bnhcrc.com.au

APPLICATION OF STATISTICAL TECHNIQUES TO PYROLYSIS-GC-MS DATA FROM SOIL TO IDENTIFY THE IMPACT OF FIRE

Malcolm Possell, Mana Gharun, Tina Bell
University of Sydney
Bushfire and Natural Hazards CRC

Version	Release history	Date
1.0	Initial release of document	26/09/2016

Australian Government
Department of Industry, Innovation and Science

Business
Cooperative Research
Centres Programme

Disclaimer:
The University of Sydney and the Bushfire and Natural Hazards CRC advise that the information contained in this publication comprises general statements based on scientific research. The reader is advised and needs to be aware that such information may be incomplete or unable to be used in any specific situation. No reliance or actions must therefore be made on that information without seeking prior expert professional, scientific and technical advice. To the extent permitted by law, the University of Sydney and the Bushfire and Natural Hazards CRC (including its employees and consultants) exclude all liability to any person for any consequences, including but not limited to all losses, damages, costs, expenses and any other compensation, arising directly or indirectly from using this publication (in part or in whole) and any information or material contained in it.

Publisher:
Bushfire and Natural Hazards CRC

August 2016

Citation: M Possell, M Gharun, T Bell (2016) Application of statistical techniques to pyrolysis-GC-MS data from soil to identify the impact of fire, Bushfire and Natural Hazards CRC

Cover: Unburnt vegetation near Canberra. Photo T Bell.

TABLE OF CONTENTS

ABSTRACT 3

END USER STATEMENT 4

INTRODUCTION 5

METHODS 9

1. Site description 9
2. Sampling protocol 9
3. Pyrolysis-GC-MS 10
4. Statistical analysis 10

RESULTS 12

DISCUSSION 16

ACKNOWLEDGEMENTS 19

REFERENCES 20

ABSTRACT

Soil organic matter has strong effects on many soil properties such as water holding capacity, soil structure and stability, nutrient availability and cation exchange capacity. Therefore, characterising soil organic matter is necessary to improve soil management. Pyrolysis coupled to gas chromatography-mass spectrometry (pyr-GC-MS) is one of many techniques that have been successfully used in this characterisation. However, a major limitation of pyr-GC-MS is that generates large amounts of mass-spectrometry data preventing fast, high throughput data analysis. This hinders our ability to identify compounds in complex matrices such as SOM that could be useful for predicting their characteristics. In this study, we aimed to investigate whether it was possible to rapidly identify significant differences among pyr-GC-MS data from soil from burnt and unburnt areas using an unsupervised statistical approach and identify the specific features that cause them. Of nearly 400 useful compounds extracted from the pyr-GC-MS data, only 15 were found to be necessary to classify between burnt and unburnt soil. We discuss how these features could be useful in the classification of soil disturbance such as fire or, potentially, as a quantitative measure of fire impact (intensity or severity).

END USER STATEMENT

Felipe Aires, *Fire and Incident Management, Office of Environment and Heritage, NSW*

The complexity of organic materials found in soils make most methodologies aiming to identify the components too expensive, time consuming and complex and often requires a specialist capable of interpreting the results.

This study demonstrated the potential of using a rapid automated, processing of pyrolysis GC-MS data to identify compounds that are useful in characterising soil from burnt or unburnt plots.

Development of future work should focus on producing operational products capable of using these newly developed technologies to assess post-fire severity and intensity and its impacts on soil carbon. This would allow a more tailored and efficient approach to carbon management by land managers.

INTRODUCTION

Soil organic matter (SOM) is a complex, heterogeneous mixture of organic materials derived from plants and animals at different stages of decomposition and degree of association with the soil mineral matrix (Buurman and Roscoe, 2011). It represents the main terrestrial carbon pool and is an essential component of the global carbon cycle (Eglin *et al*., 2010). Soil organic matter has strong effects on many soil properties such as water holding capacity, soil structure and stability, nutrient availability and cation exchange capacity (Schlesinger, 1986). Consequently, the precise characterisation of SOM is necessary to determine the mechanisms involved in its stabilisation and to predict its dynamics to be able to provide recommendations for improving soil management (Derenne and Quenea, 2015).

