# Fourier transform infrared spectroscopy as a tracer of organic matter sources in lake sediments

Charles Maxson[1,2,3*]
John Tibby[1,2]
Jonathan Marshall[4]
Matthew Kent[3]
Jonathan Tyler[2,5]
Cameron Barr[1,2]
Glenn McGregor[4]
Haidee Cadd[5,6,7]
Cameron Schulz[4]
Barry H Lomax[3]

[1] Geography, Environment and Population, University of Adelaide, Adelaide, Australia.
[2] Sprigg Geobiology Centre, University of Adelaide, Adelaide, Australia.
[3] School of Biosciences, University of Nottingham, Sutton Bonington, LE12 5RD, United Kingdom.
[4] Queensland Department of Environment and Science, Brisbane, Queensland, Australia.
[5] Department of Earth Sciences, University of Adelaide, Adelaide, Australia.
[6] ARC Centre of Excellence for Australian Biodiversity and Heritage (CABAH), School of Environment and Atmospheric Sciences, University of Wollongong, Wollongong, Australia.
[7] Chronos $^{14}C$ Carbon-Cycle Facility, UNSW Sydney, Sydney, New South Wales, Australia.

***Corresponding author**
Email: charles.maxsoniv@adelaide.edu.au; ORCID iD: 0000-0003-0092-7802

<u>Co-author emails:</u>
john.tibby@adelaide.edu.au
jonathan.marshall@des.qld.gov.au
matthew.kent@nottingham.ac.uk
jonathan.tyler@adelaide.edu.au
cameron.barr@adelaide.edu.au
glenn.mcgregor@des.qld.gov.au
haidee_cadd@uow.edu.au
cameron.schulz@des.qld.gov.au
barry.lomax@nottingham.ac.uk

Keywords: FTIR, Sediment organic matter sources, Lake sediments, Palaeolimnology, Random Forests

Word count: 5511 (6 figures, 3 tables)

**Abstract**

The source of sedimentary organic matter in lakes can help to elucidate climate and catchment variation and processes that reflect lake development. Common techniques for tracing sediment organic matter sources, such as the stable isotopes and elemental concentrations of C and N, can be too imprecise to identify the specific provenance of organic matter. By contrast, organic geochemical techniques such as gas or liquid chromatography and nuclear magnetic resonance provide detailed organic molecular characterisation but are both expensive and time consuming. Fourier Transform Infrared (FTIR) spectroscopy is a rapid, non-destructive, and well-established method for determining the constituents of lake sediments. However, the potential for identifying the sources of organic matter in lake sediments has not been fully explored. In this study, we assess the extent to which FTIR can be used to identify varying organic matter sources through analysis of modern autotrophs from Blue Lake, North Stradbroke Island, Australia. We investigated spectral processing techniques to identify the approach that could most accurately classify autotroph samples. Three autotroph groups were correctly classified 90% of the time. Processed spectra then became the basis of a model that used multivariate random forests to estimate sediment organic matter composition source from a sediment record from Blue Lake that spans the last 7500 years. FTIR-based estimates suggested that throughout the history of the lake, algae contributed the highest amount of organic matter to the sediment samples. These results allow a refinement of a previous study of C:N and $\delta^{13}C$ from the same core and suggests that alterations in C:N and, particularly, $\delta^{13}C$ reflect chemical changes in algae through time. This study demonstrates that FTIR spectroscopy is a promising tool to elucidate sources of sediment organic matter in lake sediments.

## 1. Introduction

Organic matter is an important fraction of lacustrine sediments (Meyers and Teranes 2002). It is comprised of a complex mixture of organic compounds such as lipids, carbohydrates, and proteins produced by organisms that once lived in, and around, lakes. The accumulation of these compounds records important information about past lake dynamics, catchment processes, and climate (Meyers 1994; Meyers and Ishiwatari 1993; Leng and Marshall 2004; Anderson et al. 2018; Atahan et al. 2015; Cadd et al. 2018; McGowan et al. 2018).

Methods such as the analysis of the isotopic and elemental composition of sediment organic matter are well established and can be powerful tracers of sediment organic matter origin in lacustrine studies (Leng et al. 2006; Leng and Marshall 2004; Maxson et al. in review). However, the interpretation of organic matter sources from these proxies can be ambiguous. Alternative methods that provide chemical data to identify specific sources of organic matter include methods such as fatty acid analysis (Xu et al. 2019), *n*-alkanes (Bush and McInerney 2013; Tao et al. 2016), gas chromatography-mass spectrometry (Bravo et al. 2017; Ninnes et al. 2017), nuclear magnetic resonance (Simpson et al. 2018; Simpson and Hatcher 2004), or ancient DNA (Anderson-Carpenter et al. 2011; Ahmed et al. 2018). However, these methods are often costly, time consuming, and can require large amounts of sediment.

Fourier transform infrared (FTIR) spectroscopy is a widely used method in material and soil science (Allen et al. 1994; Artz et al. 2008; Beć et al. 2020; High and Penkman 2020; Vogel et al. 2008; Rosén et al. 2010) and is a rapid, non-destructive, and cost-effective method for measuring properties of sediment (Korsman et al. 2002). FTIR spectroscopy produces a vibrational spectrum due to the interaction of infrared light wavelengths with chemical bonds which allows for the identification of constituent functional groups in a sample (Colthup 2012; Beć et al. 2020). Information on multiple biological, geochemical, and sedimentological properties can be obtained relatively quickly from a single spectrum, and samples do not require complex preparation. However, the complex mixture of organic and mineralogical constituents of lacustrine sediments can overlap and obscure individual constituent signals. Typically, partial least squares regressions have been used to tie biogeochemical information to FTIR spectral information (Rosén et al. 2010; Vogel et al. 2008; Cadd et al. 2020). This technique has been utilised to estimate the concentration of total organic carbon (TOC), total nitrogen (TN), biogenic silica, and charcoal in lake sediment cores (Meyer-Jacob et al. 2014; Rosén et al. 2010; Vogel et al. 2008; Cadd et al. 2020; Constantine et al. 2021). These studies highlight specific regions of sediment spectra that are tied directly to organic matter (Rosén et al. 2010; Vogel et al. 2008), which could enable the identification of the source of sedimentary organic material. However, the use of PLS methods to measure C:N via FTIR is laborious and time consuming, and FTIR measurements of isotopes requires an electron microscope (Hachtel et al. 2019). Therefore, alternative data processing techniques are needed to enable identification of sediment organic matter sources with FTIR spectra.

Random forests is a machine learning technique for classification and regression that operates by constructing – or 'growing' – multiple decision trees that classify elements based on the aggregate of individual trees (Breiman 2001; Ho 1995, 1998). Random forests can be used on large datasets because random sub-sampling allows trees to grow without suffering from overfitting (Ho 1995; Breiman 2001). Furthermore, random forest models are a particularly useful method to interrogate high dimensional datasets like FTIR spectral data. Random forests have the potential to associate relevant FTIR signatures to contemporary lake organic matter sources and for resulting models to then infer past contributions from these sources using lake sediment FTIR spectra.

Here we explore, for the first time, a method for distinguishing the sources of organic matter in a lacustrine sediment sequence based on FTIR spectroscopic data. We use a dataset of modern autotrophs from Blue Lake, North Stradbroke Island, Australia to train a random forest model. Then, mixtures of contemporary autotrophs are used as an independent test of model predictive ability. Finally, the model is applied to sediment core samples to infer past sediment organic matter sources in Blue Lake. Our results are compared to pre-existing C:N and $\delta^{13}C$ data from the same sediments to explore differences between these proxies (Maxson et al. in review). FTIR data may be able to refine interpretation of the Holocene sedimentary organic matter record from Blue Lake which has been hampered by the ambiguous nature of C:N and $\delta^{13}C$ from wetlands on North Stradbroke Island (Cadd et al. 2018; Maxson et al. in review).

## 2. Methods

### 2.1 Study site

North Stradbroke Island (27°27’12”S, 153°28’56”E; Figure 1) is the second largest sand island on Earth and is part of the massive south-east Queensland coastal dune fields (Patton et al. 2019). Blue Lake is a groundwater “window” lake that intersects with the regional groundwater table within the sand mass of the island (Barr et al. 2013). The connection to the groundwater table has been invoked to explain the inference, derived from sediment proxy analyses, that Blue Lake has changed little since its inception during the Holocene and that it is one of the most stable lake systems in Australia (Barr et al. 2013). This hypothesis has recently been refined by new interpretations of the organic matter record from Blue Lake that have shown that lake nutrient concentrations responded to changes in climate over the past 7,500 years (Maxson et al. in review).

The vegetation of the Blue Lake catchment is comprised of open sclerophyll woodland, dominated by *Eucalyptus* and Casuarinaceae tree species (Moss et al. 2013), and includes *Banksia*, *Melaleuca*, and various heath species as understorey vegetation (Barr et al. 2013). Aquatic vegetation (collectively referred to as ‘aquatic macrophytes’) is dominated by the emergent macrophyte *Lepironia articulata* (Retz.) Domin and the submerged *Eleocharis difformis* S.T.Blake, with small populations of

*Cycnogeton* procerus (R.Br.) Mering & Kadereit, *Gahnia* sp., and *Myriophyllum* sp.. Cyanobacteria and diatoms (collectively referred to as 'algae') primarily grow epiphytically on aquatic macrophytes and epiphytically on fallen trees in the lake. Epiphytic algae can become detached and form macroscopic 'balls' which are found on the lake floor. Generally, these algal 'balls' are found in the centre of the lake, far from any aquatic macrophyte stand or bed (Maxson et al. in review).

## 2.2 Autotroph collection and classification

Autotroph samples were collected at sites every 200 metres around the perimeter of the lake (Maxson et al. in review). Samples include terrestrial leaves and wood (*Banksia* sp., Casuarinaceae, and *Eucalyptus* sp.), aquatic macrophytes (*Lepironia articulata*, *Eleocharis difformis, Cycnogeton procerus, Myriophyllum* sp., and *Gahnia* sp.), and algae dominated by cyanobacteria (*Hapalosiphon pumilus* Kirchner ex Bornet & Flahault, *Symphyonema karboorum* G.B.McGregor, and *Scytonema mirabile* Bornet). Samples were separated for identification and for FTIR analysis. For identification, algal material was preserved *in situ* with 2% buffered formaldehyde. Sub-samples were mounted on flat slides and observed at 400× magnification. A total of 300 algal units were identified from each sample to the lowest taxonomic rank and results expressed as proportional abundance (Barbour et al. 1999). For FTIR determination, fresh samples were freeze dried and ground into a fine powder using a mortar and pestle for analysis. Additional samples included woody debris of unknown origin collected from the shallow, near shore environments of the lake.

Autotroph data were categorised at several hierarchical levels to test whether FTIR data could successfully identify each different level, namely: individual sample, genera, and group (Table 1). The autotroph groups were algae, aquatic macrophytes, and terrestrial plants. Approximately half of the samples were not identified to species level (Table 1), and therefore groupings by species were not used. Terrestrial samples were split into 'leaf' and 'stick' samples to identify differences in the sample types.

## 2.3 Mixtures of autotroph groups

To test if the composition of organic matter could be accurately determined from FTIR spectra, we created mixes of autotroph material. For each autotroph group (algae, aquatic macrophyte, and terrestrial; Table 1), a one-gram representative sample was made by mixing equal parts of all the individual samples of that group. For example, in the algae group, which contained nine individual samples, 1/9 g of each individual algae sample was combined to create the representative algae sample.

The representative autotroph group samples were then used to create amalgamations with known concentrations of each group. Three types of mixtures were created, one with 50% dominance, one with 80% dominance, and one with equal weighting (33% by weight of each group). This was repeated for each of the three groups, resulting in seven unique sample mixtures (Table 2).

## 2.4 Sediment core samples

Sediments were sub-sampled from a 2.4 metre core from Blue Lake that has previously been analysed for C:N and $\delta^{13}C$ (Barr et al. 2013; Maxson et al. in review) (Table 3). The core had largely homogenous sediments with no distinguishing features (Barr et al. 2013; Maxson et al. in review). Seven samples were chosen to represent points of major changes in the C:N and $\delta^{13}C$ of sediment organic matter. Two additional surface samples were collected near the core site (Surface 1) and 200 metres north of the core site (Surface 2) and were also analysed via FTIR (Table 3) (Maxson et al. in review). Surface samples were not analysed for C:N and $\delta^{13}C$. All samples were freeze dried and ground to a fine powder in a mortar and pestle prior to FTIR analysis.

## 2.5 FTIR analysis and data processing

### *2.5.1 Sample analysis*

For each sample (autotroph, mixture, and sediment), a small amount of material (<1 mg) was suspended in acetone to allow for easy mounting. The suspension was pipetted onto a 0.5 mm-thick calcium fluoride disc allowing time for the acetone to evaporate prior to analysis. IR spectra were generated using a Cary 670 FTIR spectrometer integrated with a Cary 620 FTIR microscope (Agilent, Santa Clara, CA, USA). The FTIR microscope was fitted with a 15× Vis/IR objective and a 64 × 64-pixel, liquid nitrogen cooled focal plane array detector purged with dry air. Each pixel constituted an area of 5.5 $\mu m^2$, providing a total scanning area of 352 $\mu m^2$. Scanning was conducted in transmission mode at a resolution of 4 $cm^{-1}$ over a spectral range of 4000 to 950 $cm^{-1}$. Three replicate measurements per sample disc were collected. Replicates were determined by averaging the 128 scans collected per measurement with the Resolutions Pro software. Background spectra were collected and automatically subtracted prior to each set of replicates.

### *2.5.2 Autotroph spectra processing and classification*

Data processing was based on Jardine et al. (2019a), with publicly available code (Jardine et al. 2019b). Data processing was carried out in R v.3.6.3 (R Core Team 2020) using the packages baseline v.1.2-3 (Liland et al. 2010), caret v.6.0-86 (Kuhn et al. 2016), class v.7.3-17 (Venables and Ripley 2002), corrplot v.0.84 (Wei and Simko 2017), e1071 v.1.7-3 (Meyer et al. 2014), prospectr v.0.2.0 (Stevens et al. 2020), MASS v. 7.3-51.6 (Venables and Ripley 2002), and RColorBrewer v1.1-2 (Neuwirth and Brewer 2014). Spectra were z-score standardized to correct for differences in absorbance intensity by subtracting the mean value and dividing by the standard deviation (Duarte et al. 2004; Jardine et al. 2015). Baseline drift was reduced by subtracting a second-order polynomial baseline from each spectrum (Jardine et al. 2019a). The z-score standardized, baseline corrected data are referred to as 'unprocessed spectra'. Replicates of individual samples were split into training (2/3) and validation (1/3) datasets to test the classification potential of the data.

We used *k* nearest neighbour (*k*-nn) classification to identify groups of samples based on the similarity of spectra (Jardine et al. 2019a; Julier et al. 2016; Varmuza and Filzmoser 2016). The parameter *k* is the number of most similar spectra, or 'nearest neighbours', used for classification and is manually selected. When $k = 1$, the object is assigned to the class of its nearest neighbour. When $k > 1$ the classification is determined by the average of the selected number of nearest neighbours. Classification in the case of a tie is determined randomly (Venables and Ripley 2002). Euclidean distance was used to determine between-sample spectral similarity.

In addition to *k*-nn classification, data processing consisted of generating first and second derivatives of spectral data and applying Savitzky-Golay smoothing (Zimmermann and Kohler 2013) to find which data derivative had the highest classification potential. Data derivatives are used because they can amplify the signals in spectra to better differentiate samples. However, data derivatives generally amplify random noise in the data (Tsai and Philpot 1998). Savitzky-Golay smoothing is a technique that applies an algorithm which approximates the spectra by polynomial least-square fitting to a moving window of size $\omega$. To smooth the random noise in the data derivatives, Savitzky-Golay smoothing was applied to optimise the signal to noise ratio and smoothed data are referred to as 'processed spectra'.

Leave-one-out cross-validation (LOOCV) was conducted on the training dataset of unprocessed, first, and second derivative data to determine the best combination of parameters ($k$ and $\omega$) for classification (Jardine et al. 2019a). LOOCV removes one sample from a dataset and runs a model on the remaining datapoints ($n$-1). This process is repeated $n$ times and classification accuracy can then be calculated based on the percent correct classification across the LOOCV procedure. A range of window sizes ($\omega = 5$ to $\omega = 43$) and nearest neighbours ($k = 1$ to $k = 20$) were tested with LOOCV. The Savitzky-Golay polynomial order ($p$) was fixed at three for simplicity, because $p$ must be larger than the order of the derived data (two being the highest in this case).

The best performing combination of parameters for $k$ and $\omega$ was applied to the training dataset to create a model. This model was applied to the validation dataset and classification accuracy was calculated from the percentage of correctly predicted classifications divided by total classifications. We used confusion matrices (Stehman 1997) to examine the inter- and intra-taxon patterns of classification accuracy (Jardine et al. 2019a). Confusion matrices compare the value predicted by the LOOCV to the reference. A diagonal line in the confusion matrix shows 100% correct classification, and deviations from the diagonal line can show how samples have been misclassified. Hierarchical cluster analysis with the Euclidean distance metric and average linkage method as well as principal components analysis (PCA) were used to visualise relationships of data clusters.

## 2.6 Estimating organic matter origin in sediment samples

Autotroph data were used to predict sediment organic matter composition using multivariate random forest analysis. The R (R Core Team 2020) packages 'vegan' v.2.5-6 (Oksanen et al. 2019), ‘randomForestSRC’ v.2.9.3 (Ishwaran et al. 2007) and ‘randomForestExplainer’ v.0.10.1 (Paluszynska et al. 2020) were used. To compare autotroph and sediment samples, spectral data were truncated to 1800 – 1200 $cm^{-1}$, the spectral region that contains the majority of sediment organic matter compounds (Rosén et al. 2010; Vogel et al. 2008). A PCA was run to determine if plant and sediment data could be compared. Methods and results may be found in the Supplementary Material (Figure S7). Random forests operate by bootstrapping parameters from a dataset to build prediction trees, with approximately one-third of the original dataset omitted from the bootstrap to build each individual tree (Breiman 2001; Biau and Scornet 2016). Each tree becomes a ‘test set’ of the original data. Aggregating trees against the unused data from the bootstrap gives an estimate of model error (out-of-bag error) without requiring an independent validation dataset (Biau and Scornet 2016).

