csv(tab, "prediction.RF.isolated.csv")
(tab[1,5] + tab[5,1] / sum(tab))
```

```
#Plot the Variable importance
par(mar = c(3, 1 , 1 ,1))
varImpPlot(isolated.rf,
        type = 1,
        sort = TRUE,
        main = "Variable Importance Plot",
        cex = 0.75)


# Run with all isolated data
set.seed(2)
isolated.rf1 <- randomForest(Species ~ ., data = isolated, ntree = 100,
                    importance = TRUE, proximity = TRUE, do.trace = TRUE)

isolated.rf1
#Look at error matrix
plot(isolated.rf1)
print(isolated.rf1)

#Plot the Variable importance
par(mar = c(4, 1 , 1 ,1))
varImpPlot(isolated.rf1,
        type = 1,
        sort = TRUE,
        main = "Variable Importance Plot")

par(mar = c(3, 3, 1, 3))
```

```
# Plot final dotchart
# All data
varImpPlot(isolated.rf1,
           type = 1,
           sort = TRUE,
           main = "Variable Importance Plot",
           cex = 0.75,
           mgp=c(2,1,.5))

#See if trained data can predict unlabelled test data
#make copy of testing_data
testing_data1 <- testing_data

# Actual Species names
Species_1 <- testing_data1[1]

#Remove the sample labels
testing_data1 <- testing_data1[,-1]

#Unlabel data
new_data <- data.frame(testing_data1[,-1])

#Predict for accuracy
new_data$predictedlabel <- predict(isolated.rf, new_data)
new_data$predictedlabel
Predicted <- as.character(new_data$predictedlabel)
Actual <- as.character(testing_data1$Species)

#Cbind the predicted labels with the known species labels from test data
new_data1 <- as.data.frame(cbind(Predicted, Species_1))

#View results as csv
write.csv(new_data1, "prediciton_name_data_isolated_final.csv")

```

```
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
```

```
# Create new Species labels
Species <- c(rep("Agros", 1),
             rep("Antho", 1),
             rep("Desch", 1),
             rep("Festu", 1),
             rep("Molin", 1))

# Factorise
Species <- as.factor(Species)

# Bind the two together
impdf <- cbind(Species, impdf)

# Remove the sample labels
impdf <- impdf[,-2]

colnames(impdf) <- (gsub("X","",colnames(impdf)))

#Melt all data together
melt <- melt(impdf)

#Plot boxplot of MDA as x axis
p <- ggplot(melt, aes(factor(variable), value, fill = Species))
p + geom_boxplot() + facet_wrap(~variable, scale="free") +
  theme(axis.text.x  = element_blank())
```

```
#boxplot of wavenumbers in order of MDA dotchart
colnames(impdf)
varimporder <- d %>% dplyr::select(Species, X1641.35472, X786.92447, X622.98187,
                                   X1676.07197, X727.13364, X1450.40981,
                                  X1072.37747, X1691.50187, X1467.76844,
                                   X1134.09703, X794.63942, X688.55891,  X1461.98223,
                                   X800.42563,  X866.00267,  X821.64173,  X819.71299,
                                    X626.83935,  X613.33819,  X815.85552,  X632.62556,
                                   X858.28772,  X1068.51999, X607.55198)

colnames(varimporder) <- (gsub("X","",colnames(varimporder)))

#melt all the data together
melt <- melt(varimporder)
boxplot(melt, value ~ variable)