Fire affects the carbon balance of terrestrial biomes with immediate release of carbon dioxide (CO_2), carbon monoxide (CO), methane (CH_4), volatile organic compounds (VOCs) and particulate matter (PM) into the atmosphere during the consumption of fuel (Urbanski *et al*., 2009). The carbon balance can also be changed by modifying and redistributing carbon stocks held in partially combusted heavier fuels (i.e. wood converted to charcoal) and in the soil (Volkova and Weston, 2013; Possell *et al*., 2015). Carbon within the SOM can be oxidised during fires with carbon losses varying with fire intensity (Knicker, 2007). Post-fire changes in carbon pools are due to alteration in the activity of microorganisms responsible for decomposition of organic matter and uptake of CO_2 via photosynthesis by vegetation regrowth. Pyrogenic carbon (thermochemically altered carbon derived during combustion) ranges from large pieces of charred biomass, to charcoal, and soot and ash particles often submicron in diameter (Hammes *et al*., 2007). The amount of pyrogenic carbon produced by fires and deposited on soil surfaces is a small proportion of the fuel consumed (typically less than 3%; Jenkins *et al*., 2014) but it represents an important pathway by which carbon can be rendered inert and accumulate in soils over time (Forbes *et al*., 2006). The amount and type of pyrogenic carbon deposited on and later incorporated into the soil is influenced by the type of pre-existing vegetation, the spatial distribution of plant species, the density of plant material as well as fuel, weather conditions, fire intensity and duration (Knicker, 2007). As many of these factors are heterogeneous across the landscape, the characterisation of the effects of fire on SOM becomes a challenging task.

There are a number of analytical methods that have been used to characterise the composition of SOM and these have been reviewed in depth by Derenee and Tu (2014). In brief, these methods are used to examine the nature of chemical functions (e.g. nuclear magnetic resonance (NMR) and Fourier transform infrared (FTIR)); molecular identification of complex organic mixtures (e.g. Fourier transform ion cyclotron resonance mass spectrometry (FT-ICR MS)); spatial elemental and isotopic analysis using x-ray microscopy or secondary ion mass spectrometry (SIMS); and degradation techniques that cleave the organic matter to provide molecular information (Derenne and Quenea, 2015).

Thermal degradation through the use of a pyrolysis unit has been a popular analytical technique and methodologies using pyrolysis have recently been reviewed by Derenne and Quenea (2015). An important thermal degradation technique is pyrolysis coupled to gas chromatography-mass spectrometry (pyr-GC-MS), which allows for identification of the compounds that make up the pyrolysate (the mixture of products generated by pyrolysis). However, despite this being one of the more common methods used, it has some disadvantages. Firstly, the compounds can only be identified if they are amenable to separation by gas chromatography. This amenability is influenced by the type of chromatography column and the temperature and pressure settings used on the column to enact the separation of the pyrolysate (Sáiz-Jiménez, 1994). Another drawback is the difficulty in producing quantitative data because of matrix effects (the combined effect of all components of the sample other than the analyte on the measurement of the quantity) and the large number of standards that are required for extremely complicated pyrolysates (often hundreds of different compounds), most of which are commercially unavailable (Derenne and Quenea, 2015). Thus, semi-quantitative approaches based on the peak area of the two most common ion fragments for a compound have been utilised to allow comparison among chromatograms based on the percentage of the total quantifiable peak area (e.g. Buurman *et al*., 2007).

A further limitation with pyr-GC-MS is the vast amount of data that it generates. A typical electron-impact GC-MS output is represented by two components. The chromatogram displays the mixture as separated by the GC, and each peak on the chromatogram normally corresponds to the elution of a distinct molecule, which is characterised by retention time. For each point on the GC chromatogram, a mass spectrum is obtained by fragmentation, using electron impact, in the mass-spectrometer chamber. A mass spectrum is represented by a histogram displaying the intensity of each fragment as a ratio of its mass-to-charge. Each chromatogram often produces hundreds of peaks with their own retention time and ions over a range of masses and intensities leading to the identification of several hundreds of compounds (Buurman *et al*., 2007; Buurman and Roscoe, 2011; Schellekens *et al*., 2009; 2011; Vancampenhout *et al*., 2008; 2009). Although the generation of considerable amounts of mass-spectrometry data is not unique to pyr-GC-MS, it presents a bottleneck to fast, high throughput data analysis. Attempts at automated analyses of pyr-GC-MS data, although successful, have demonstrated that how the analyses are done, both in terms of software and hardware, can have a significant impact on the detection and hence quantification of compounds (Wenig and Odermatt, 2010; Tolu *et al*., 2015). The amount of data generated from a small number of samples can also limit subsequent statistical analysis of datasets through techniques such as principal components analysis because there are too few samples relative to the number of variables (compounds) identified. This limits the ability to identify compounds in complex matrices such as SOM that could be useful for predicting their characteristics. However, once a suite of compounds have been identified, these compounds could be concentrated on providing a more streamlined, quantifiable and potentially faster analysis by pyr-GC-MS.