Hyperparameters of the random forests model can be tuned against the out-of-bag error to find the optimal values of each. Parameters generally used are *mtry* (number of features scanned before splitting), *ntree* (number of trees to grow), and *node_size* (minimal number of terminal nodes). Using the R package ‘caret’ v6.0-86 (Kuhn et al. 2016; Brownlee 2020), each hyperparameter was tested over a range of values: mtry (1-89; 89 is the number of predictors in the dataset); ntree (100-2000); and node_size (1-15) to find the value of each that produced the highest accuracy in the model (Supplementary Material, Figure S5 and S6). The best model parameters were ntree = 500, node_size = 1, and mtry = 9 with an out-of-bag error of six percent (Supplementary Material, Figure S5 and S6).

Multivariate random forest analysis applied to autotroph samples was used to train a regression model which was subsequently used to estimate sediment organic matter composition. Validation of the model was quantified by the out-of-bag error and tested on the mixtures of known compositions. Predicted composition of sediment data was calculated and plotted against sediment age to compare with previously measured C:N and $\delta^{13}C$ data (Maxson et al. in review).

## 3. Results

3.1 Autotroph classification

Autotroph FTIR spectra show distinct peaks and unique signatures (Figure 2). The k-nearest neighbour classification accuracy of the validation dataset was highest at the group level (96%; Figure 3a) lower at the genus level (82%), and the lowest at the individual sample level (71%) (Supplementary Material, Figure S1 and S3). Parameters with the highest classification accuracy were second derivative smoothing with $k = 1$ and $\omega = 35$. Three groupings in the samples became apparent from the analyses: algae, aquatic macrophytes, and terrestrial plants (Figure 3). Aquatic macrophyte samples (*Myriophyllum, Eleocharis difformis, Gahnia, Lepironia articulata,* and *Cycnogeton procerus*) were the most frequently misclassified, as other macrophytes, (Supplementary Material, Figure S1 and S3) at the individual sample and genus levels compared to the other sample types. The uncertainty of the aquatic macrophyte classification represented a substantial proportion of the uncertainty of all samples. However, only one aquatic macrophyte (*Eleocharis difformis*) sample was misclassified as a non-macrophyte (*Symphyonema karboora*, a cyanobacterium.

Terrestrial samples were only misclassified as other terrestrial samples, except for one Casuarinaceae leaf and one Casuarinaceae stick sample, which were misclassified as *Symphyonema karboorum* (Supplementary Material, Figure S1 and S3). Generally, the leaf samples were misclassified more frequently than the wood samples. Algae (*Hapalosiphon pumilus, Scytonema mirabile*, and *Symphyonema karboorum*) were the best classified of the three groups. There were no misclassifications in the validation dataset at any level. *Symphyonema karboorum* samples were misidentified as other algae in the training datasets (Supplementary Material, Figure S1 and S3).

Processed spectra principal component analyses show distinct clusters (Figure 3c). Algae and terrestrial samples are the two most dispersed groups. Algae have low PC 1 scores, and terrestrial samples have high PC 1 scores, with both groups spread evenly across PC 2. Aquatic macrophytes plot between algae and terrestrial clusters, near the PC space origin. PC 1 and 2 loadings show variation is mostly in the wavenumbers 1800 – 950 $cm^{-1}$ with small loadings in the 3000 - 2800 $cm^{-1}$ in PC 2 (Figure 3d).

Cluster analysis generally agrees with the PCA and with data clustered into three groups: terrestrial, aquatic macrophyte, and algal samples (Figure 3e). The algal cluster contains only algae samples, and the aquatic macrophyte cluster contains one misclassified algal sample. The terrestrial cluster contains five misclassified macrophyte samples and three misclassified algal samples (Figure 3e). The misclassified macrophyte, *Myriophyllum*, and algal samples are anomalous, with no obvious explanation for why they cluster with terrestrial samples. *Lepironia* is cellulose and lignin rich and physiologically robust (Stephens and Sharp 2009), which may explain why it clusters with terrestrial samples, rather than other aquatic macrophytes.

## 3.2 Random forest model performance and estimation

Testing of hyperparameter values led to a range of out-of-bag error between 5 and 15 percent, with an average value of 10 percent. This indicates that model classification accuracy varied between 85 and 95 percent at all values and combinations of the selected model hyperparameters and is interpreted to suggest good model performance.

Mixture composition estimation successfully identified the dominant autotroph in four of six mixes with a dominant autotroph, and poor estimation of the equal mixture (Table 2). In the equal and 50 percent mixes, aquatic macrophytes were consistently overestimated with algae and terrestrial underestimated. In the 80 percent mixes, dominant plant type was identified, and estimated composition was within out-of-bag error (10%) of the actual value (Table 2).

The predicted composition of sediment organic matter sources was calculated (Table 3) and plotted against sediment age (Figure 6a) to compare with previous qualitative estimates of sediment composition (Barr et al. 2013 and Maxson et al. submitted). FTIR-based sediment source composition estimated that algae was the dominant source of organic matter throughout the record, ranging between 74 and 90 percent of the organic matter. One seemingly anomalous sediment sample at ~3.8 ka estimated algae at 20 percent and terrestrial composition at 70 percent. Terrestrial and aquatic macrophyte estimates were generally similar and varied between two  and 19 percent of the sediment organic matter.

## 4. Discussion

### 4.1 FTIR-based identification of autotrophs

FTIR spectra of autotroph samples show distinct peaks that have previously been identified in algae and vascular plants (Cocozza et al. 2003; Filip and Bielek 2002; Grube et al. 2006; Haberhauer et al. 1998; Her et al. 2004; Kansiz et al. 1999; Murdock and Wetzel 2009; Nelson 1991; San-Blas et al. 2011; Traoré et al. 2016; Yang et al. 2007; Zaccheo et al. 2002; Zeroual et al. 1995; Beć et al. 2020; Gonzalez-Torres et al. 2017). These include a broad OH band from 3600 – 3000 $cm^{-1}$, aliphatic compound peaks at 2925 $cm^{-1}$ and 2850 $cm^{-1}$, C=O stretching at 1740 $cm^{-1}$, amide I (C=O stretching) at 1640 $cm^{-1}$, lignin or cellulose (absorbed O-H or conjugated C-O) in the band 1650 – 1600 $cm^{-1}$, amide II (C-N stretching and N-H deformation) at 1550 $cm^{-1}$, and polysaccharides in the 1180 – 950 $cm^{-1}$ region (Figure 2a).

The major differences between autotroph groups are seen in the spectral region at wavenumbers lower than 1800 $cm^{-1}$ (Figure 3d) and will be referred to as the 'total organic carbon (TOC)' region. This region is dense with chemical information typical of organic materials due to the number and variety of bonds to carbon (and nitrogen) located in this part of the spectrum. Clear separation between algae and terrestrial plants can be seen in the polysaccharide, lignin, and amide I regions of the spectra (regions 7, 5, and 4 respectively in Figure 2a). The differences in the polysaccharide region between algae and terrestrial plants suggest subtle differences in the polysaccharides of each. Namely, vascular plants have a higher relative abundance of structural polysaccharides (cellulose or lignin; Figures 2a and 4a) than algae (Meyers 1994; Liu et al. 2020), and algae have a higher relative abundance of storage polysaccharides (generally glycogen) than vascular plants (San-Blas et al. 2011; Kansiz et al. 1999). Specifically, algae from Queensland wallum lakes form a bulky, water stable polysaccharide gel in which the cells live (McGregor 2018). Algae have a higher relative abundance of nitrogen than vascular plants which is in proportion to the high relative abundance of proteins in the algal matrix (Meyers and Teranes 2002; Meyers and Ishiwatari 1993). Aquatic macrophytes have peaks in both the lignin and amide I regions and a strong cellulose signal in the polysaccharide region of the spectra, which is similar to terrestrial plants (Figure 2). This suggests aquatic macrophytes are intermediate in chemical composition between algae and terrestrial plants in terms of polysaccharides and proteins.

The spectral region of 1800 – 1200 $cm^{-1}$ contains most of the information that permits discrimination between groups (Figure 2 and 3). The high classification accuracy using full (96%; Figure 3a, Section 3.1) or truncated spectra (90%; Section 3.2) supports this interpretation. Exclusion of the polysaccharide region (1200 – 1000 $cm^{-1}$) may be the cause of the lower classification accuracy, given the differences between samples in this spectral region, while higher wavenumber peaks (> 1800 $cm^{-1}$) are generally similar between samples (Supplementary Material, Figure S2 and S4; Figure 2). Peaks in amide I and II, lignin, and lipid spectral regions (Figure 2), as well as small differences in amide IV (proteins; 1235 $cm^{-1}$) and aromatic hydrocarbons (amide III proteins; 1460 – 1455 $cm^{-1}$) (Beć et al. 2020), most likely explain differences between autotroph groups in the 1800 – 1200 $cm^{-1}$ region. The marked physiological and ecological differences between sample types, i.e. algae, aquatic macrophyte, and terrestrial (Maxson et al. in review) are a likely driver of the strong model prediction (Figure 3).This suggests our data groups in an intuitive way, and that groupings are identified based on known differences between autotroph types.

Our analyses of controlled mixtures of the different plants and algae show that FTIR has promise for identifying sediment organic matter sources. When the proportions of autotrophs are similar (i.e. the 50% mixes), estimation can be difficult (Table 2). However, when the organic matter compositions are dominated by one autotroph group (i.e. the 80% mixes) model performance improves (Table 2). This is because the mixtures tend to become more 'macrophyte-like' (i.e. intermediate between algae and terrestrial plants) when proportions of autotroph groups are similar. These similarities are most likely explained by the amide I, lignin, and cellulose peaks in the spectra (4, 5, and 7 in Figure 2). Since algal data show no peak in lignin and a peak in amide I, with the reverse occurring in terrestrial plants (Figure 2), mixing algal and terrestrial organic material produces peaks in both. Since macrophytes have peaks in both amide I and lignin, any mixture with similar compositions of the three autotroph groups would appear to be more 'macrophyte-like' than actually the case. Similarly, algae have low cellulose concentrations, terrestrial material have high cellulose concentrations, and macrophytes have cellulose concentrations intermediate between the two (Meyers and Teranes 2002; Liu et al. 2020) so mixing of the three groups with similar concentrations leads to a more 'macrophyte-like' spectral signature. Hence, aquatic macrophyte chemical similarities to both algae and terrestrial plants accounts for their overestimation in all the mixes (Table 2). Caution is therefore needed when interpreting the model results, especially when high concentrations of macrophytes are present. Higher wavenumber inclusion does not change the out-of-bag error of the model, most likely due to similarities of group spectra outside the TOC region (Figure 2). Lower wavenumbers (900 – 600 $cm^{-1}$) indicate that only carbohydrate (Beć et al. 2020;

High and Penkman 2020) or inorganic peaks like carbonate or biogenic silica (Meyer-Jacob et al. 2014) are present, so these wavenumbers may be of little use in differentiating mixtures of autotrophs.

4.2 Sediment organic matter composition in Blue Lake

The FTIR data suggest that Blue Lake fossil sediments are dominated by biogenic silica and TOC (Figure 4). The sediment spectral data are very similar to pure biogenic silica spectra, especially in the 1000–1200 $cm^{-1}$ region (Figure 4). However, a PCA containing both plant and sediment data show they overlap in ordination space (Supplementary material, Figure S7). Furthermore, PC 1 and 2 loadings (Figure 5c,d) show sediment spectra have strong signals in the TOC spectral region (Meyer-Jacob et al. 2014; Rosén et al. 2010; Liu et al. 2013), which indicates the influence of organic matter in these samples. Importantly, biogenic silica and TOC spectra only overlap in the 1200 – 1000 $cm^{-1}$ region (Figure 4) (Meyer-Jacob et al. 2014; Rosén et al. 2010; Vogel et al. 2008), so any signal the 1800 – 1200 $cm^{-1}$ range is likely to derive from organic matter in the sediments rather than linkages to biogenic silica.

When inferring environmental change from sedimentary organic matter it is important to consider diagenesis, because it can alter sediment organic matter composition over time. Diagenetic processes preferentially degrade proteins and amino acids, so tend to degrade algal remains more than terrestrial plants, due to their higher proportion of proteins relative to vascular plants (Li et al. 2013; Meyers and Teranes 2002; Meyers 1994). However, the two most prominent algal species in Blue Lake, *Hapalosiphon pumilus* and *Symphyonema karboorum*, have resistant polysaccharide sheaths that may protect their remains from degradation (McGregor 2018). Structural polysaccharides such as cellulose and lignin can be resistant to degradation (High et al. 2016; High et al. 2013). Reduced absorbance in the TOC region of sediment spectra (Figure 4), relative to the autotroph data (Figure 2), suggests sediment organic matter has degraded in Blue Lake, or that the organic matter signal is dwarfed relative to that from biogenic silica (Figure 4). Loadings of PCs 1 and 2 of the processed sediment spectra (Figure 5c,d) have similar peaks to the loadings of the equivalent PCs of the autotroph processed spectra (Figure 3d), which indicates organic compounds have at least been partially preserved. In summary, these data suggest that diagenetic changes may have altered organic matter in Blue Lake, but the chemical differences between the organic matter sources remain intact and can be used to trace their sources through time.

### 4.3 Application of FTIR to palaeolimnological studies

An important motivation of this study was to further interrogate sources of sedimentary carbon, C:N and $\delta^{13}C$ records due to the ambiguity of sedimentary C:N and $\delta^{13}C$ from Blue Lake (Maxson et al. in review; Barr et al. 2013). This ambiguity characterises oligotrophic lakes on North Stradbroke Island and Fraser Island where nitrogen limitation drives algal C:N values higher (Cadd et al. 2018; Maxson et al. in review), making distinctions between algae and vascular plants using C:N difficult. The FTIR-based method described herein shows promise in this regard, as it was able to distinguish between modern autotroph groups (Figure 3). However, the imprecise estimation of the composition of mixtures of autotrophs highlights some of the limitations of the FTIR method. Namely, the method performs poorly when there is no single dominant source of organic matter, however, performance increases as one source becomes dominant (Table 2).

In the context of the out-of-bag model error (10%) the FTIR based estimates suggest algal dominance throughout the Blue Lake record (Figure 6a) and that algal, aquatic macrophyte, and terrestrial concentrations remained largely stable, with one exception being the estimated high terrestrial concentration at ~3.8 ka. TOC data indicate a prolonged (~200 year) increase in organic matter in this part of the record (Figure 6b). Algal under prediction in the modern group mixes (Table 2) implies that we can have confidence in the organic matter inferences due to their high inferred algal proportions. However, the excursion to terrestrial dominance at ~3.8 ka disagrees with the C:N data, which does not indicate significant amounts of terrestrial material (Maxson et al. in review). Without more FTIR data, it is difficult to confidently confirm whether a temporary alteration in organic source occurred. Further study is needed surrounding the 3.8 ka event.

The inferred high algal contribution to the Blue Lake organic matter record is supported by the higher relative abundance of sediments consisting of diatoms (Figure 4a), the prevalence of algae found in the present-day lake (Maxson et al. in review), and the relative stability of Blue Lake through time (Barr et al. 2013). Moreover, it further suggests that degradation of algal material has not led to algae being underrepresented by FTIR. Indeed, sedimentary C:N and $\delta^{13}C$ data are generally most similar to average modern algal C:N and $\delta^{13}C$ composition (Maxson et al. in review). FTIR results suggest that C:N and $\delta^{13}C$ data reflect changes in algal chemistry through time, rather than indicating different sources of organic matter (Figure 6c,d). The shift in C:N and $\delta^{13}C$ at 4.2 ka, therefore, most likely reflects an increase in lake nutrient concentrations leading to less nitrogen limitation in algae, reducing algal C:N values (Maxson et al. in review). This interpretation supports previous studies from Blue Lake and North Stradbroke Island that indicate a reduction in rainfall (Barr et al. 2019) led to an increase in water residence time that subsequently promoted higher nutrient availability in Blue Lake (Maxson et al. in review).

Interpreting the source of organic matter in lake sediments is valuable to a range of applications in palaeoecology and palaeoclimatology, however determining organic sediment sources can present a significant challenge. This study tested the hypothesis that FTIR spectra can be used to discriminate lake sediment organic matter. Overall, the data presented here support this hypothesis. While there may be issues related to organic matter preservation, complex mixtures of sediment, and organic source chemical similarities, this study suggests that FTIR spectroscopy is a valuable tool for identifying organic matter sources in lacustrine sediments.

## 5. **Conclusions**

We presented the first use of FTIR spectral data to compare modern autotroph material with fossil sediment samples to infer the source of sediment organic matter in lake sediments. We demonstrated that processing FTIR spectra using Savitzky-Golay smoothing on data derivatives can enhance classification accuracy when using *k*-nearest neighbour classification and that random forests can classify autotroph samples at the group level with high accuracy. Testing the model on mixes of autotrophs showed that it can better identify compositions of mixtures with a dominant source, rather than thosewith more similar compositions of multiple sources. Application of the model to fossil sediments indicated that Blue Lake's sediment organic matter has been dominated by algae throughout its lifetime, and C:N and $\delta^{13}C$ shifts most likely reflect chemical changes in algae, possibly associated with changed nutrient availability driven by precipitation. This study demonstrates that FTIR based studies have the potential to more accurately trace sediment organic matter origins than those using C:N and $\delta^{13}C$. FTIR has the potential to be applied to sediment records to trace sources of sediment organic matter particularly where contemporary autotroph data are available. This study highlights the potential of FTIR as a new tool in palaeolimnological studies as an efficient, non-destructive means of identifying specific sources of organic matter that will enable more accurate inferences of lake and catchment evolution, climate changes, and variability.