p <- ggplot(melt, aes(factor(variable), value, fill = Species))
p + geom_boxplot() + facet_wrap(~variable, scale="free") +
  labs(x = "Wavenumber", y = "Relative Intensity") +
theme(axis.text.x  = element_blank())

```

References

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

Supplementary Material SM3

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

SM3.1.1 Agrostis

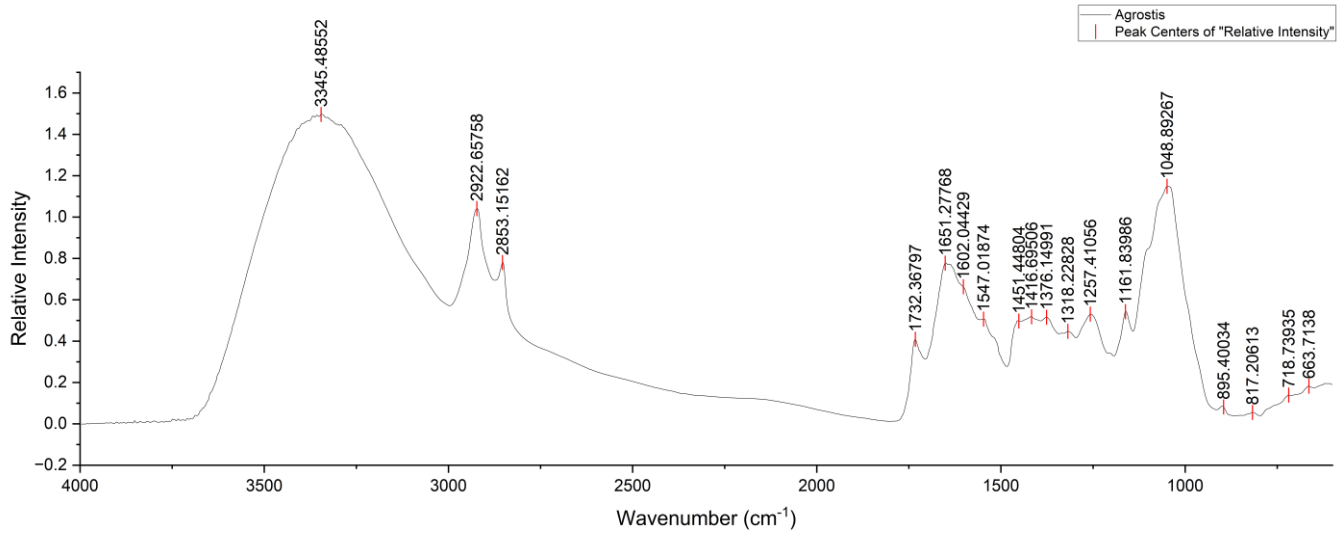