Despite the apparent limitations of pyr-GC-MS, it has proved to be a useful method for evaluating the impact of heat on certain components of SOM such as cellulose, pectin, lignin, protein and amino acids. Furans, pyranones, anhydrosugars and 5-hydroxymethylfurfural are the major products of pyrolysis of cellulose at temperatures <300 °C (Ralph and Hatfield, 1991; Bassilakis *et al*., 2001). As temperatures increase, these compounds are replaced by polycyclic aromatic hydrocarbons as the pyrolysis of lignin produces mainly substituted methoxyphenols, with the highest yields achieved at pyrolysis temperatures between 500 and 600 °C (Knicker, 2007). Amino acids and proteins generate pyrrole-type and nitrogen-containing heterocyclic compounds when pyrolysed at low temperatures (200–300 °C) (Chiavari and Galletti, 1992; Britt *et al*., 2004). As pyrolysis temperatures increase above 500 °C, peptides generates polynuclear aromatic structures containing nitrogen (Sharma *et al.* 2003). In studies where SOM has been analysed to determine the effects of fire on soil, the maximum temperature that the soil reaches has been shown to have an effect on the amount and composition of the SOM. For instance, de la Rosa *et al*. (2008) showed that when the temperature of SOM is raised to 520 °C there is an enrichment of heterocyclic nitrogen compounds and aliphatic nitriles. This has also been demonstrated in SOM from *Eucalyptus* and *Pinus pinaster* forests (de la Rosa *et al*., 2012). The compounds produced at higher temperatures have chemically-bound carbon and nitrogen that is far less available for plant and microbial uptake. The study of de la Rosa *et al*. (2012) also revealed a reduction in the amount of isoprenoids in soil (organic compounds of plant origin) due to fire. These few studies highlight that there is great complexity in the characterisation of the effect of fire on SOM. However, there are particular compounds that may be useful as marker compounds for how intense or severe a fire was in a similar way to how methoxy-phenols are used as marker compounds for woodsmoke (Hawthorne *et al*., 1988).

The type of pyrogenic carbon generated during a fire can be used to trace fire history at a particular site. For instance, partially-charred litter from a Florida Scrub Oak ecosystem was evident in the soil organic matter for at least ten years after fire (Alexis *et al*., 2012). Charred solid residues, often referred to as black carbon, are considered to be one of the most recalcitrant forms of organic carbon (Schmidt and Noack, 2000). The slow decomposition of this black carbon has been useful in investigating the fire history of colluvial soils in north-west Spain for the past 8,500 years (Kaal *et al*., 2008a; 2008b; 2008c). This series of studies and that of Kaal *et al*. (2009) demonstrate that pyr-GC-MS of black carbon predominantly produces benzene, toluene, C_2-benzenes, polyaromatic hydrocarbons and benzonitriles. However, these compounds can also be found in the pyrolysate of soil organic matter, albeit at different proportions (de la Rosa *et al*., 2008; 2012). It is clear that a robust method is needed to differentiate between the origins of these compounds when using pyr-GC-MS.

In this study, we aimed to investigate whether it was possible to rapidly identify significant differences among pyr-GC-MS chromatograms using an unsupervised approach that does not require manual scrutiny of all peaks in all of the chromatograms nor complicated optimisation of software or hardware. We applied this approach to pyr-GC-MS chromatograms of soil from burnt and unburnt areas and then used an ensemble learning method to identify the features of the chromatograms that caused the differences. We discuss how these features could be useful in the classification of soil disturbance such as fire or, potentially, as a quantitative measure of fire impact (intensity or severity).

METHODS

1. SITE DESCRIPTION

Sites (n = 4) were identified within prescribed burns conducted in 2015 in the Australian Capital Territory, Australia (Table 1). Elevation of the sites ranged from 760–1300 m above sea level and the climate of the study area is broadly described as cool temperate. Sites were classified as low woodland and tall open forest dominated by Brittle Gum (*Eucalyptus mannifera*), Red Box (*E. polyanthemos*), White Gum (*E. rossi*), Apple Box (*E. bridgesiana*), Narrow-leaved Peppermint (*E. radiata*) and Broad-leaved Peppermint (*E. dives*). The understorey consisted mostly of Bracken (*Pteridium esculentum*), Grey Tussock Grass (*Poa sieberiana*) and Native Blackthorn (*Bursaria spinosa* subsp. *lasiophylla*). Soils at all sites are categorised as Kurosols according to the Australian Soil Classification (Isbell, 2016).