## Acknowledgements

We acknowledge Minjerribah (North Stradbroke Island) and the surrounding waters as Quandamooka Country and thank the Quandamooka Yoolooburrabee Aboriginal Corporation for permission to undertake research on Country. Financial support for the project was provided by Australian Research Council Discovery Projects DP150103875 and DP190102782 and Australian Research Council Linkage Project LP0990124. We acknowledge and thank Carsten Meyer-Jacob for his contribution of biogenic silica data and Ben Stuart-Koster, Ben Gilby, and Katharine Miller for advice on Random Forest modelling.

## References


Ahmed E, Parducci L, Unneberg P, Ågren R, Schenk F, Rattray JE, Han L, Muschitiello F, Pedersen MW, and Smittenberg RH (2018) Archaeal community changes in Lateglacial lake sediments: Evidence from ancient DNA. Quat Sci Rev 181: 19-29

Allen DT, Palen EJ, Haimov MI, Hering SV, and Young JR (1994) Fourier transform infrared spectroscopy of aerosol collected in a low pressure impactor (LPI/FTIR): method development and field calibration. Aerosol Sci Tech 21: 325-42

Anderson-Carpenter LL, McLachlan JS, Jackson ST, Kuch M, Lumibao CY, and Poinar HN (2011) Ancient DNA from lake sediments: bridging the gap between paleoecology and genetics. BMC Evol Bio 11: 1-15

Anderson NJ, Leng MJ, Osburn CL, Fritz SC, Law AC, and McGowan S (2018) A landscape perspective of Holocene organic carbon cycling in coastal SW Greenland lake-catchments. Quat Sci Rev 202: 98-108

Artz RR, Chapman SJ, Robertson AJ, Potts JM, Laggoun-Défarge F, Gogo S, Comont L, Disnar J-R, and Francez A-J (2008) FTIR spectroscopy can be used as a screening tool for organic matter quality in regenerating cutover peatlands. Soil Biol Biochem 40: 515-27

Atahan P, Heijnis H, Dodson J, Grice K, Le Metayer P, Taffs K, Hembrow S, Woltering M, and Zawadzki A (2015) Pollen, biomarker and stable isotope evidence of late Quaternary environmental change at Lake McKenzie, southeast Queensland. J Paleolimnol 53: 139-56

Barbour MT, Gerritsen J, Snyder BD, and Stribling JB (1999) Rapid bioassessment protocols for use in streams and wadeable rivers: periphyton, benthic macroinvertebrates and fish. US Environmental Protection Agency, Office of Water, Washington DC

Barr C, Tibby J, Leng M, Tyler J, Henderson A, Overpeck J, Simpson G, Cole J, Phipps S, and Marshall J (2019) Holocene El Niño–Southern Oscillation variability reflected in subtropical Australian precipitation. Sci Rep 9: 1-9

Barr C, Tibby J, Marshall J, McGregor G, Moss P, Halverson G, and Fluin J (2013) Combining monitoring, models and palaeolimnology to assess ecosystem response to environmental change at monthly to millennial timescales: the stability of Blue Lake, North Stradbroke Island, Australia. Freshw Biol 58: 1614-30

Beć KB, Grabska J, Bonn GK, Popp M, and Huck CW (2020) Principles and applications of vibrational spectroscopic imaging in plant science: a review. Frontiers Plant Sci 11: 1-27

Biau G, and Scornet E (2016) A random forest guided tour. Test 25: 197-227

Bravo AG, Bouchet S, Tolu J, Björn E, Mateos-Rivera A, and Bertilsson S (2017) Molecular composition of organic matter controls methylmercury formation in boreal lakes. Nat Comm 8: 1-9

Breiman L (2001) Random forests. Mach Learning 45: 5-32

Brownlee J (2020). Tune maching learning algorithms in R (random forest case study). https://machinelearningmastery.com/tune-machine-learning-algorithms-in-r/.

Bush RT, and McInerney FA (2013) Leaf wax *n*-alkane distributions in and across modern plants: implications for paleoecology and chemotaxonomy. Geochim Cosmo Acta 117: 161-79

Cadd HR, Tibby J, Barr C, Tyler J, Unger L, Leng MJ, Marshall JC, McGregor G, Lewis R, and Arnold LJ (2018) Development of a southern hemisphere subtropical wetland (Welsby Lagoon, south-east Queensland, Australia) through the last glacial cycle. Quat Sci Rev 202: 53-65

Cadd HR, Tyler J, Tibby J, Baldock J, Hawke B, Barr C, and Leng MJ (2020) The potential for rapid determination of charcoal from wetland sediments using infrared spectroscopy. Palaeogeogr Palaeoclimatol Palaeoecol 542: 1-13

Cocozza C, D'orazio V, Miano T, and Shotyk W (2003) Characterization of solid and aqueous phases of a peat bog profile using molecular fluorescence spectroscopy, ESR and FT-IR, and comparison with physical properties. Org Geochem 34: 49-60

Colthup N (2012) Introduction to infrared and Raman spectroscopy. Elsevier

Constantine IM, Mooney S, Hibbert B, Marjo C, Bird M, Cohen T, Forbes M, McBeath A, Rich A, and Stride J (2021) Using charcoal, ATR FTIR and chemometrics to model the intensity of pyrolysis: exploratory steps towards characterising fire events. Sci Total Environ: 147052

Duan D, Zhang D, Yang Y, Wang J, Chen Ja, and Ran Y (2017) Source, composition, and environmental implication of neutral carbohydrates in sediment cores of subtropical reservoirs, South China. Biogeosciences 14: 4009-22

Duarte IF, Barros A, Almeida C, Spraul M, and Gil AM (2004) Multivariate analysis of NMR and FTIR data as a potential tool for the quality control of beer. J Agric Food Chem 52: 1031-38

Filip Z, and Bielek P (2002) Susceptibility of humic acids from soils with various contents of metals to microbial utilization and transformation. Biol Fertil Soils 36: 426-33

Gonzalez-Torres A, Rich A, Marjo C, and Henderson R (2017) Evaluation of biochemical algal floc properties using Reflectance Fourier-Transform Infrared Imaging. Algal Res 27: 345-55

Grube M, Lin J-G, Lee P, and Kokorevicha S (2006) Evaluation of sewage sludge-based compost by FT-IR spectroscopy. Geoderma 130: 324-33

Haberhauer G, Rafferty B, Strebl F, and Gerzabek M (1998) Comparison of the composition of forest soil litter derived from three different sites at various decompositional stages using FTIR spectroscopy. Geoderma 83: 331-42

Hachtel JA, Huang J, Popovs I, Jansone-Popova S, Keum JK, Jakowski J, Lovejoy TC, Dellby N, Krivanek OL, and Idrobo JC (2019) Identification of site-specific isotopic labels by vibrational spectroscopy in the electron microscope. Science 363: 525-28

Hatcher PG, Ravin A, Behar F, and Baudin F (2014) Diagenesis of organic matter in a 400 m organic rich sediment core from offshore Namibia using solid state 13C NMR and FTIR. Org Geochem 75: 8-23

Her N, Amy G, Park H-R, and Song M (2004) Characterizing algogenic organic matter (AOM) and evaluating associated NF membrane fouling. Water Res 38: 1427-38

High K, Milner N, Panter I, Demarchi B, and Penkman K (2016) Lessons from Star Carr on the vulnerability of organic archaeological remains to environmental change. Proc Natl Acad Sci 113: 12957-62

High K, and Penkman K (2020) A review of analytical methods for assessing preservation in waterlogged archaeological wood and their application in practice. Herit Sci 8: 1-33

High K, Penkman K, Milner N, and Panter I. 2013. Fading Star: Towards understanding the effects of acidification on organic remains (wood) at Star Carr. In Proceedings of the 12th ICOM-CC-WOAM Conference

Ho TK. 1995. Random decision forests. In Proceedings of 3rd international conference on document analysis and recognition 278-82 1: 278-82

——— (1998) The random subspace method for constructing decision forests. IEEE Transactions on pattern analysis and machine intelligence 20: 832-44

Ishwaran H, Kogalur UB, Blackstone EH, and Lauer MS (2007) Random survival forests for R. R News 7: 25-31

Jardine PE, Fraser WT, Lomax BH, and Gosling WD (2015) The impact of oxidation on spore and pollen chemistry. J Micropaleontol 34: 139-49

Jardine PE, Gosling WD, Lomax BH, Julier A, and Fraser WT (2019a) Chemotaxonomy of domesticated grasses: a pathway to understanding the origins of agriculture. J Micropaleontol: 83-95

——— (2019b). Data and code for "Chemotaxonomy of domesticated grasses: a pathway to understanding the origins of agriculture". https://doi.org/10.6084/m9.figshare.8046395.

Julier AC, Jardine PE, Coe AL, Gosling WD, Lomax BH, and Fraser WT (2016) Chemotaxonomy as a tool for interpreting the cryptic diversity of Poaceae pollen. Rev Palaeobot Palynol 235: 140-47

Kansiz M, Heraud P, Wood B, Burden F, Beardall J, and McNaughton D (1999) Fourier transform infrared microspectroscopy and chemometrics as a tool for the discrimination of cyanobacterial strains. Phytochemistry 52: 407-17

Korsman T, Renberg I, DÅBakk E, and Nilsson MB (2002) Near-infrared spectrometry (NIRS) in palaeolimnology. in Tracking environmental change using lake sediments. Springer, 299-317

Kuhn M, Wing J, Weston S, Williams A, Keefer C, Englehardt A, Cooper T, Mayer Z, Kenkel B, Benesty M, Sescarbeau R, Ziem A, Scrucca L, Tang Y, and Candan C (2016) Classification and Regression Training. R package version 6.0-86.

Leng MJ, Lamb AL, Heaton TH, Marshall JD, Wolfe BB, Jones MD, Holmes JA, and Arrowsmith C (2006) Isotopes in lake sediments. in Melanie J Leng (ed.) Isotopes in Palaeoenvironmental Research. Springer, Dordecht, 147-84

Leng MJ, and Marshall JD (2004) Palaeoclimate interpretation of stable isotope data from lake sediment archives. Quat Sci Rev 23: 811-31

Li H, Minor EC, and Zigah PK (2013) Diagenetic changes in Lake Superior sediments as seen from FTIR and 2D correlation spectroscopy. Org Geochem 58: 125-36

Liland KH, Almøy T, and Mevik B-H (2010) Optimal choice of baseline correction for multivariate calibration of spectra. Appl Spectrosc 64: 1007-16

Liu S, He Z, Tang Z, Liu L, Hou J, Li T, Zhang Y, Shi Q, Giesy JP, and Wu F (2020) Linking the molecular composition of autochthonous dissolved organic matter to source identification for freshwater lake ecosystems by combination of optical spectroscopy and FT-ICR-MS analysis. Sci Total Environ 703: 1-20

Liu X, Colman SM, Brown ET, Minor EC, and Li H (2013) Estimation of carbonate, total organic carbon, and biogenic silica content by FTIR and XRF techniques in lacustrine sediments. J Paleolimnol 50: 387-98

Maxson CR, Tibby J, Barr C, Tyler J, Leng MJ, Lomax BH, Marshall J, McGregor G, Schulz C, and Cadd HR (in review) Ecological response and climate sensitivity of a groundwater fed lake on subtropical North Stradbroke Island, Queensland, Australia over the last 7,500 years

McGowan S, Anderson NJ, Edwards ME, Hopla E, Jones V, Langdon PG, Law A, Solovieva N, Turner S, and van Hardenbroek M (2018) Vegetation transitions drive the autotrophy–heterotrophy balance in Arctic lakes. Limnol Oceanogr Lett 3: 246-55

Mcgregor GB (2018) Freshwater Cyanobacteria of North-Eastern Australia: 3. Nostocales. Phytotaxa 359: 1-166

Meyer-Jacob C, Vogel H, Gebhardt A, Wennrich V, Melles M, and Rosén P (2014) Biogeochemical variability during the past 3.6 million years recorded by FTIR spectroscopy in the sediment record of Lake El'gygytgyn, Far East Russian Arctic. Clim Past 10: 209-20

Meyer D, Dimitriadou E, Hornik K, Weingessel A, Leisch F, Chang C, and Lin C (2014) e1071: Misc functions of the Department of Statistics (e1071), TU Wien. R Package Version 1

Meyers PA (1994) Preservation of elemental and isotopic source identification of sedimentary organic matter. Chem Geol 114: 289-302

Meyers PA, and Ishiwatari R (1993) Lacustrine organic geochemistry—an overview of indicators of organic matter sources and diagenesis in lake sediments. Org Geochem 20: 867-900

Meyers PA, and Teranes JL (2002) Sediment organic matter. in William M Last and John P Smol (eds.) Tracking environmental change using lake sediments. Springer, Dordecht, 239-69

Moss PT, Tibby J, Petherick L, McGowan H, and Barr C (2013) Late Quaternary vegetation history of North Stradbroke Island, Queensland, eastern Australia. Quat Sci Rev 74: 257-72

Murdock JN, and Wetzel DL (2009) FT-IR microspectroscopy enhances biological and ecological analysis of algae. Appl Spectrosc Rev 44: 335-61

Nelson WH (1991) Modern techniques for rapid microbiological analysis. VCH

Neuwirth E, and Brewer RC (2014) ColorBrewer palettes. R Package Version: 1.1-2

Ninnes S, Tolu J, Meyer-Jacob C, Mighall TM, and Bindler R (2017) Investigating molecular changes in organic matter composition in two Holocene lake-sediment records from central Sweden using pyrolysis-GC/MS. J Geophys Res Biogeosci 122: 1423-38

Oksanen J, Blanchet FG, Friendly M, Kindt R, Legendre P, McGlinn D, Minchin PR, O'Hara RB, Simpson GL, Solymos P, Stevens MHH, and ES, and Wagner H (2019) vegan: Community Ecology Package. R package version 2.5-6. https://CRAN.R-project.org/package=vegan
Paluszynska A, Biecek P, and Jiang Y (2020) Explaining and Visualizing Random Forests in Terms of Variable Importance
Patton NR, Ellerton D, and Shulmeister J (2019) High-resolution remapping of the coastal dune fields of south east Queensland, Australia: a morphometric approach. J Maps 15: 578-89
R Core Team (2020) R: A language and environment for statistical computing
Rosén P, Vogel H, Cunningham L, Reuss N, Conley DJ, and Persson P (2010) Fourier transform infrared spectroscopy, a new method for rapid determination of total organic and inorganic carbon and biogenic silica concentration in lake sediments. J Paleolimnol 43: 247-59
San-Blas E, Guerra M, Portillo E, Esteves I, Cubillán N, and Alvarado Y (2011) ATR/FTIR characterization of Steinernema glaseri and Heterorhabditis indica. Vibrat Spectro 57: 220-28
Simpson AJ, Simpson MJ, and Soong R (2018) Environmental nuclear magnetic resonance spectroscopy: an overview and a primer. Analytical Chem 90: 628-39
Simpson MJ, and Hatcher PG (2004) Determination of black carbon in natural organic matter by chemical oxidation and solid-state 13C nuclear magnetic resonance spectroscopy. Org Geochem 35: 923-35
Stehman SV (1997) Selecting and interpreting measures of thematic classification accuracy. Remote Sens Environ 62: 77-89
Stephens K, and Sharp D (2009) The Flora of North Stradbroke Island. State of Queensland, Environmental Protection Agency
Stevens A, Ramirez-Lopez L, Stevens MA, and Rcpp L (2020) Package 'prospectr'. R Package Version
Tao S, Eglinton TI, Montluçon DB, McIntyre C, and Zhao M (2016) Diverse origins and pre-depositional histories of organic matter in contemporary Chinese marginal sea sediments. Geochim Cosmo Acta 191: 70-88
Traoré M, Kaal J, and Cortizas AM (2016) Application of FTIR spectroscopy to the characterization of archeological wood. Spectrochim Acta Part A Mol Biomol Spectrosc 153: 63-70
Tsai F, and Philpot W (1998) Derivative analysis of hyperspectral data. Remote Sens Environ 66: 41-51
Varmuza K, and Filzmoser P (2016) Introduction to multivariate statistical analysis in chemometrics. CRC press, Boca Raton
Venables WN, and Ripley BD (2002) Modern Applied Statistics with S. Springer, New York
Vogel H, Rosén P, Wagner B, Melles M, and Persson P (2008) Fourier transform infrared spectroscopy, a new cost-effective tool for quantitative analysis of biogeochemical properties in long sediment records. J Paleolimnol 40: 689-702
Wei T, and Simko V (2017) "corrplot": Visualization of a Correlation Matrix. R package Version 0.84
Xu X, Li W, Deng H, Fujibayashi M, Nomura M, Nishimura O, and Wang G (2019) Approach deliberation for source identification of sedimentary organic matters via comparing freshwater lakes with multi-ecotypes. Sci Total Environ 649: 327-34
Yang H, Yan R, Chen H, Lee DH, and Zheng C (2007) Characteristics of hemicellulose, cellulose and lignin pyrolysis. Fuel 86: 1781-88
Zaccheo P, Cabassi G, Ricca G, and Crippa L (2002) Decomposition of organic residues in soil: experimental technique and spectroscopic approach. Org Geochem 33: 327-45
Zeroual W, Manfait M, and Choisy C (1995) FT-IR spectroscopy study of perturbations induced by antibiotic on bacteria (Escherichia coli). Pathologie-biologie 43: 300-05
Zimmermann B, and Kohler A (2013) Optimizing Savitzky-Golay parameters for improving spectral resolution and quantification in infrared spectroscopy. Appl Spectrosc 67: 892-902

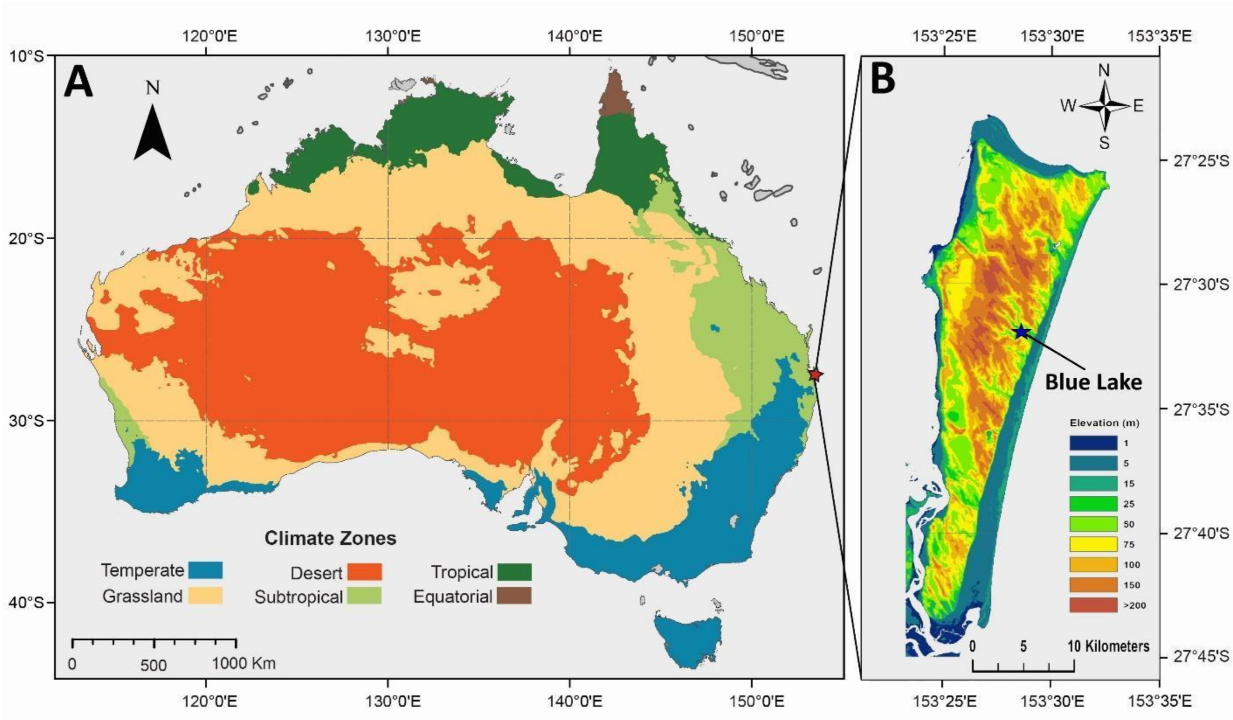


*Figure 1:* (A) Map of Australian climate zones with North Stradbroke Island location. (B) North Stradbroke Island elevation with Blue Lake location.