Figure SM3.1.1: Averaged FT-IR spectra of Agrostis with peak numbers included.

3.1.2 Anthoxanthum odoratum

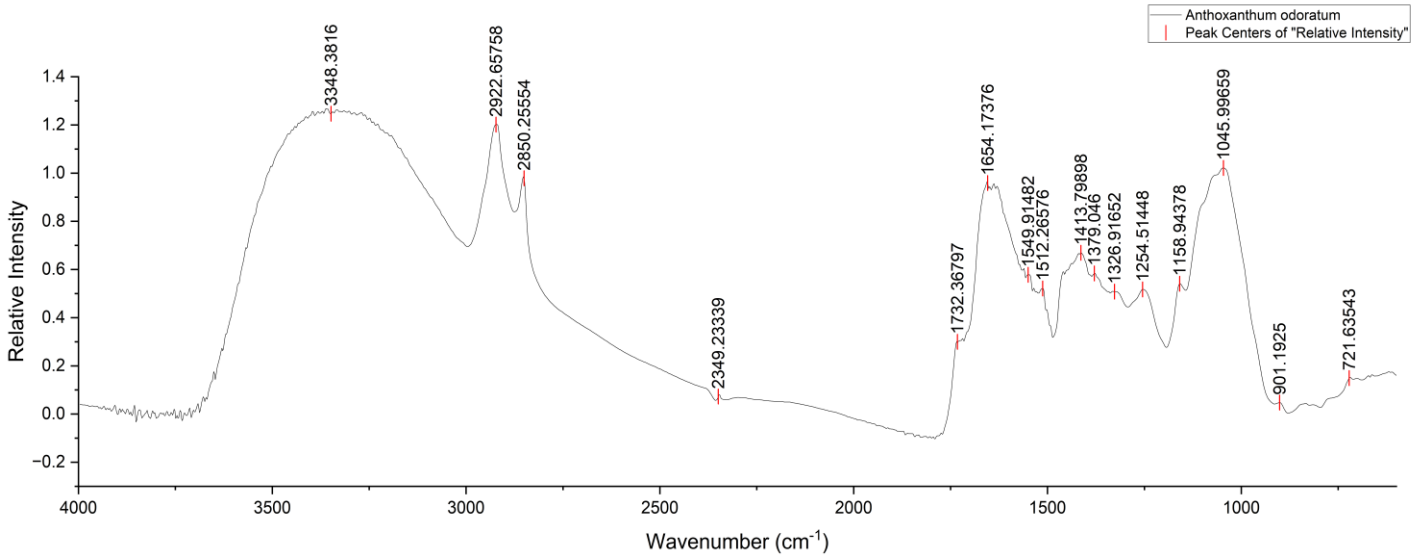


Figure SM3.1.2: Averaged FT-IR spectra of Anthoxanthum odoratum with peak numbers included.

3.1.3 Deschampsia cespitosa

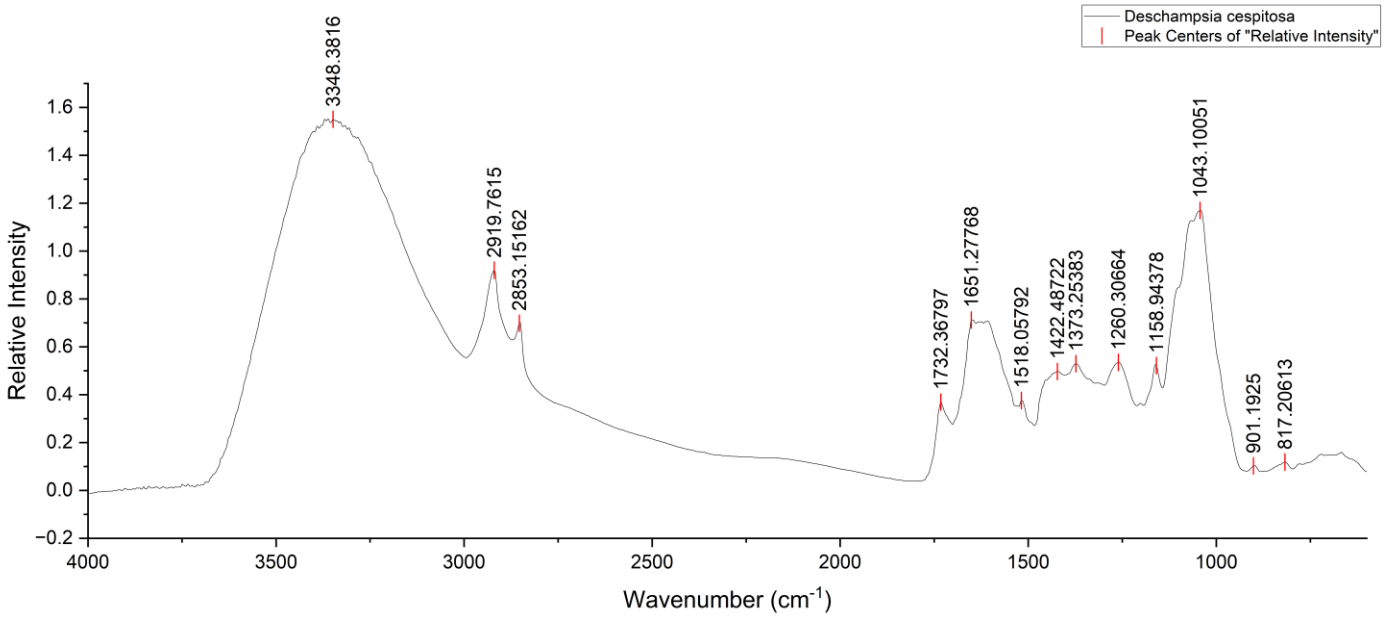


Figure SM3.1.3: Averaged FT-IR spectra of Deschampsia cespitosa with peak numbers included.

3.1.4 Festuca ovina