TABLE 1 DESCRIPTION OF THE STUDY SITES IN THE AUSTRALIAN CAPITAL TERRITORY, AUSTRALIA AND PRESCRIBED BURNING OPERATIONS. A.S.L. = ABOVE SEA LEVEL; N/A = NO RECORDED FIRES SINCE RECORDS BEGAN IN 1902 (NSW LAND AND PROPERTY, 2016).

Burn name	State	Longitude	Latitude	Mean elevation (m a.s.l.)	Ignition date	Date of previous burn
Googong	ACT	-35.52	149.29	767	11/3/2015	N/A
Tidbinbilla	ACT	-35.46	148.90	869	17/3/2015	January 2003
Cotter	ACT	-35.60	148.80	1234	30/3/2015	January 2003
Lone Pine	ACT	-35.88	148.94	1271	-	N/A

2. SAMPLING PROTOCOL

At each site, six circular plots (22.5 m radius; 1590.4 m^2) were established at 20–50 m from the access road and at least 500 m apart. Three plots were located in the area burnt by the prescribed burn and three plots were located nearby in an adjacent unburnt area. Factors such as spatial proximity, dominant canopy species, tree size and density distribution, slope and aspect were considered before selecting the adjacent burnt and unburnt plots to ensure that biophysical differences in plot properties were minimised. Within each circular plot, four circular subplots (radius = 5 m) located 5 m along the north-south and east-west axes of each of the larger plots, as measured from the centre point, were established as described in Possell *et al.* (2015).

For consistency among sites, soil samples were taken at a random point within the north subplot of each plot. Surface material, including ash and charcoal in burnt plots and leaf litter (fine fuel) in unburnt plots, was carefully removed to expose the underlying soil mineral layer. The top 10 cm of soil was collected using a steel core (4.37 cm diameter x 10 cm depth). Soil samples were stored in zip locked bags, cooled and transferred to the laboratory and sieved to 2 mm before air-drying for several days and ground.

3. PYROLYSIS-GC-MS

Pyrolysis was done using a Gerstel pyrolysis module for the Gerstel Thermal Desorption Unit (TDU; Gerstel, Mülheim an der Ruhr, Germany). Approximately 5 mg of each sample was purged with ultra-high purity helium (BOC Ltd, North Ryde, NSW, Australia) at 60 °C for 3 minutes to eliminate air and residual moisture from the sample. Samples were heated by the TDU at 12 °C s^{-1} to 300 °C before pyrolysis at 600 °C for 20 seconds. Pyrolysis products were carried by the helium through a programmed temperature vaporisation (PTV) inlet (CIS-4; Gerstel) installed in an Agilent 7890 GC (Agilent Technologies Pty Ltd, Mulgrave, Australia). The PTV inlet was held at 300 °C with a 25:1 split ratio. Pyrolysis products were separated on a HP-5MS capillary column (30 m x 0.25 mm, 0.25 µm film thickness; Agilent) which was connected to a two-way splitter with makeup gas (Agilent). A restrictor column of deactivated fused silica (1.44 m x 0.18 mm; Agilent), connected to the outlet of the splitter, transferred the pyrolysis products to a mass selective detector (Model 5975C; Agilent). Ultra-high purity helium was used as carrier gas (flow rate through the HP5-MS column was 2.3 ml min^{-1} and 4 ml min^{-1} through the restrictor column). The initial oven temperature of the GC was 40 °C, held for 1 minute, then heated at a rate of 5 °C min^{-1} to 300 °C, and held isothermal for 15 minutes. The temperature of the GC-MS interface was 280 °C, the MS ion source 230 °C and the quadrupole 150 °C. The detector, in electron impact mode (70 eV), scanned the range of 45–650 *m/z*. Operation of the GC-MS was controlled by Agilent Chemstation (version E.02.01.117) and the pyrolysis module and TDU by Maestro (version 1.4.26.40/3.5; Gerstel).