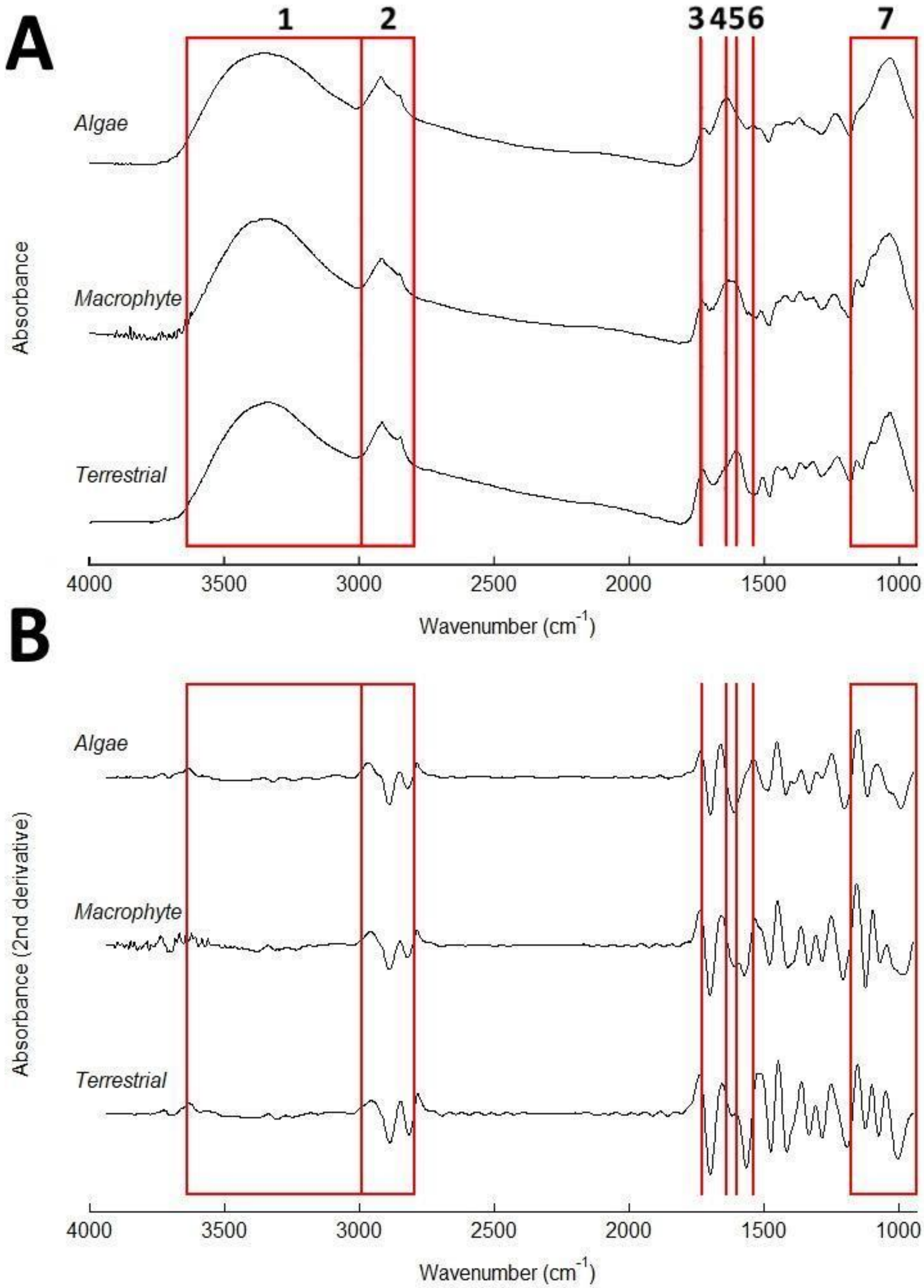


*Figure 2*: A) Unprocessed mean spectra of group data. B) is the smoothed and derived group data. Numbers in A refer to OH zone (1) (Beć et al. 2020), aliphatic compounds (2) (Beć et al. 2020; Zaccheo et al. 2002), lipids or fatty acids (3) (High and Penkman 2020; Haberhauer et al. 1998), amide I (4) (Murdock and Wetzel 2009; Gonzalez-Torres et al. 2017), Lignin (5) (High and Penkman 2020; Traoré et al. 2016), amide II (6) (Beć et al. 2020; Gonzalez-Torres et al. 2017), and polysaccharides (7) (Beć

et al. 2020; Traoré et al. 2016; San-Blas et al. 2011; Kansiz et al. 1999).

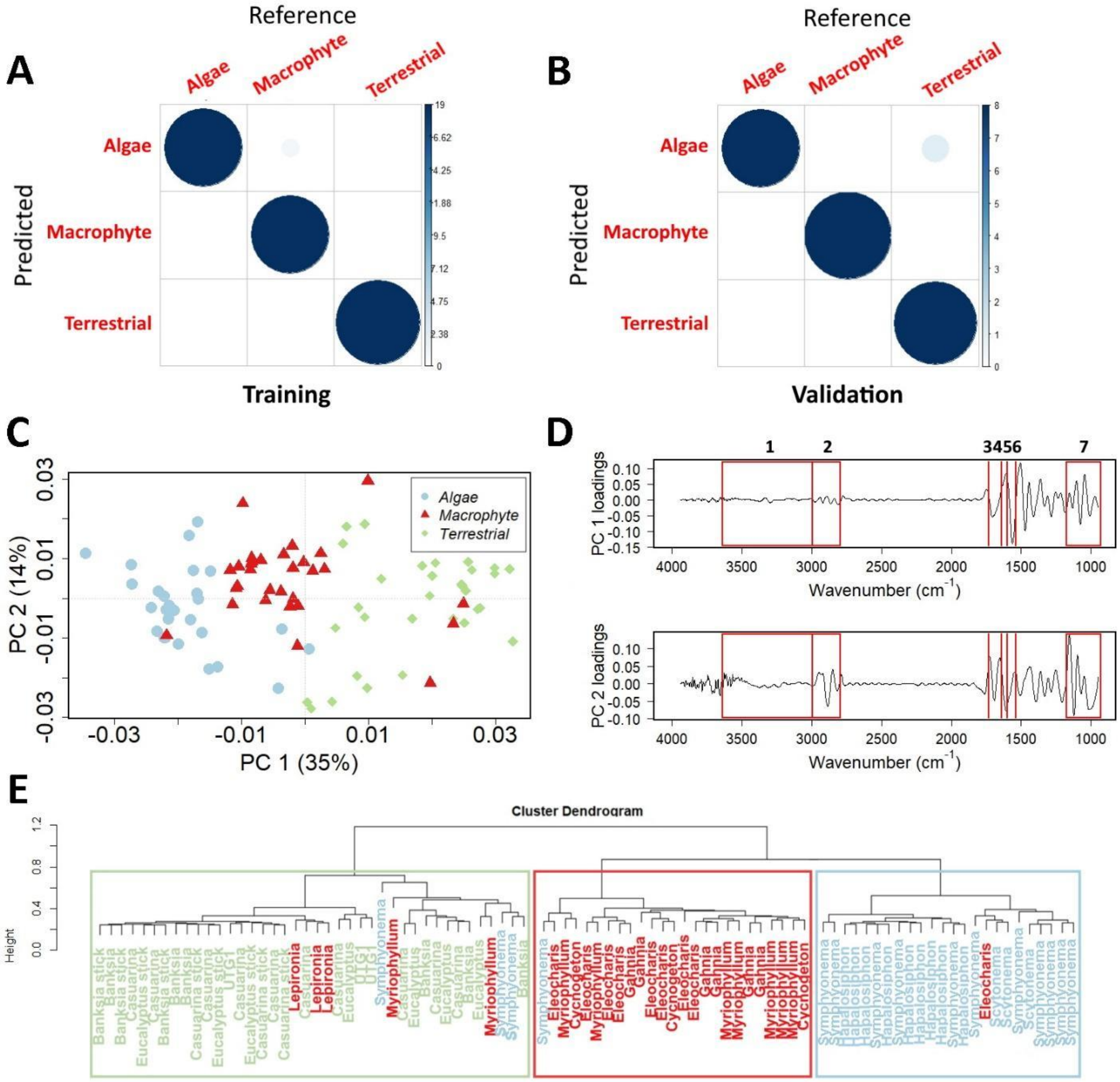


*Figure 3:* Autotroph group processed data validation and training confusion matrices (A and B), PCA (C), PC 1 and 2 (D), and cluster analysis (E). PC 1 and 2 loadings labelled with organic matter spectra from Figure 1. Cluster analysis data labelled with genera, with boxes that correspond to group PCA colours in (C), to highlight how autotrophs are grouped.

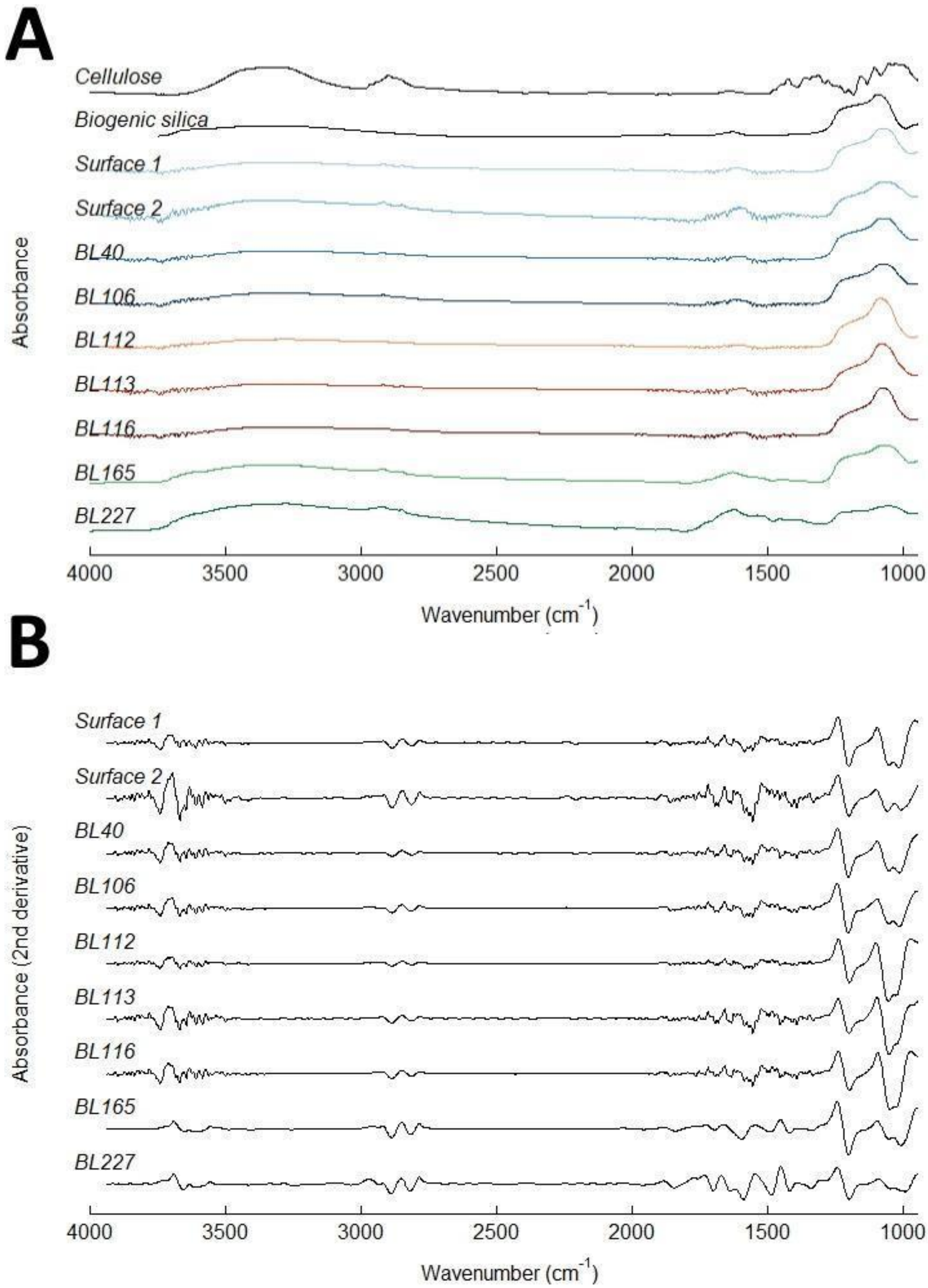


*Figure 4:* A) Sediment mean unprocessed stack with cellulose and biogenic silica spectra (Meyer-Jacob et al. 2014) for comparison. B) Smoothed and derived sediment data.

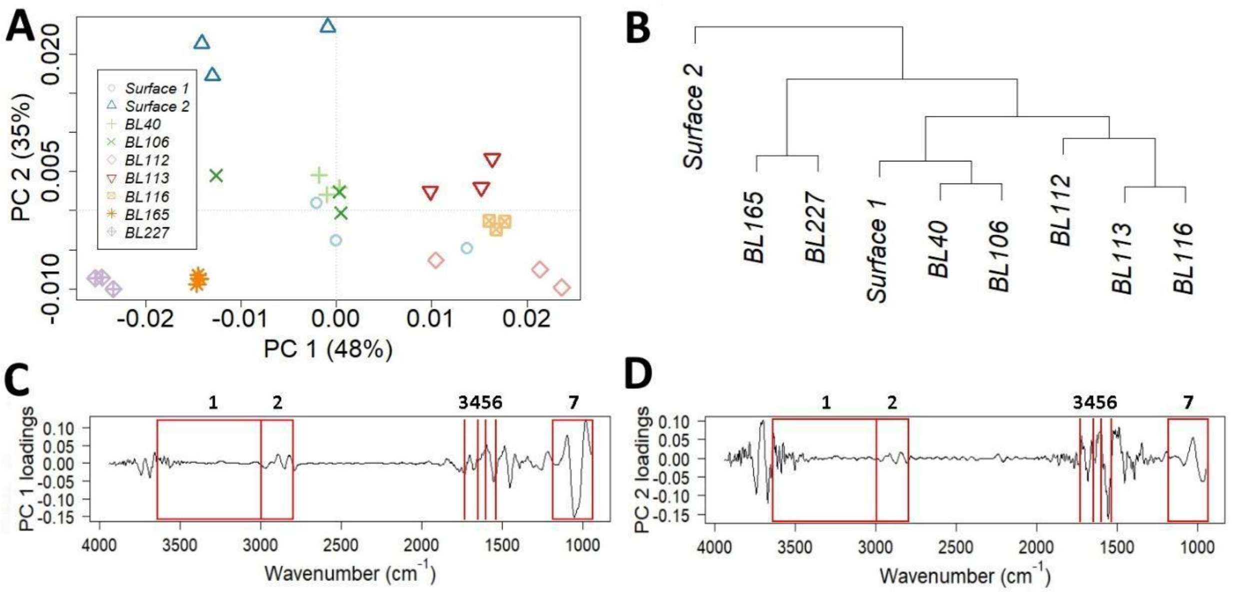


*Figure 5:* Sediment processed spectra (A) PCA, (B) hierarchical clustering, (C) PC 1, and (D) PC 2. PC 1 and 2 loadings labelled with organic matter spectra from Figure 1 and represent the same spectra as Figure 2.