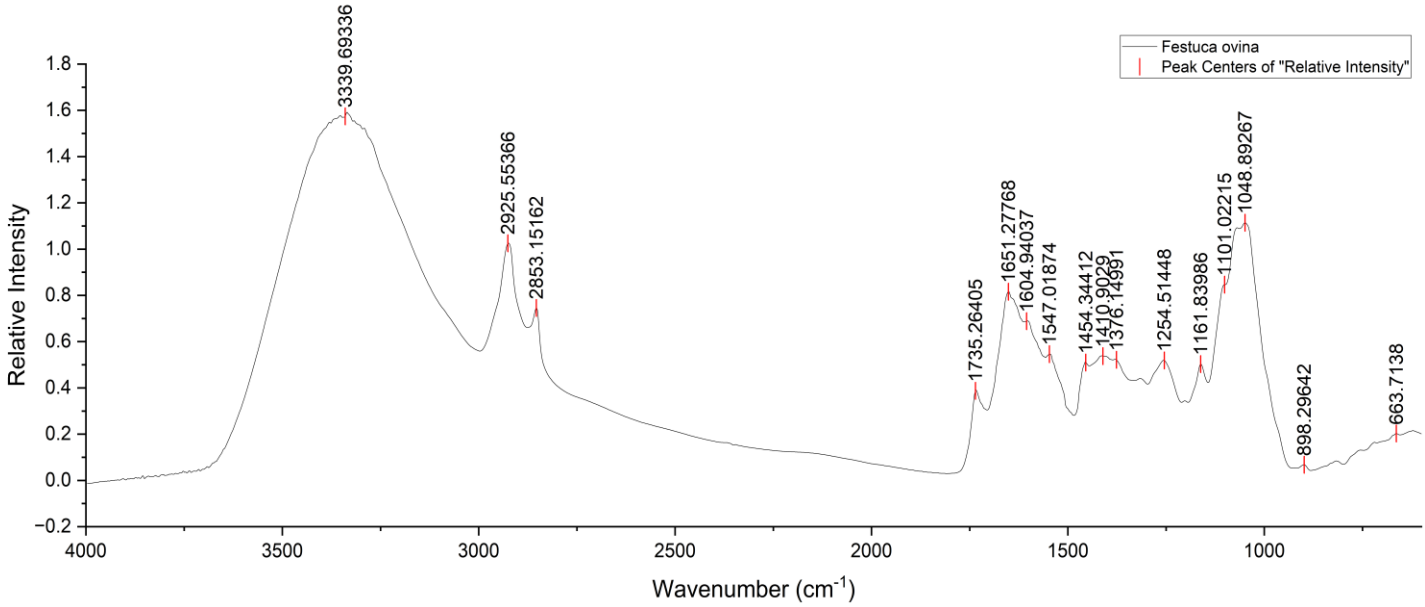


Figure SM3.1.4: Averaged FT-IR spectra of Festuca ovina with peak numbers included.

3.1.5 Molinia caerulea

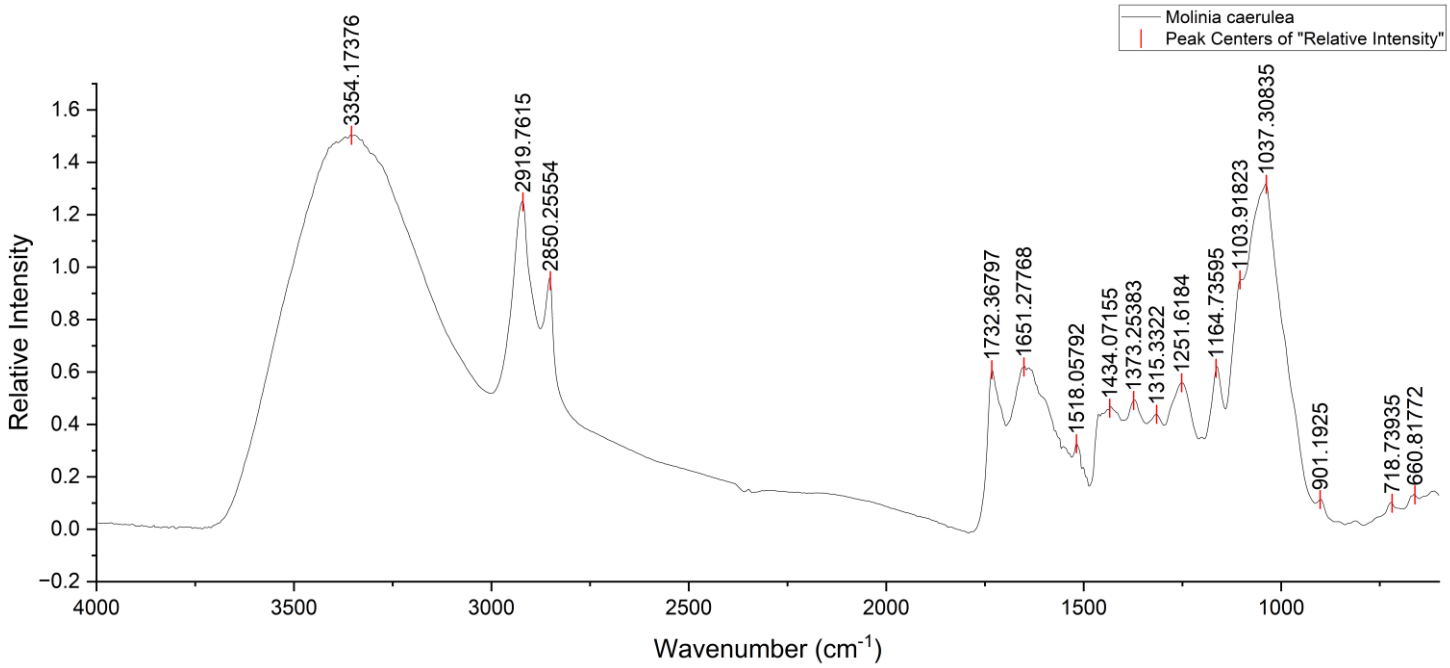


Figure SM3.1.5: Averaged FT-IR spectra of Molinia caerulea with peak numbers included.

SM3.2 Second derivatives of averaged grass spectra

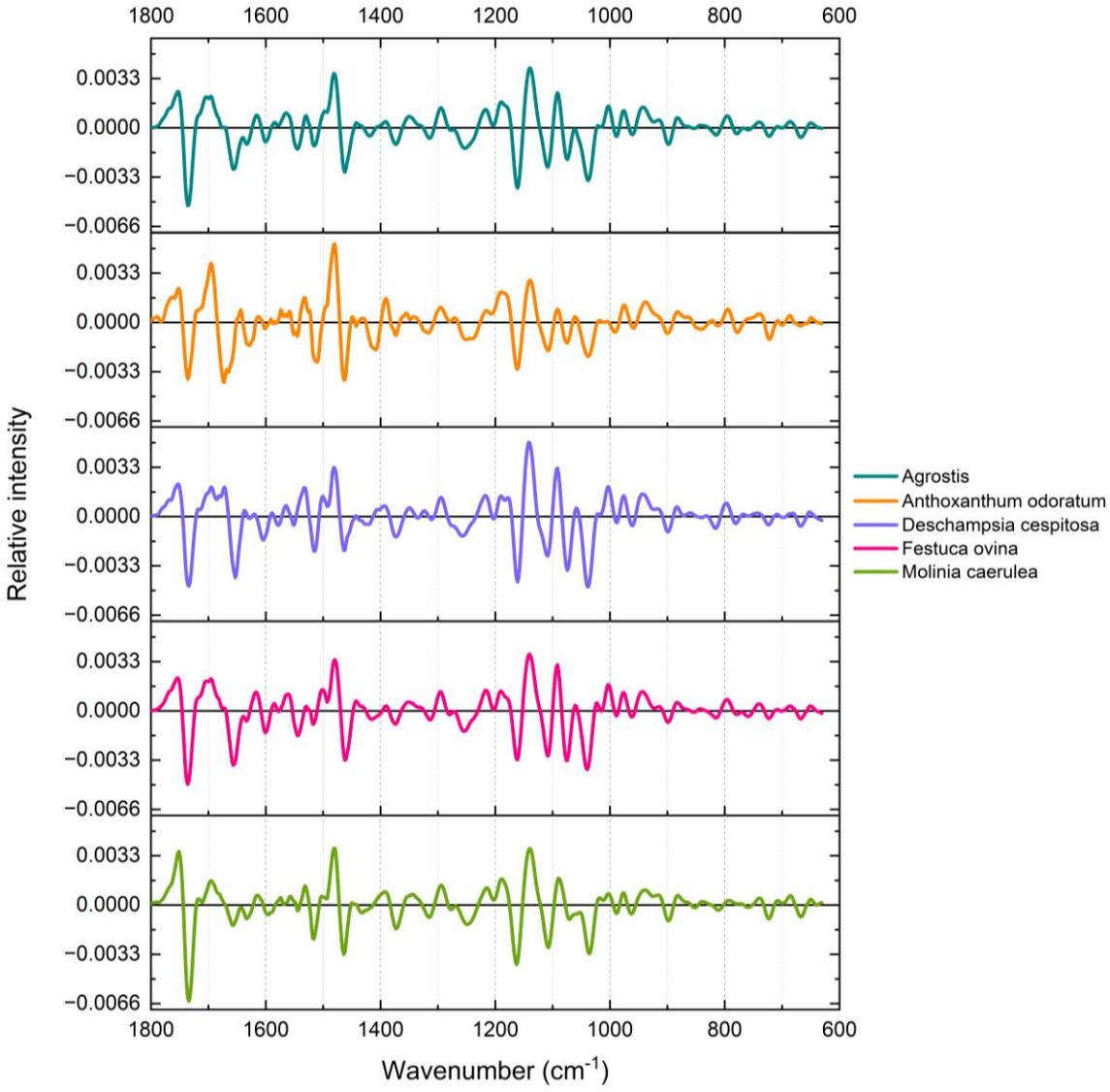


Figure SM3.2: Averaged Savitzky-Golay smoothed, second derivative FT-IR spectra of the five moorland grass species (Agrostis is genus).