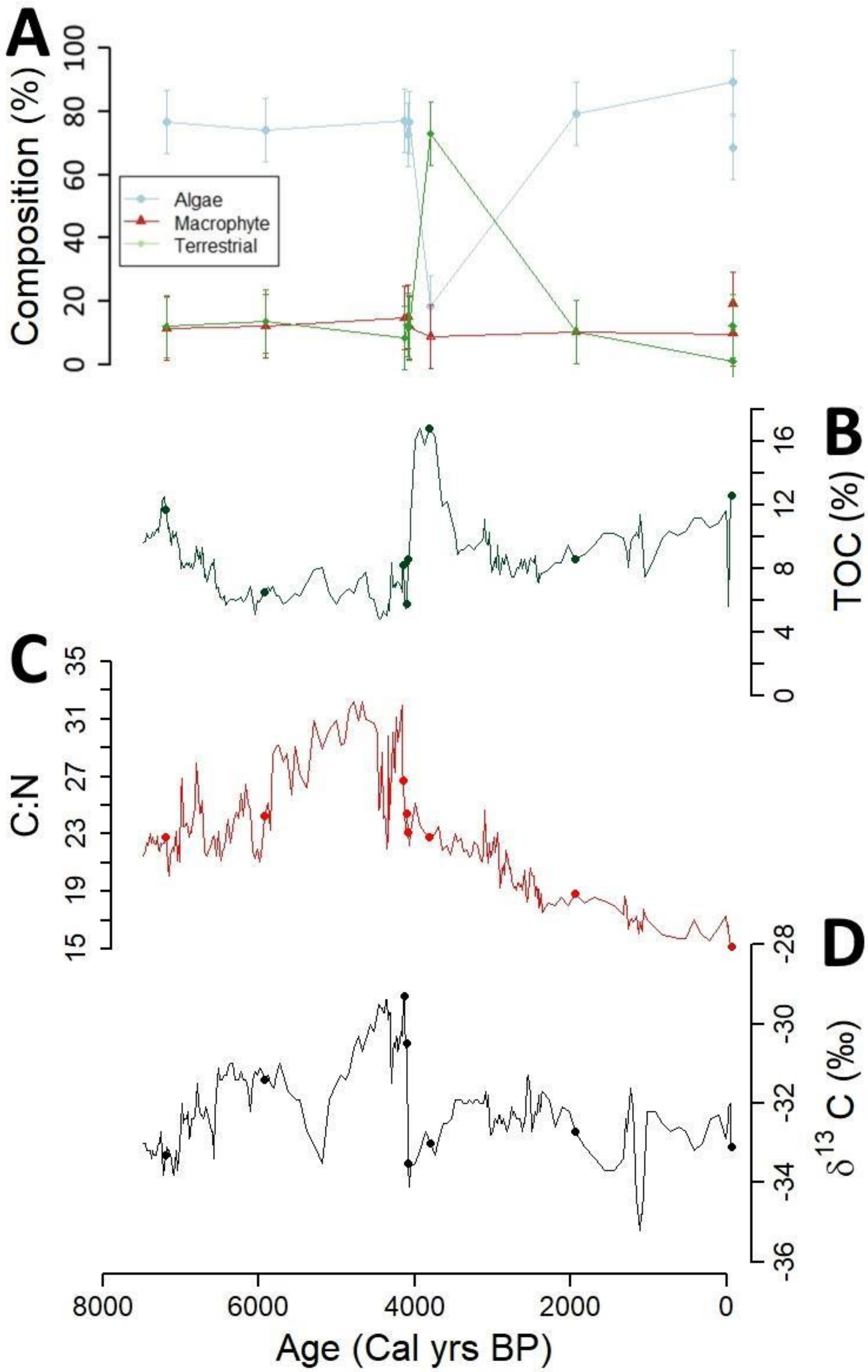


*Figure 6:* Predicted sediment constituents of terrestrial plants, aquatic macrophytes and algae, plotted against age (A) and compared to core sediment TOC (B), C:N (C), and $\delta^{13}C$ (D) of Maxson et al (in review). Error bars on composition estimates represent out-of-bag error of the model (10%). Dots on TOC, C:N, and $\delta^{13}C$ data represent the location of FTIR sediment samples. Sediment surface samples (Surface 1 and 2) were taken in the vicinity of the core (~20 metres away; Surface 1) and 150 metres north of the core (Surface 2).

**Table 1:** Plant and algae classification. Algal samples list dominant species in each sample with percent abundance of that species, if available.

| **Plant and algal samples** | | |
|---|---|---|
| Group | Taxon (percent abundance) | Notes |
| algae | *Symphyonema karboorum (>90%)* | Collected from in lake woody debris |
| algae | *Hapalosiphon pumilus (97%)* | Algal ball from deep lake |
| algae | *Hapalosiphon pumilus (>90%)* | Collected from *Eleocharis difformis* |
| algae | *Symphyonema karboorum (95%)* | |
| algae | *Symphyonema karboorum (89%)* | |
| algae | *Scytonema mirabile (>95%)* | Collected from in lake woody debris |
| algae | *Symphyonema karboorum* | |
| algae | *Symphyonema karboorum* | |
| algae | *Hapalosiphon pumilus (91%)* | Collected from *Lepironia articulata* |
| terrestrial | Unknown | Wood debris collected from in the lake |
| terrestrial | *Casuarina* sp. | *Casuarina* stick sample |
| terrestrial | *Casuarina* sp. | *Casuarina* leaf sample |
| terrestrial | *Casuarina* sp. | *Casuarina* leaf sample |
| terrestrial | *Casuarina* sp. | *Casuarina* stick sample |
| terrestrial | *Eucalyptus* sp. | *Eucalyptus* leaf sample |
| terrestrial | *Eucalyptus* sp. | *Eucalyptus* stick sample |
| terrestrial | *Banksia* sp. | *Banksia* stick sample |
| terrestrial | *Banksia* sp. | *Banksia* leaf sample |
| terrestrial | *Banksia* sp. | *Banksia* leaf sample |
| Aquatic macrophyte | *Eleocharis difformis* | |
| Aquatic macrophyte | *Eleocharis difformis* | |
| Aquatic macrophyte | *Eleocharis difformis* | |
| Aquatic macrophyte | *Myriophyllum* sp. | Collected from shallow, near shore environment |
| Aquatic macrophyte | *Myriophyllum* sp. | |
| Aquatic macrophyte | *Cycnogeton procerus* | |
| Aquatic macrophyte | *Gahnia* sp. | |
| Aquatic macrophyte | *Lepironia articulata* | |

**Table 2**: Mixes and their composition with random forest model predicted composition in bold.

| Mixture | Algae (%) | Macrophyte (%) | Terrestrial (%) |
|---|---|---|---|
| Equal | 33 **(33)** | 33 **(48)** | 33 **(19)** |
| Algae 50 | 50 **(24)** | 25 **(44)** | 25 **(32)** |
| Macrophyte 50 | 25 **(13)** | 50 **(76)** | 25 **(11)** |
| Terrestrial 50 | 25 **(5)** | 25 **(56)** | 50 **(39)** |
| Algae 80 | 80 **(70)** | 10 **(21)** | 10 **(9)** |
| Macrophyte 80 | 10 **(13)** | 80 **(74)** | 10 **(13)** |
| Terrestrial 80 | 10 **(2)** | 10 **(16)** | 80 **(82)** |

**Table 3:** Sediment sample depth and age with random forest model estimated percent composition by plant type. Surface samples are represented as a depth of 0 cm. Cal yr BP: calibrated year before present (where "present" equals 1950 CE).

| **Core samples** | | | | |
|---|---|---|---|---|
| Depth (cm) | Age (Cal yr BP) | Algae (%) | Macrophytes (%) | Terrestrial (%) |
| Surface 1 | -68 | 90 | 8 | 2 |
| Surface 2 | -68 | 69 | 19 | 12 |
| 40 | 1945 | 79 | 12 | 9 |
| 106 | 3804 | 20 | 10 | 70 |
| 112 | 4085 | 76 | 13 | 10 |
| 113 | 4099 | 74 | 16 | 10 |
| 116 | 4140 | 80 | 13 | 7 |
| 165 | 5921 | 75 | 14 | 12 |
| 227 | 7197 | 77 | 13 | 9 |

Supplementary Information.

All data and the code for the Random Forest model can be found on Figshare at: https://figshare.com/projects/Fourier_transform_infrared_spectroscopy_as_a_tracer_of_organic_matter_sources_in_lake_sediments/100088

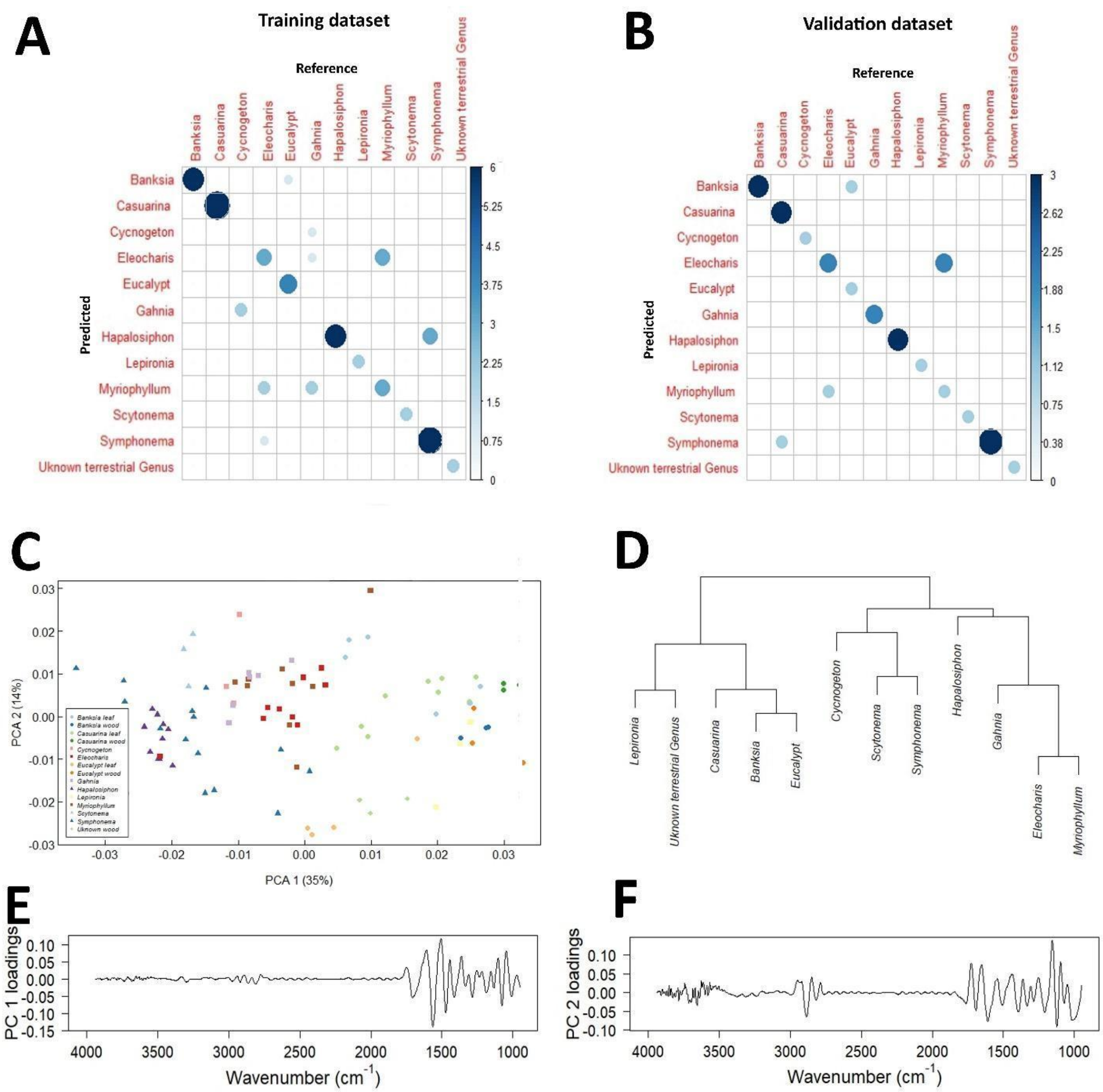


*Figure S1:* Plant genera processed data training and validation confusion matrices (A and B), PCA (C), cluster analysis (D), PC 1 and 2 (E and F). Code available from (Jardine et al. 2019).

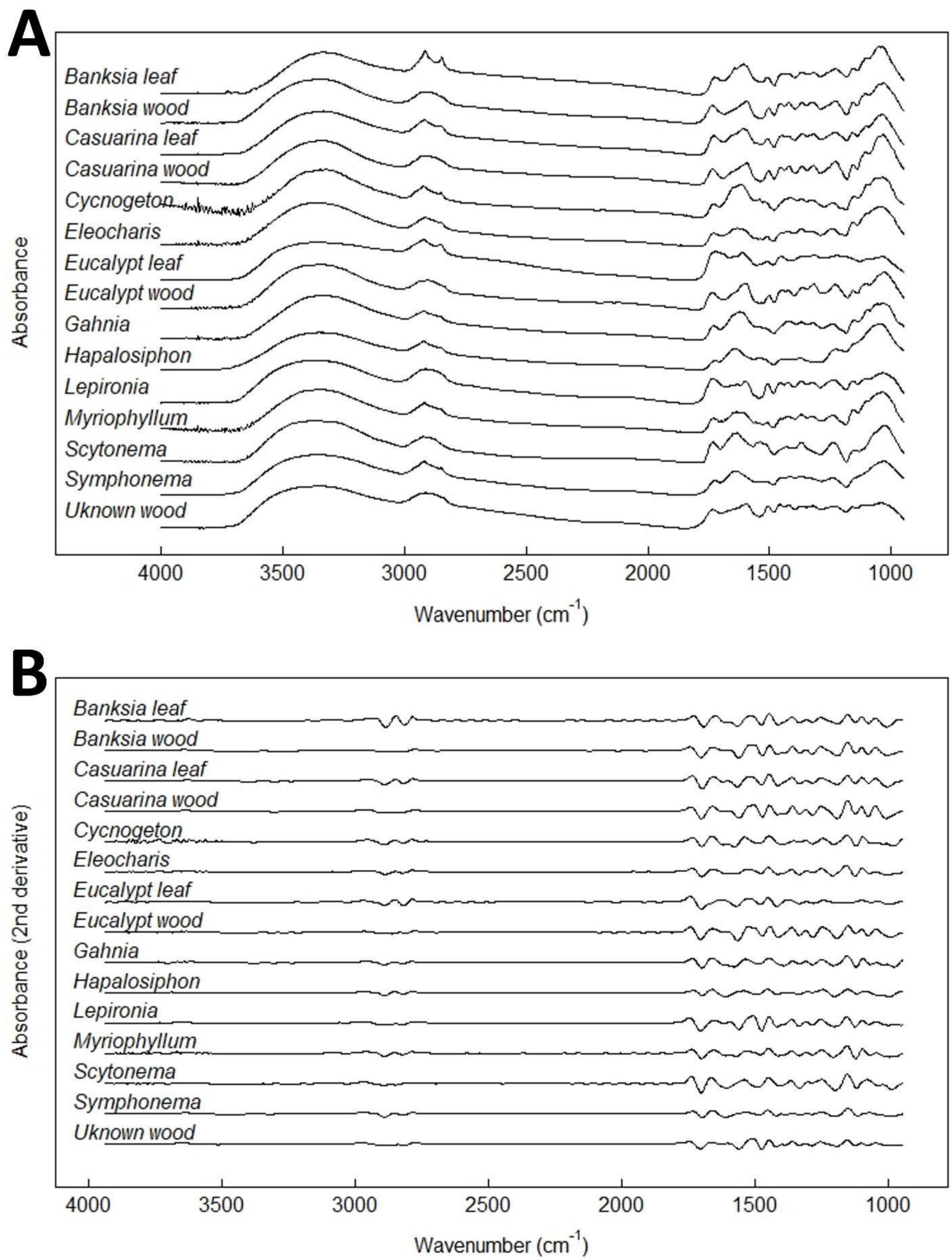


*Figure S2:* Unprocessed mean spectra of genera data (A). Smoothed and derived genera data (B).

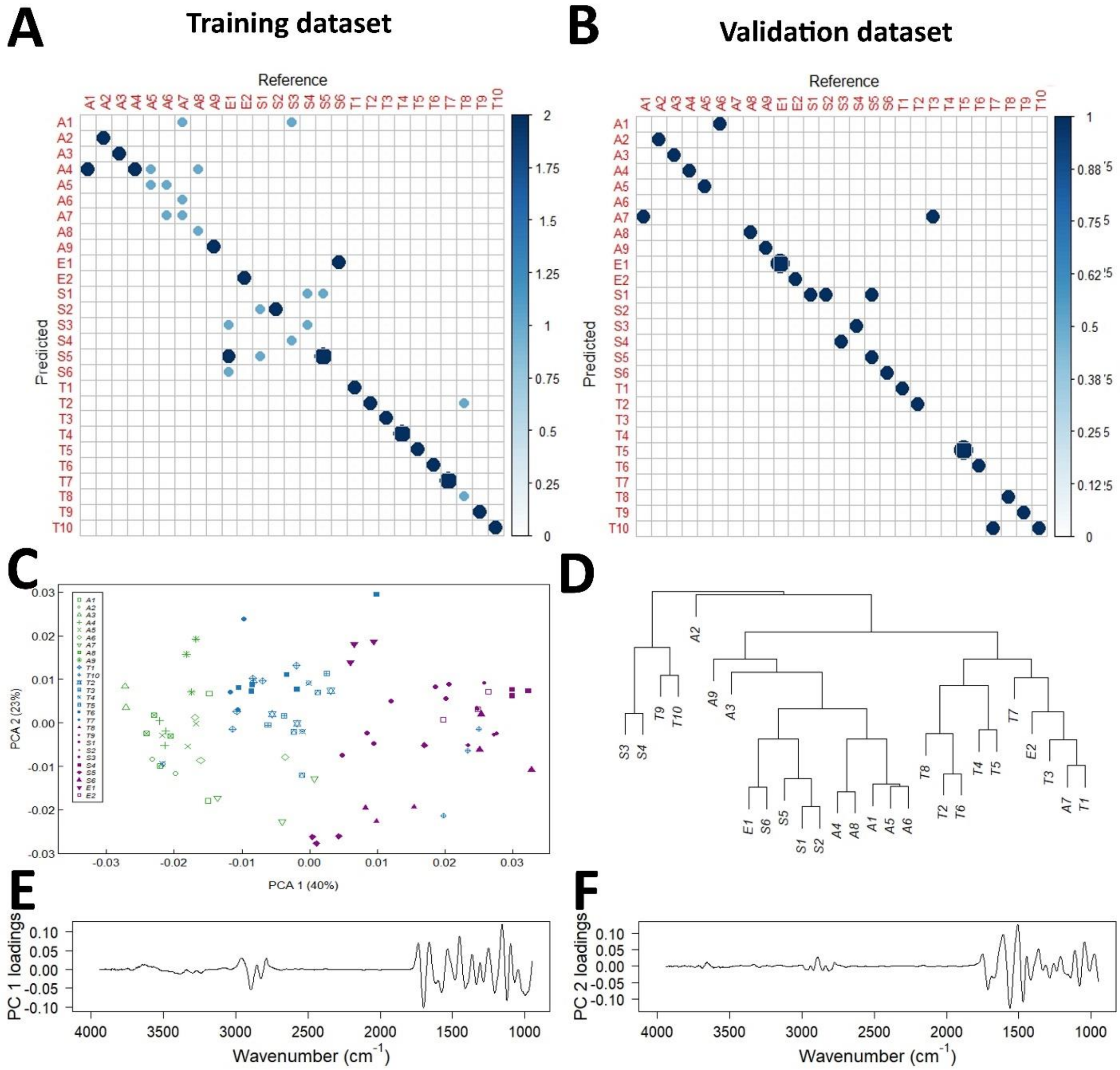


*Figure S3:* Plant individual processed data training and validation confusion matrices (A and B), PCA (C), cluster analysis (D), PC 1 and 2 (E and F). Macrophytes are separated into emergent (E) and submerged (S) species. Plants identified in Table S3. Code available from (Jardine et al. 2019).

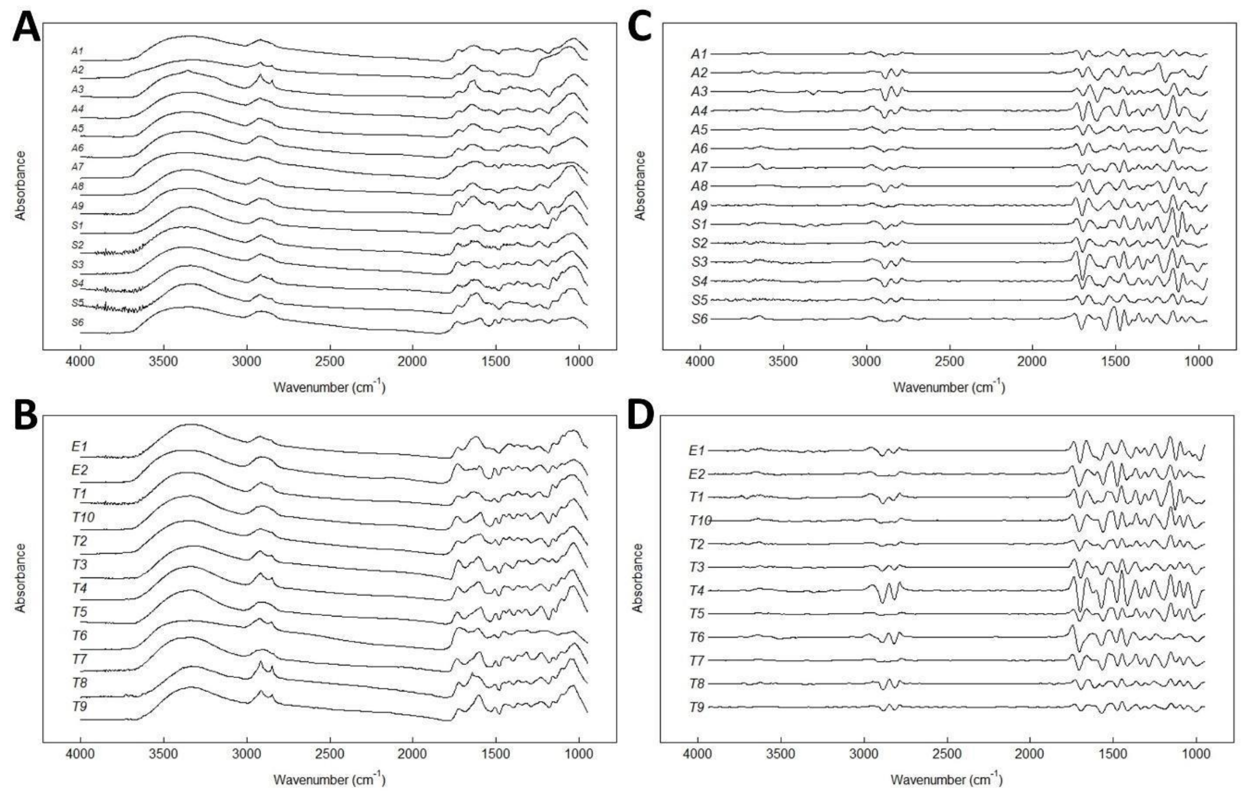


*Figure S4:* Unprocessed mean spectra of individual data, split into algae and submerged macrophytes (A) and emergent macrophytes and terrestrial (B) samples. Smoothed and derived species data split into algae and submerged macrophytes (C) and emergent macrophytes and terrestrial (D) samples. Emergent species are *Myriophyllum* sp., *Eleocharis* sp., and *Cycnogeton* sp. Emergent species are *Lepironia articulata* and *Gahnia* sp.

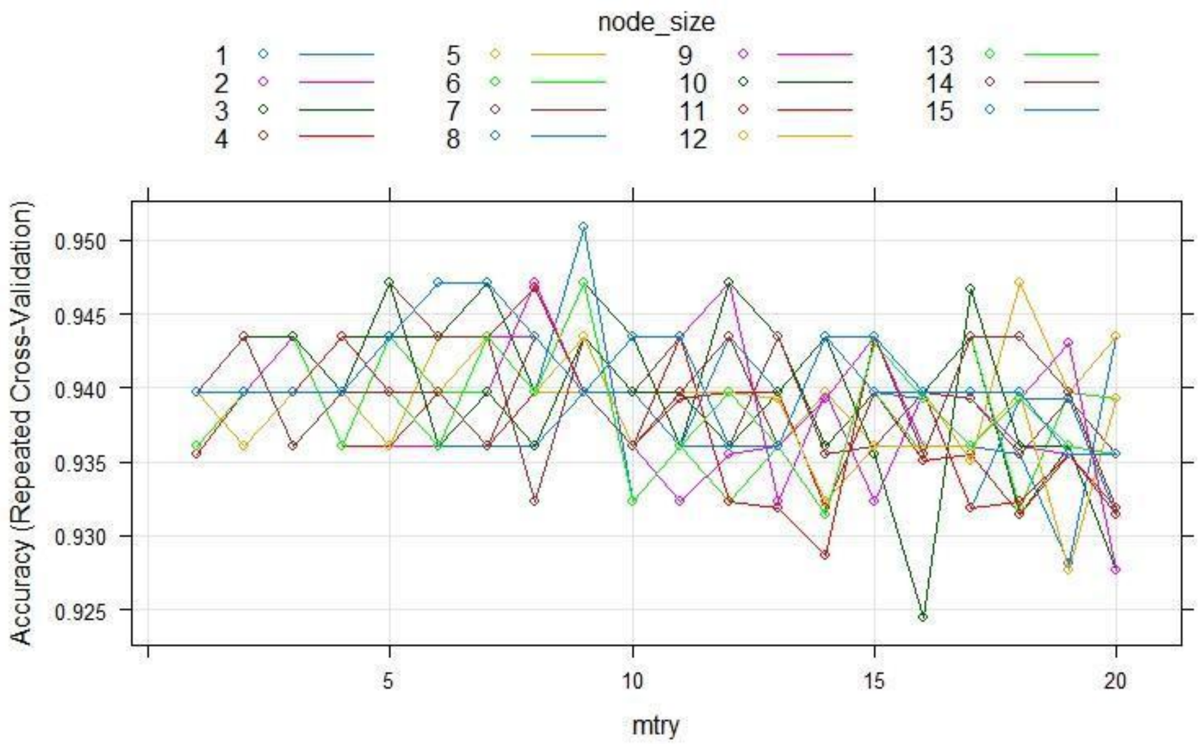


*Figure S5:* Testing of mtry and node_size with ntree held constant.

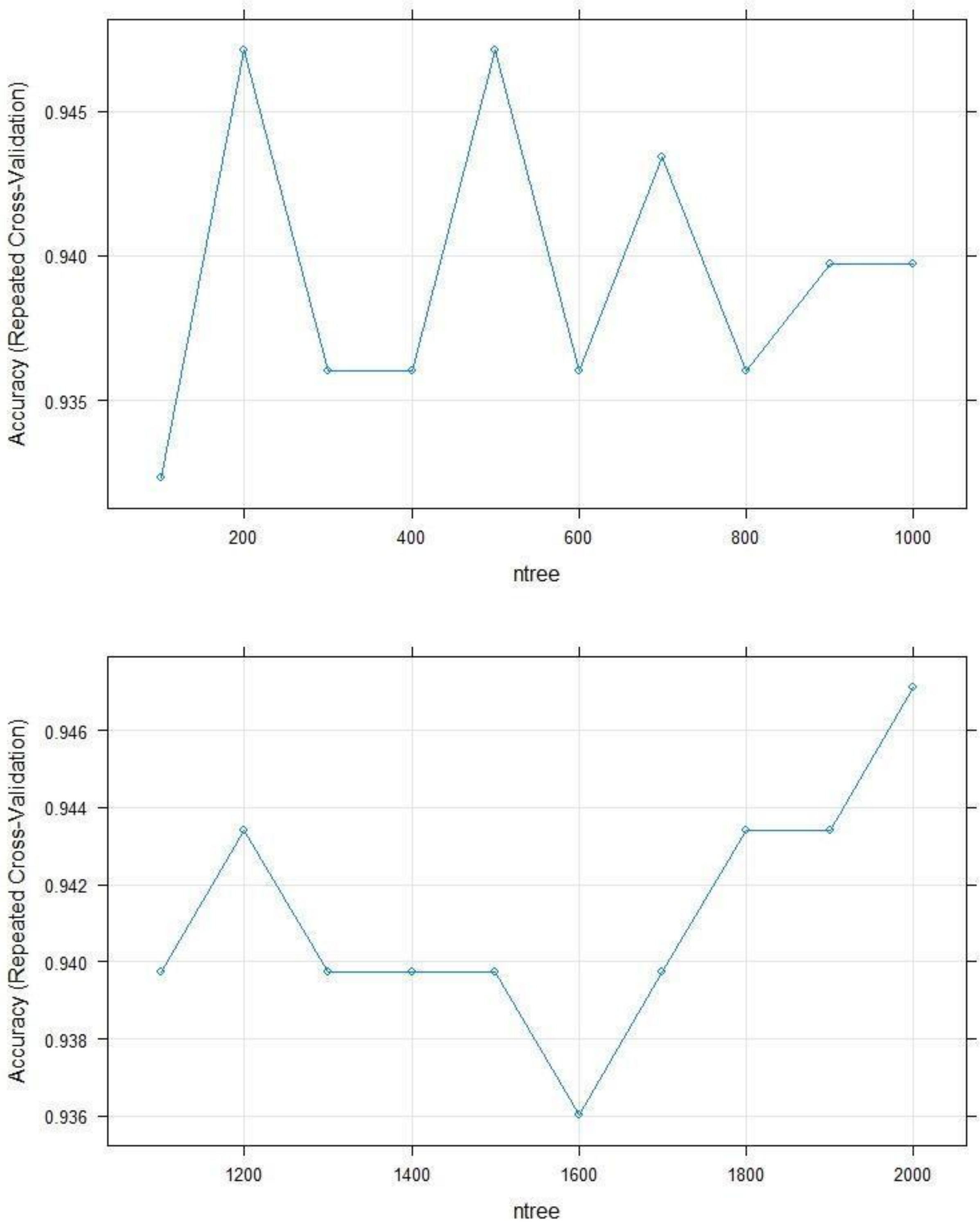


*Figure S6:* Testing of ntree with marry and node_size held constant.

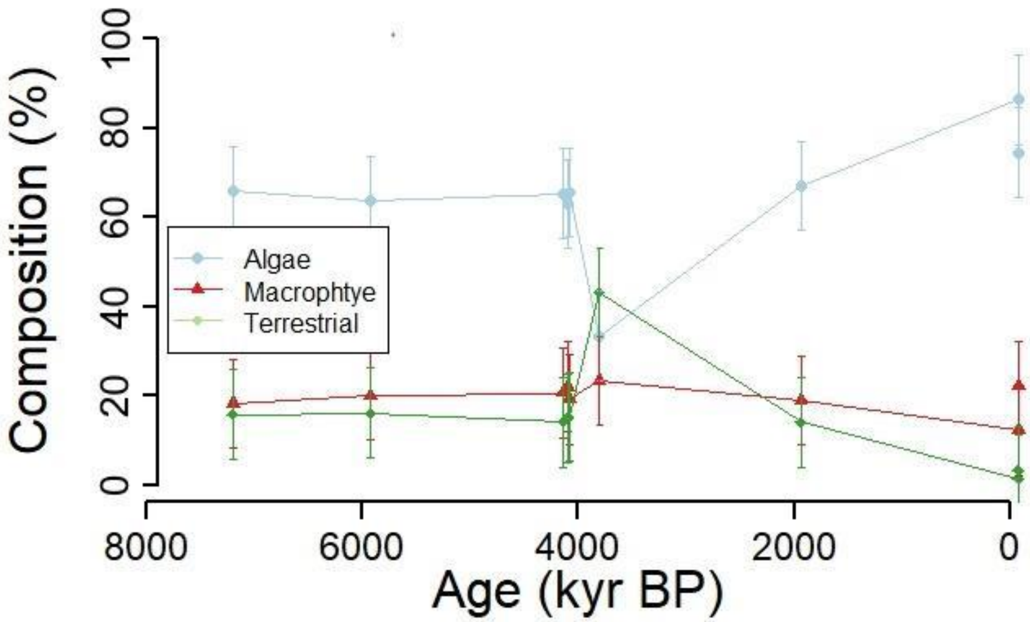


*Figure S6:* Data further truncated to 1800 – 1300 cm$^{-1}$ to test for any potential influence of biogenic silica on the spectra.

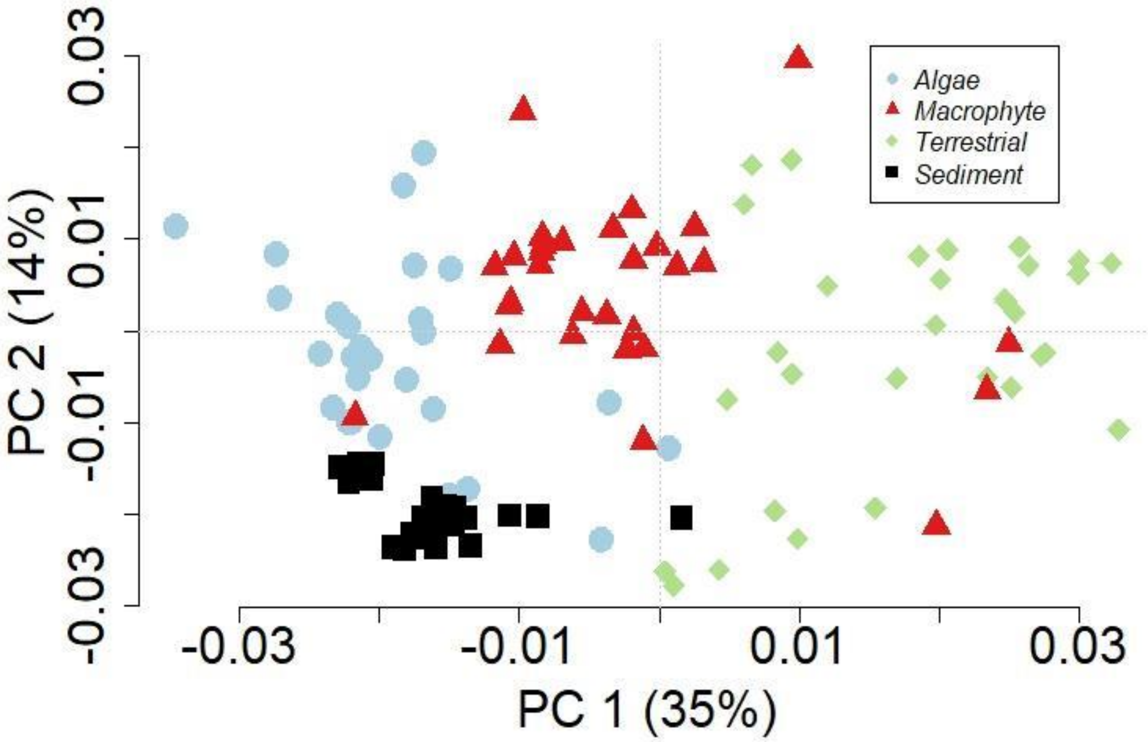


*Figure S7:* PCA biplot of modern autotrophs with the location of predicted fossil sediment PC sample scores based on a model trained by plant PC scores.

**Sediment PC score estimation**

Methods

Two separate principal component analyses were run on the plant and sediment data in R (R Core Team 2020). In the R package ‘vegan’ v.2.5-6 (Oksanen et al. 2019), the function ‘predict’ was used to predict the PC scores of the sediment in the PC space of the plant samples.