SM3.3: HCA of non-differentiated spectra

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

SM3.4: HCA of second derivative spectra

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

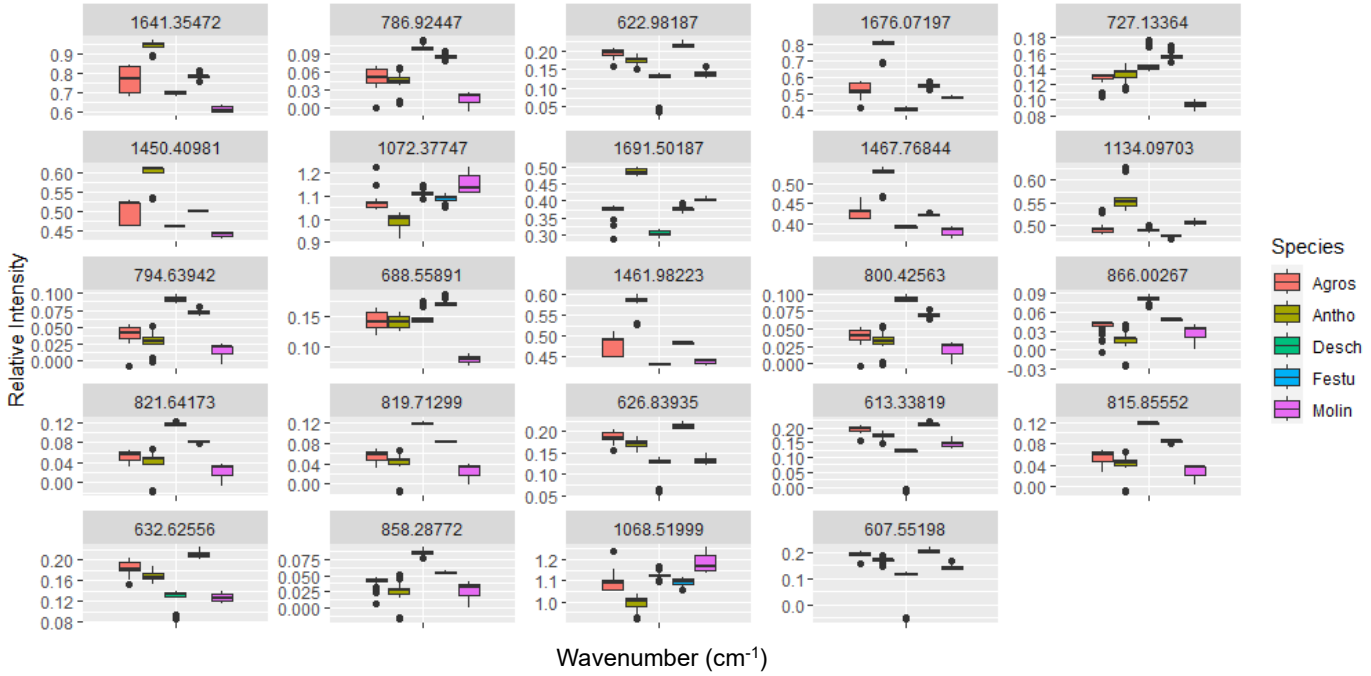
SM3.5: Confusion matrix of bagged randomForest model

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

SM3.6: Boxplots of top 24 wavenumbers identified by randomForest variable importance measure (MeanDecreaseAccuracy)

Figure SM3.6: Boxplots of the top 24 important wavenumbers identified by looped randomForest model. Y- axis is relative intensity, and x – axis is species boxplots (colourcoded), plot visually describes within and between species variation.



SM3.7: Confusion matrix of refined randomForest model

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