Results

The sediment data plots mostly in the PC space of algae, with an excursion of a couple of datapoints into the terrestrial PC space (Fig. S7), almost exactly the result of our random forest model (Fig. 6A). The difference is the RF model can predict proportion of each plant group in the sediment.

| mtry | node_size | ntree | Accuracy | Kappa | AccuracySD | KappaSD |
|---|---|---|---|---|---|---|
| 6 | 4 | 200 | 0.94713 | 0.920564 | 0.0646382 | 0.097062 |
| 6 | 4 | 500 | 0.94713 | 0.920564 | 0.0709191 | 0.106474 |
| 6 | 4 | 2000 | 0.94713 | 0.920564 | 0.0709191 | 0.106474 |
| 6 | 4 | 700 | 0.943426 | 0.915008 | 0.0709642 | 0.106534 |
| 6 | 4 | 1200 | 0.943426 | 0.915008 | 0.0709642 | 0.106534 |
| 6 | 4 | 1800 | 0.943426 | 0.915008 | 0.0709642 | 0.106534 |
| 6 | 4 | 1900 | 0.943426 | 0.915008 | 0.0709642 | 0.106534 |
| 6 | 4 | 900 | 0.939722 | 0.909453 | 0.0765855 | 0.114953 |
| 6 | 4 | 1000 | 0.939722 | 0.909453 | 0.0765855 | 0.114953 |
| 6 | 4 | 1100 | 0.939722 | 0.909453 | 0.0765855 | 0.114953 |
| 6 | 4 | 1300 | 0.939722 | 0.909453 | 0.0765855 | 0.114953 |
| 6 | 4 | 1400 | 0.939722 | 0.909453 | 0.0765855 | 0.114953 |
| 6 | 4 | 1500 | 0.939722 | 0.909453 | 0.0765855 | 0.114953 |
| 6 | 4 | 1700 | 0.939722 | 0.909453 | 0.0765855 | 0.114953 |
| 6 | 4 | 300 | 0.936019 | 0.903897 | 0.081648 | 0.122536 |
| 6 | 4 | 400 | 0.936019 | 0.903897 | 0.081648 | 0.122536 |
| 6 | 4 | 600 | 0.936019 | 0.903897 | 0.081648 | 0.122536 |
| 6 | 4 | 800 | 0.936019 | 0.903897 | 0.081648 | 0.122536 |
| 6 | 4 | 1600 | 0.936019 | 0.903897 | 0.081648 | 0.122536 |
| 6 | 4 | 100 | 0.932315 | 0.898342 | 0.08625 | 0.12943 |

**Table S1:** Results of ntree testing, organised by highest accuracy

**Table S2:** mtry and node_size test results, organised by highest accuracy

| mtry | ntree | node_size | Accuracy | Kappa | AccuracySD | KappaSD |
|---|---|---|---|---|---|---|
| 9 | 500 | 1 | 0.950833 | 0.926119 | 0.064369 | 0.096665 |
| 5 | 500 | 7 | 0.94713 | 0.920564 | 0.070919 | 0.106474 |
| 5 | 500 | 10 | 0.94713 | 0.920564 | 0.070919 | 0.106474 |
| 6 | 500 | 15 | 0.94713 | 0.920564 | 0.070919 | 0.106474 |
| 7 | 500 | 3 | 0.94713 | 0.920564 | 0.064638 | 0.097061 |
| 7 | 500 | 15 | 0.94713 | 0.920564 | 0.064638 | 0.097061 |
| 8 | 500 | 2 | 0.94713 | 0.920564 | 0.064638 | 0.097061 |
| 9 | 500 | 3 | 0.94713 | 0.920564 | 0.064638 | 0.097061 |
| 9 | 500 | 13 | 0.94713 | 0.920564 | 0.070919 | 0.106474 |
| 12 | 500 | 2 | 0.94713 | 0.920564 | 0.070919 | 0.106474 |
| 12 | 500 | 3 | 0.94713 | 0.920564 | 0.070919 | 0.106474 |
| 18 | 500 | 5 | 0.94713 | 0.920564 | 0.070919 | 0.106474 |
| 8 | 500 | 11 | 0.946759 | 0.920137 | 0.071202 | 0.106808 |
| 17 | 500 | 10 | 0.946667 | 0.9202 | 0.078545 | 0.116983 |
| 2 | 500 | 10 | 0.943426 | 0.915008 | 0.076729 | 0.115175 |
| 2 | 500 | 14 | 0.943426 | 0.915008 | 0.076729 | 0.115175 |
| 3 | 500 | 2 | 0.943426 | 0.915008 | 0.076729 | 0.115175 |

| | | | | | | |
|---|---|---|---|---|---|---|
| 3 | 500 | 4 | 0.943426 | 0.915008 | 0.076729 | 0.115175 |
| 3 | 500 | 6 | 0.943426 | 0.915008 | 0.076729 | 0.115175 |
| 3 | 500 | 9 | 0.943426 | 0.915008 | 0.076729 | 0.115175 |
| 3 | 500 | 10 | 0.943426 | 0.915008 | 0.076729 | 0.115175 |
| 4 | 500 | 1 | 0.943426 | 0.915008 | 0.076729 | 0.115175 |
| 4 | 500 | 3 | 0.943426 | 0.915008 | 0.076729 | 0.115175 |
| 4 | 500 | 11 | 0.943426 | 0.915008 | 0.08209 | 0.123211 |
| 5 | 500 | 3 | 0.943426 | 0.915008 | 0.076729 | 0.115175 |
| 5 | 500 | 6 | 0.943426 | 0.915008 | 0.070964 | 0.106534 |
| 5 | 500 | 15 | 0.943426 | 0.915008 | 0.076729 | 0.115175 |
| 6 | 500 | 3 | 0.943426 | 0.915008 | 0.076729 | 0.115175 |
| 6 | 500 | 5 | 0.943426 | 0.915008 | 0.070964 | 0.106534 |
| 6 | 500 | 7 | 0.943426 | 0.915008 | 0.070964 | 0.106534 |
| 7 | 500 | 1 | 0.943426 | 0.915008 | 0.070964 | 0.106534 |
| 7 | 500 | 5 | 0.943426 | 0.915008 | 0.070964 | 0.106534 |
| 7 | 500 | 7 | 0.943426 | 0.915008 | 0.070964 | 0.106534 |
| 7 | 500 | 9 | 0.943426 | 0.915008 | 0.070964 | 0.106534 |
| 7 | 500 | 11 | 0.943426 | 0.915008 | 0.070964 | 0.106534 |
| 7 | 500 | 12 | 0.943426 | 0.915008 | 0.070964 | 0.106534 |
| 7 | 500 | 13 | 0.943426 | 0.915008 | 0.070964 | 0.106534 |
| 8 | 500 | 9 | 0.943426 | 0.915008 | 0.070964 | 0.106534 |
| 8 | 500 | 14 | 0.943426 | 0.915008 | 0.070964 | 0.106534 |
| 8 | 500 | 15 | 0.943426 | 0.915008 | 0.070964 | 0.106534 |
| 9 | 500 | 4 | 0.943426 | 0.915008 | 0.070964 | 0.106534 |
| 9 | 500 | 7 | 0.943426 | 0.915008 | 0.076729 | 0.115175 |
| 9 | 500 | 10 | 0.943426 | 0.915008 | 0.070964 | 0.106534 |
| 9 | 500 | 12 | 0.943426 | 0.915008 | 0.070964 | 0.106534 |
| 10 | 500 | 3 | 0.943426 | 0.915008 | 0.070964 | 0.106534 |

| | | | | | | |
|---|---|---|---|---|---|---|
| 10 | 500 | 15 | 0.943426 | 0.915008 | 0.070964 | 0.106534 |
| 11 | 500 | 2 | 0.943426 | 0.915008 | 0.070964 | 0.106534 |
| 11 | 500 | 7 | 0.943426 | 0.915008 | 0.070964 | 0.106534 |
| 11 | 500 | 11 | 0.943426 | 0.915008 | 0.070964 | 0.106534 |
| 11 | 500 | 15 | 0.943426 | 0.915008 | 0.070964 | 0.106534 |
| 12 | 500 | 8 | 0.943426 | 0.915008 | 0.070964 | 0.106534 |
| 12 | 500 | 14 | 0.943426 | 0.915008 | 0.070964 | 0.106534 |
| 13 | 500 | 3 | 0.943426 | 0.915008 | 0.076729 | 0.115175 |
| 13 | 500 | 7 | 0.943426 | 0.915008 | 0.070964 | 0.106534 |
| 14 | 500 | 1 | 0.943426 | 0.915008 | 0.070964 | 0.106534 |
| 14 | 500 | 6 | 0.943426 | 0.915008 | 0.070964 | 0.106534 |
| 14 | 500 | 8 | 0.943426 | 0.915008 | 0.070964 | 0.106534 |
| 14 | 500 | 10 | 0.943426 | 0.915008 | 0.070964 | 0.106534 |
| 14 | 500 | 14 | 0.943426 | 0.915008 | 0.070964 | 0.106534 |
| 14 | 500 | 15 | 0.943426 | 0.915008 | 0.070964 | 0.106534 |
| 15 | 500 | 9 | 0.943426 | 0.915008 | 0.070964 | 0.106534 |
| 15 | 500 | 11 | 0.943426 | 0.915008 | 0.070964 | 0.106534 |
| 15 | 500 | 14 | 0.943426 | 0.915008 | 0.076729 | 0.115175 |
| 15 | 500 | 15 | 0.943426 | 0.915008 | 0.070964 | 0.106534 |
| 17 | 500 | 3 | 0.943426 | 0.915008 | 0.070964 | 0.106534 |
| 17 | 500 | 6 | 0.943426 | 0.915008 | 0.070964 | 0.106534 |
| 17 | 500 | 14 | 0.943426 | 0.915008 | 0.070964 | 0.106534 |
| 18 | 500 | 14 | 0.943426 | 0.915008 | 0.070964 | 0.106534 |
| 20 | 500 | 1 | 0.943426 | 0.915008 | 0.070964 | 0.106534 |
| 20 | 500 | 5 | 0.943426 | 0.915008 | 0.070964 | 0.106534 |
| 15 | 500 | 13 | 0.942963 | 0.914362 | 0.078563 | 0.117721 |
| 19 | 500 | 9 | 0.942963 | 0.914644 | 0.078563 | 0.11702 |
| 1 | 500 | 1 | 0.939722 | 0.909453 | 0.081956 | 0.123004 |

| | | | | | | |
|---|---|---|---|---|---|---|
| 1 | 500 | 2 | 0.939722 | 0.909453 | 0.081956 | 0.123004 |
| 1 | 500 | 4 | 0.939722 | 0.909453 | 0.081956 | 0.123004 |
| 1 | 500 | 5 | 0.939722 | 0.909453 | 0.081956 | 0.123004 |
| 1 | 500 | 6 | 0.939722 | 0.909453 | 0.081956 | 0.123004 |
| 1 | 500 | 7 | 0.939722 | 0.909453 | 0.081956 | 0.123004 |
| 1 | 500 | 8 | 0.939722 | 0.909453 | 0.081956 | 0.123004 |
| 1 | 500 | 10 | 0.939722 | 0.909453 | 0.081956 | 0.123004 |
| 1 | 500 | 12 | 0.939722 | 0.909453 | 0.081956 | 0.123004 |
| 1 | 500 | 14 | 0.939722 | 0.909453 | 0.081956 | 0.123004 |
| 1 | 500 | 15 | 0.939722 | 0.909453 | 0.081956 | 0.123004 |
| 2 | 500 | 1 | 0.939722 | 0.909453 | 0.081956 | 0.123004 |
| 2 | 500 | 2 | 0.939722 | 0.909453 | 0.081956 | 0.123004 |
| 2 | 500 | 3 | 0.939722 | 0.909453 | 0.081956 | 0.123004 |
| 2 | 500 | 4 | 0.939722 | 0.909453 | 0.081956 | 0.123004 |
| 2 | 500 | 6 | 0.939722 | 0.909453 | 0.081956 | 0.123004 |
| 2 | 500 | 7 | 0.939722 | 0.909453 | 0.081956 | 0.123004 |
| 2 | 500 | 8 | 0.939722 | 0.909453 | 0.081956 | 0.123004 |
| 2 | 500 | 9 | 0.939722 | 0.909453 | 0.081956 | 0.123004 |
| 2 | 500 | 11 | 0.939722 | 0.909453 | 0.081956 | 0.123004 |
| 2 | 500 | 12 | 0.939722 | 0.909453 | 0.081956 | 0.123004 |
| 2 | 500 | 13 | 0.939722 | 0.909453 | 0.081956 | 0.123004 |
| 2 | 500 | 15 | 0.939722 | 0.909453 | 0.081956 | 0.123004 |
| 3 | 500 | 1 | 0.939722 | 0.909453 | 0.081956 | 0.123004 |
| 3 | 500 | 3 | 0.939722 | 0.909453 | 0.081956 | 0.123004 |
| 3 | 500 | 5 | 0.939722 | 0.909453 | 0.081956 | 0.123004 |
| 3 | 500 | 7 | 0.939722 | 0.909453 | 0.081956 | 0.123004 |
| 3 | 500 | 8 | 0.939722 | 0.909453 | 0.081956 | 0.123004 |
| 3 | 500 | 11 | 0.939722 | 0.909453 | 0.081956 | 0.123004 |

| | | | | | | |
|---|---|---|---|---|---|---|
| 3 | 500 | 12 | 0.939722 | 0.909453 | 0.081956 | 0.123004 |
| 3 | 500 | 13 | 0.939722 | 0.909453 | 0.081956 | 0.123004 |
| 3 | 500 | 15 | 0.939722 | 0.909453 | 0.081956 | 0.123004 |
| 4 | 500 | 2 | 0.939722 | 0.909453 | 0.081956 | 0.123004 |
| 4 | 500 | 5 | 0.939722 | 0.909453 | 0.081956 | 0.123004 |
| 4 | 500 | 7 | 0.939722 | 0.909453 | 0.081956 | 0.123004 |
| 4 | 500 | 8 | 0.939722 | 0.909453 | 0.076586 | 0.114953 |
| 4 | 500 | 9 | 0.939722 | 0.909453 | 0.081956 | 0.123004 |
| 4 | 500 | 10 | 0.939722 | 0.909453 | 0.081956 | 0.123004 |
| 4 | 500 | 12 | 0.939722 | 0.909453 | 0.081956 | 0.123004 |
| 4 | 500 | 13 | 0.939722 | 0.909453 | 0.081956 | 0.123004 |
| 4 | 500 | 14 | 0.939722 | 0.909453 | 0.081956 | 0.123004 |
| 4 | 500 | 15 | 0.939722 | 0.909453 | 0.081956 | 0.123004 |
| 5 | 500 | 1 | 0.939722 | 0.909453 | 0.076586 | 0.114953 |
| 5 | 500 | 8 | 0.939722 | 0.909453 | 0.081956 | 0.123004 |
| 5 | 500 | 9 | 0.939722 | 0.909453 | 0.076586 | 0.114953 |
| 5 | 500 | 11 | 0.939722 | 0.909453 | 0.076586 | 0.114953 |
| 5 | 500 | 12 | 0.939722 | 0.909453 | 0.081956 | 0.123004 |
| 5 | 500 | 13 | 0.939722 | 0.909453 | 0.081956 | 0.123004 |
| 5 | 500 | 14 | 0.939722 | 0.909453 | 0.076586 | 0.114953 |
| 6 | 500 | 4 | 0.939722 | 0.909453 | 0.076586 | 0.114953 |
| 6 | 500 | 6 | 0.939722 | 0.909453 | 0.076586 | 0.114953 |
| 6 | 500 | 11 | 0.939722 | 0.909453 | 0.076586 | 0.114953 |
| 6 | 500 | 12 | 0.939722 | 0.909453 | 0.076586 | 0.114953 |
| 6 | 500 | 14 | 0.939722 | 0.909453 | 0.076586 | 0.114953 |
| 7 | 500 | 2 | 0.939722 | 0.909453 | 0.076586 | 0.114953 |
| 7 | 500 | 6 | 0.939722 | 0.909453 | 0.076586 | 0.114953 |
| 7 | 500 | 10 | 0.939722 | 0.909453 | 0.076586 | 0.114953 |

| | | | | | | |
|---|---|---|---|---|---|---|
| 8 | 500 | 1 | 0.939722 | 0.909453 | 0.076586 | 0.114953 |
| 8 | 500 | 3 | 0.939722 | 0.909453 | 0.076586 | 0.114953 |
| 8 | 500 | 4 | 0.939722 | 0.909453 | 0.076586 | 0.114953 |
| 8 | 500 | 5 | 0.939722 | 0.909453 | 0.076586 | 0.114953 |
| 8 | 500 | 12 | 0.939722 | 0.909453 | 0.076586 | 0.114953 |
| 8 | 500 | 13 | 0.939722 | 0.909453 | 0.076586 | 0.114953 |
| 9 | 500 | 2 | 0.939722 | 0.909453 | 0.076586 | 0.114953 |
| 9 | 500 | 5 | 0.939722 | 0.909453 | 0.076586 | 0.114953 |
| 9 | 500 | 6 | 0.939722 | 0.909453 | 0.081956 | 0.123004 |
| 9 | 500 | 8 | 0.939722 | 0.909453 | 0.076586 | 0.114953 |
| 9 | 500 | 9 | 0.939722 | 0.909453 | 0.076586 | 0.114953 |
| 9 | 500 | 11 | 0.939722 | 0.909453 | 0.076586 | 0.114953 |
| 9 | 500 | 14 | 0.939722 | 0.909453 | 0.076586 | 0.114953 |
| 9 | 500 | 15 | 0.939722 | 0.909453 | 0.076586 | 0.114953 |
| 10 | 500 | 6 | 0.939722 | 0.909453 | 0.076586 | 0.114953 |
| 10 | 500 | 7 | 0.939722 | 0.909453 | 0.076586 | 0.114953 |
| 10 | 500 | 8 | 0.939722 | 0.909453 | 0.076586 | 0.114953 |
| 10 | 500 | 10 | 0.939722 | 0.909453 | 0.076586 | 0.114953 |
| 11 | 500 | 5 | 0.939722 | 0.909453 | 0.081956 | 0.123004 |
| 11 | 500 | 10 | 0.939722 | 0.909453 | 0.076586 | 0.114953 |
| 11 | 500 | 12 | 0.939722 | 0.909453 | 0.070809 | 0.106295 |
| 11 | 500 | 14 | 0.939722 | 0.909453 | 0.070809 | 0.106295 |
| 12 | 500 | 4 | 0.939722 | 0.909453 | 0.076586 | 0.114953 |
| 12 | 500 | 5 | 0.939722 | 0.909453 | 0.076586 | 0.114953 |
| 12 | 500 | 12 | 0.939722 | 0.909453 | 0.076586 | 0.114953 |
| 12 | 500 | 13 | 0.939722 | 0.909453 | 0.076586 | 0.114953 |
| 13 | 500 | 1 | 0.939722 | 0.909453 | 0.070809 | 0.106295 |
| 13 | 500 | 4 | 0.939722 | 0.909453 | 0.081956 | 0.123004 |

| | | | | | | |
|---|---|---|---|---|---|---|
| 13 | 500 | 8 | 0.939722 | 0.909453 | 0.070809 | 0.106295 |
| 13 | 500 | 10 | 0.939722 | 0.909453 | 0.076586 | 0.114953 |
| 14 | 500 | 2 | 0.939722 | 0.909008 | 0.070809 | 0.106651 |
| 14 | 500 | 5 | 0.939722 | 0.909453 | 0.070809 | 0.106295 |
| 15 | 500 | 1 | 0.939722 | 0.909453 | 0.076586 | 0.114953 |
| 15 | 500 | 3 | 0.939722 | 0.909453 | 0.070809 | 0.106295 |
| 15 | 500 | 4 | 0.939722 | 0.909453 | 0.070809 | 0.106295 |
| 15 | 500 | 6 | 0.939722 | 0.909453 | 0.076586 | 0.114953 |
| 15 | 500 | 8 | 0.939722 | 0.909453 | 0.076586 | 0.114953 |
| 16 | 500 | 2 | 0.939722 | 0.909453 | 0.070809 | 0.106295 |
| 16 | 500 | 3 | 0.939722 | 0.909453 | 0.070809 | 0.106295 |
| 16 | 500 | 5 | 0.939722 | 0.909453 | 0.070809 | 0.106295 |
| 16 | 500 | 7 | 0.939722 | 0.909453 | 0.076586 | 0.114953 |
| 16 | 500 | 8 | 0.939722 | 0.909453 | 0.076586 | 0.114953 |
| 16 | 500 | 11 | 0.939722 | 0.909453 | 0.070809 | 0.106295 |
| 16 | 500 | 15 | 0.939722 | 0.909453 | 0.070809 | 0.106295 |
| 17 | 500 | 15 | 0.939722 | 0.909453 | 0.070809 | 0.106295 |
| 18 | 500 | 12 | 0.939722 | 0.909453 | 0.070809 | 0.106295 |
| 18 | 500 | 15 | 0.939722 | 0.909453 | 0.076586 | 0.114953 |
| 19 | 500 | 5 | 0.939722 | 0.909453 | 0.076586 | 0.114953 |
| 19 | 500 | 6 | 0.939722 | 0.909453 | 0.076586 | 0.114953 |
| 19 | 500 | 14 | 0.939722 | 0.909453 | 0.070809 | 0.106295 |
| 11 | 500 | 4 | 0.939259 | 0.909089 | 0.083654 | 0.124716 |
| 13 | 500 | 12 | 0.939259 | 0.908807 | 0.0784 | 0.117473 |
| 14 | 500 | 9 | 0.939259 | 0.908807 | 0.0784 | 0.117473 |
| 16 | 500 | 1 | 0.939259 | 0.909089 | 0.0784 | 0.116784 |
| 16 | 500 | 13 | 0.939259 | 0.909089 | 0.0784 | 0.116784 |
| 17 | 500 | 2 | 0.939259 | 0.908362 | 0.0784 | 0.117793 |

| | | | | | | |
|---|---|---|---|---|---|---|
| 17 | 500 | 7 | 0.939259 | 0.909089 | 0.0784 | 0.116784 |
| 18 | 500 | 8 | 0.939259 | 0.908807 | 0.0784 | 0.117473 |
| 18 | 500 | 9 | 0.939259 | 0.909089 | 0.0784 | 0.116784 |
| 18 | 500 | 13 | 0.939259 | 0.909089 | 0.0784 | 0.116784 |
| 19 | 500 | 7 | 0.939259 | 0.908807 | 0.0784 | 0.117473 |
| 19 | 500 | 8 | 0.939259 | 0.908807 | 0.0784 | 0.117473 |
| 20 | 500 | 6 | 0.939259 | 0.909089 | 0.0784 | 0.116784 |
| 20 | 500 | 12 | 0.939259 | 0.909089 | 0.0784 | 0.116784 |
| 1 | 500 | 3 | 0.936019 | 0.903897 | 0.086705 | 0.130118 |
| 1 | 500 | 9 | 0.936019 | 0.903897 | 0.086705 | 0.130118 |
| 1 | 500 | 13 | 0.936019 | 0.903897 | 0.086705 | 0.130118 |
| 2 | 500 | 5 | 0.936019 | 0.903897 | 0.086705 | 0.130118 |
| 3 | 500 | 14 | 0.936019 | 0.903897 | 0.081648 | 0.122536 |
| 4 | 500 | 4 | 0.936019 | 0.903897 | 0.081648 | 0.122536 |
| 4 | 500 | 6 | 0.936019 | 0.903897 | 0.081648 | 0.122536 |
| 5 | 500 | 2 | 0.936019 | 0.903897 | 0.081648 | 0.122536 |
| 5 | 500 | 4 | 0.936019 | 0.903897 | 0.081648 | 0.122536 |
| 5 | 500 | 5 | 0.936019 | 0.903897 | 0.081648 | 0.122536 |
| 6 | 500 | 1 | 0.936019 | 0.903897 | 0.081648 | 0.122536 |
| 6 | 500 | 2 | 0.936019 | 0.903897 | 0.081648 | 0.122536 |
| 6 | 500 | 8 | 0.936019 | 0.903897 | 0.081648 | 0.122536 |
| 6 | 500 | 9 | 0.936019 | 0.903897 | 0.081648 | 0.122536 |
| 6 | 500 | 10 | 0.936019 | 0.903897 | 0.081648 | 0.122536 |
| 6 | 500 | 13 | 0.936019 | 0.903897 | 0.081648 | 0.122536 |
| 7 | 500 | 4 | 0.936019 | 0.903897 | 0.081648 | 0.122536 |
| 7 | 500 | 8 | 0.936019 | 0.903897 | 0.081648 | 0.122536 |
| 7 | 500 | 14 | 0.936019 | 0.903897 | 0.081648 | 0.122536 |
| 8 | 500 | 6 | 0.936019 | 0.903897 | 0.081648 | 0.122536 |

| | | | | | | |
|---|---|---|---|---|---|---|
| 8 | 500 | 8 | 0.936019 | 0.903897 | 0.081648 | 0.122536 |
| 8 | 500 | 10 | 0.936019 | 0.903897 | 0.081648 | 0.122536 |
| 10 | 500 | 2 | 0.936019 | 0.903897 | 0.076256 | 0.114453 |
| 10 | 500 | 4 | 0.936019 | 0.903897 | 0.081648 | 0.122536 |
| 10 | 500 | 5 | 0.936019 | 0.903897 | 0.081648 | 0.122536 |
| 10 | 500 | 9 | 0.936019 | 0.903897 | 0.076256 | 0.114453 |
| 10 | 500 | 11 | 0.936019 | 0.903897 | 0.081648 | 0.122536 |
| 10 | 500 | 12 | 0.936019 | 0.903897 | 0.076256 | 0.114453 |
| 10 | 500 | 14 | 0.936019 | 0.903897 | 0.076256 | 0.114453 |
| 11 | 500 | 1 | 0.936019 | 0.903897 | 0.076256 | 0.114453 |
| 11 | 500 | 3 | 0.936019 | 0.903897 | 0.076256 | 0.114453 |
| 11 | 500 | 6 | 0.936019 | 0.903897 | 0.081648 | 0.122536 |
| 11 | 500 | 8 | 0.936019 | 0.903897 | 0.076256 | 0.114453 |
| 11 | 500 | 13 | 0.936019 | 0.903897 | 0.076256 | 0.114453 |
| 12 | 500 | 1 | 0.936019 | 0.903897 | 0.076256 | 0.114453 |
| 12 | 500 | 7 | 0.936019 | 0.903897 | 0.081648 | 0.122536 |
| 12 | 500 | 10 | 0.936019 | 0.903897 | 0.081648 | 0.122536 |
| 12 | 500 | 15 | 0.936019 | 0.903453 | 0.070452 | 0.106087 |
| 13 | 500 | 5 | 0.936019 | 0.903897 | 0.081648 | 0.122536 |
| 13 | 500 | 6 | 0.936019 | 0.903897 | 0.081648 | 0.122536 |
| 13 | 500 | 9 | 0.936019 | 0.903897 | 0.076256 | 0.114453 |
| 13 | 500 | 13 | 0.936019 | 0.903897 | 0.076256 | 0.114453 |
| 13 | 500 | 14 | 0.936019 | 0.903897 | 0.076256 | 0.114453 |
| 13 | 500 | 15 | 0.936019 | 0.903897 | 0.076256 | 0.114453 |
| 14 | 500 | 3 | 0.936019 | 0.903897 | 0.076256 | 0.114453 |
| 15 | 500 | 5 | 0.936019 | 0.903897 | 0.076256 | 0.114453 |
| 15 | 500 | 7 | 0.936019 | 0.903897 | 0.076256 | 0.114453 |
| 15 | 500 | 12 | 0.936019 | 0.903897 | 0.076256 | 0.114453 |

| | | | | | | |
|---|---|---|---|---|---|---|
| 16 | 500 | 9 | 0.936019 | 0.903453 | 0.070452 | 0.106087 |
| 16 | 500 | 12 | 0.936019 | 0.903897 | 0.076256 | 0.114453 |
| 17 | 500 | 1 | 0.936019 | 0.903453 | 0.076256 | 0.114761 |
| 17 | 500 | 9 | 0.936019 | 0.903897 | 0.076256 | 0.114453 |
| 17 | 500 | 12 | 0.936019 | 0.903897 | 0.076256 | 0.114453 |
| 17 | 500 | 13 | 0.936019 | 0.903897 | 0.076256 | 0.114453 |
| 18 | 500 | 2 | 0.936019 | 0.903897 | 0.076256 | 0.114453 |
| 18 | 500 | 10 | 0.936019 | 0.903897 | 0.076256 | 0.114453 |
| 19 | 500 | 3 | 0.936019 | 0.903897 | 0.076256 | 0.114453 |
| 19 | 500 | 10 | 0.936019 | 0.903897 | 0.076256 | 0.114453 |
| 19 | 500 | 13 | 0.936019 | 0.903897 | 0.076256 | 0.114453 |
| 1 | 500 | 11 | 0.935556 | 0.903104 | 0.081963 | 0.123085 |
| 12 | 500 | 9 | 0.935556 | 0.903533 | 0.083331 | 0.124238 |
| 14 | 500 | 7 | 0.935556 | 0.903251 | 0.078056 | 0.116951 |
| 15 | 500 | 10 | 0.935556 | 0.903089 | 0.078056 | 0.116576 |
| 16 | 500 | 6 | 0.935556 | 0.903533 | 0.078056 | 0.116273 |
| 16 | 500 | 14 | 0.935556 | 0.903533 | 0.083331 | 0.124238 |
| 17 | 500 | 4 | 0.935556 | 0.903533 | 0.078056 | 0.116273 |
| 18 | 500 | 1 | 0.935556 | 0.903533 | 0.083331 | 0.124238 |
| 18 | 500 | 7 | 0.935556 | 0.903533 | 0.078056 | 0.116273 |
| 19 | 500 | 2 | 0.935556 | 0.903533 | 0.078056 | 0.116273 |
| 19 | 500 | 4 | 0.935556 | 0.903533 | 0.083331 | 0.124238 |
| 19 | 500 | 11 | 0.935556 | 0.903533 | 0.078056 | 0.116273 |
| 19 | 500 | 15 | 0.935556 | 0.903533 | 0.083331 | 0.124238 |
| 20 | 500 | 7 | 0.935556 | 0.903533 | 0.078056 | 0.116273 |
| 20 | 500 | 10 | 0.935556 | 0.903533 | 0.078056 | 0.116273 |
| 20 | 500 | 13 | 0.935556 | 0.903533 | 0.083331 | 0.124238 |
| 20 | 500 | 15 | 0.935556 | 0.903251 | 0.083331 | 0.124873 |

| | | | | | | |
|---|---|---|---|---|---|---|
| 16 | 500 | 4 | 0.935093 | 0.903028 | 0.091099 | 0.13476 |
| 17 | 500 | 5 | 0.935093 | 0.903028 | 0.091099 | 0.13476 |
| 8 | 500 | 7 | 0.932315 | 0.898342 | 0.08625 | 0.12943 |
| 10 | 500 | 1 | 0.932315 | 0.898342 | 0.081164 | 0.121805 |
| 10 | 500 | 13 | 0.932315 | 0.898342 | 0.08625 | 0.12943 |
| 11 | 500 | 9 | 0.932315 | 0.898342 | 0.081164 | 0.121805 |
| 12 | 500 | 6 | 0.932315 | 0.898342 | 0.081164 | 0.121805 |
| 12 | 500 | 11 | 0.932315 | 0.897897 | 0.081164 | 0.122074 |
| 13 | 500 | 2 | 0.932315 | 0.898342 | 0.081164 | 0.121805 |
| 14 | 500 | 12 | 0.932315 | 0.898342 | 0.081164 | 0.121805 |
| 15 | 500 | 2 | 0.932315 | 0.898342 | 0.081164 | 0.121805 |
| 18 | 500 | 11 | 0.932315 | 0.898342 | 0.081164 | 0.121805 |
| 13 | 500 | 11 | 0.931852 | 0.897533 | 0.082836 | 0.123764 |
| 14 | 500 | 4 | 0.931852 | 0.897978 | 0.082836 | 0.1235 |
| 17 | 500 | 8 | 0.931852 | 0.897978 | 0.082836 | 0.1235 |
| 17 | 500 | 11 | 0.931852 | 0.897978 | 0.082836 | 0.1235 |
| 18 | 500 | 6 | 0.931852 | 0.897978 | 0.082836 | 0.1235 |
| 20 | 500 | 2 | 0.931852 | 0.897978 | 0.082836 | 0.1235 |
| 20 | 500 | 4 | 0.931852 | 0.897533 | 0.077527 | 0.115767 |
| 20 | 500 | 14 | 0.931852 | 0.897978 | 0.082836 | 0.1235 |
| 14 | 500 | 13 | 0.931389 | 0.897473 | 0.090626 | 0.134058 |
| 18 | 500 | 3 | 0.931389 | 0.897473 | 0.090626 | 0.134058 |
| 18 | 500 | 4 | 0.931389 | 0.897473 | 0.095208 | 0.141022 |
| 20 | 500 | 8 | 0.931389 | 0.897473 | 0.095208 | 0.141022 |
| 20 | 500 | 11 | 0.931389 | 0.897028 | 0.090626 | 0.134299 |
| 14 | 500 | 11 | 0.928611 | 0.892786 | 0.085627 | 0.128489 |
| 19 | 500 | 1 | 0.928148 | 0.891978 | 0.082165 | 0.122743 |
| 19 | 500 | 12 | 0.927685 | 0.891917 | 0.094606 | 0.140124 |

| | | | | | | |
|---|---|---|---|---|---|---|
| 20 | 500 | 3 | 0.927685 | 0.891917 | 0.094606 | 0.140124 |
| 20 | 500 | 9 | 0.927685 | 0.891917 | 0.094606 | 0.140124 |
| 16 | 500 | 10 | 0.924444 | 0.88614 | 0.086392 | 0.129669 |

**Table S3:** Identifiers of individual samples in figure S3.

| **Plant and algal samples** | |
|---|---|
| Individual samples | Taxon (percent abundance) |
| A1 | *Symphyonema karboorum (>90%)* |
| A2 | *Hapalosiphon pumilus (97%)* |
| A3 | *Hapalosiphon pumilus (>90%)* |
| A4 | *Symphyonema karboorum (95%)* |
| A5 | *Symphyonema karboorum (89%)* |
| A6 | *Scytonema mirabile (>95%)* |
| A7 | *Symphyonema karboorum* |
| A8 | *Symphyonema karboorum* |
| A9 | *Hapalosiphon pumilus (91%)* |
| T1 | Unknown |
| T2 | *Casuarina* sp. |
| T3 | *Casuarina* sp. |
| T4 | *Casuarina* sp. |
| T5 | *Casuarina* sp. |
| T6 | *Eucalyptus* sp. |
| T7 | *Eucalyptus* sp. |
| T8 | *Banksia* sp. |
| T9 | *Banksia* sp. |
| T10 | *Banksia* sp. |
| S1 | *Eleocharis difformis* |

| | |
|---|---|
| S2 | *Eleocharis difformis* |
| S3 | *Eleocharis difformis* |
| S4 | *Myriophyllum* sp. |
| S5 | *Myriophyllum* sp. |
| S6 | *Cycnogeton procerus* |
| E1 | *Gahnia* sp. |
| E2 | *Lepironia articulata* |

References

Oksanen J, Blanchet FG, Friendly M, Kindt R, Legendre P, McGlinn D, Minchin PR, O'Hara RB, Simpson GL, Solymos P, Stevens MHH, and ES, and Wagner H (2019) vegan: Community Ecology Package. R package version 2.5-6. https://CRAN.R-project.org/package=vegan

R Core Team (2020) R: A language and environment for statistical computing

Jardine, PE, WD Gosling, BH Lomax, A Julier, and WT Fraser. 2019. 'Data and code for "Chemotaxonomy of domesticated grasses: a pathway to understanding the origins of agriculture"'. https://doi.org/10.6084/m9.figshare.8046395.