4. STATISTICAL ANALYSIS

Post-processing of mass-spectral data was performed using MSeasy (version 5.5.3; Nicole *et al*., 2012). MSeasy is a package for R (version 3.1.2; R Core Team, 2015) that performs unsupervised data mining on GC-MS data. This program is insensitive to shift in retention times and detects putative compounds within complex metabolic mixtures through the clustering of mass spectra. Retention times were used after clustering, together with validation criteria, namely the Silhouette Width (Rousseeuw, 1987) and Dunn's Index (Dunn, 1974), for quality control of putative compounds. The package generates a fingerprinting or profiling matrix compatible with a mass spectral search program and library. Identification of the compounds corresponding to the mass spectra of the clusters was performed using NIST08 mass spectral library in NIST MS Search (NIST MS Search v.2.0f; NIST, Gaithersburg, MD). Identification was made using a combination of the library's calculated match factor (where 900 or greater is an excellent match; 800-899 is a good match; 700-799 a fair match and less than 600 a poor match; NIST, 2008) and visual comparison of the mass-spectra.

To test for differences in the fingerprinting matrix between the burnt and unburnt sites, a chemical dissimilarity matrix was calculated between pairs of individuals using the Manhattan distance

$$D_{x,y} = \sum_{i=1}^{Nm} \left| x_i - y_i \right| \qquad (1)$$

where x and *y* are two distinct individuals and N_m is the total number of compounds in the dataset. A non-parameteric multivariate analysis of variance (pMANOVA; Anderson, 2001) was used to test for differences between the fingerprinting matrices of burnt and unburnt sites with the pairings from a particular site nested together. This analysis calculates a ''pseudo-F'' ratio analogous to Fisher's F-ratio for each factor and their interactions based on the Manhattan distance matrix. The partial squared coefficient of correlation (R^2) is the percentage of variance in the chemical dissimilarity matrix that is explained by the factor, and the significance (*P* values) were calculated by performing 1000 permutations on the rows or columns of the matrices. The chemical dissimilarity matrix and pMANOVA were calculated using the package 'vegan' in R (Oksanen *et al*., 2014). To check that the results were not a function of heterogeneity of group dispersions (variances), a multivariate analogue of Levene's test for homogeneity of variances was applied to the distance measures using the 'betadisper' function within the 'vegan' package (Anderson, 2006).

To identify the variables (compounds) contributing to any observed differences in the chromatograms among the burnt and unburnt sites, Random Forests analysis (Breiman, 2001) was used for the classification using the 'randomForest' package in R (v. 4.6-10; Liaw and Wiener, 2002). We used a variable selection procedure to identify the important compounds (Genuer *et al*., 2010). This method is based on the unscaled permutation importance calculated by permuting each predictor in turn and using the difference in the prediction error (out-of-bag (OOB) error) before and after permutation as a measure of variable importance (Liaw and Wiener, 2002). The approach of Genuer *et al*. (2010) identifies a set of classifiers suitable for model interpretation by:

a) ranking all predictors using the unscaled permutation importance (averaged over 999 repetitions) calculated by Random Forests;
b) discarding noise predictors by fitting a regression tree (Therneau *et al*., 2015) to the curve of the plot of standard deviations of importance measures ordered by their mean importance. As the variability of the predictor's importance is larger for true predictors compared to noise predictors, predictors with a standard deviation less than the smallest predicted value of the regression tree model (the threshold) are discarded;
c) computing OOB error for models (using default parameters for the random forest models) starting with a model with the most important variable and adding predictors sequentially in the order of their ranking (nested models) and;
d) selecting the model with the smallest OOB error.

RESULTS

Pyrolysis-GC-MS of soil from burnt and unburnt plots produced chromatograms containing up to 388 peaks. *Prima facie* comparison of chromatograms, such as those in Figure 1, indicate that there is little difference between the pyrolysate derived from the soil of burnt and unburnt sites because of the similarity in the retention times and amplitude of the peaks present. However, due to the potential for co-elution of compounds from the GC column, the number of peaks identified does not translate to the number of compounds found. The use of MSeasy in this study identified 642 clusters (putative compounds) of which 371 were identified as having met the validation criteria. To test for differences in the fingerprinting matrix generated by MSeasy, between the burnt and unburnt sites, a chemical dissimilarity matrix was calculated between pairs of burnt and unburnt sites. Permutational MANOVA of the chemical dissimilarity matrix identified a significant difference between burnt and unburnt sites (P = 0.003) and that burning accounted for 14.3% of the difference (Table 2). This statistical approach does not identify which compounds cause the difference, requiring further examination of the data by other techniques.

FIGURE 1 EXAMPLE CHROMATOGRAMS (USING TOTAL ION COUNT; TIC) GENERATED BY PYROLYSIS-GC-MS FOR SOIL FROM (A) UNBURNT AND (B) BURNT PLOTS COLLECTED FROM TIDBINBILLA (SEE TABLE 1).

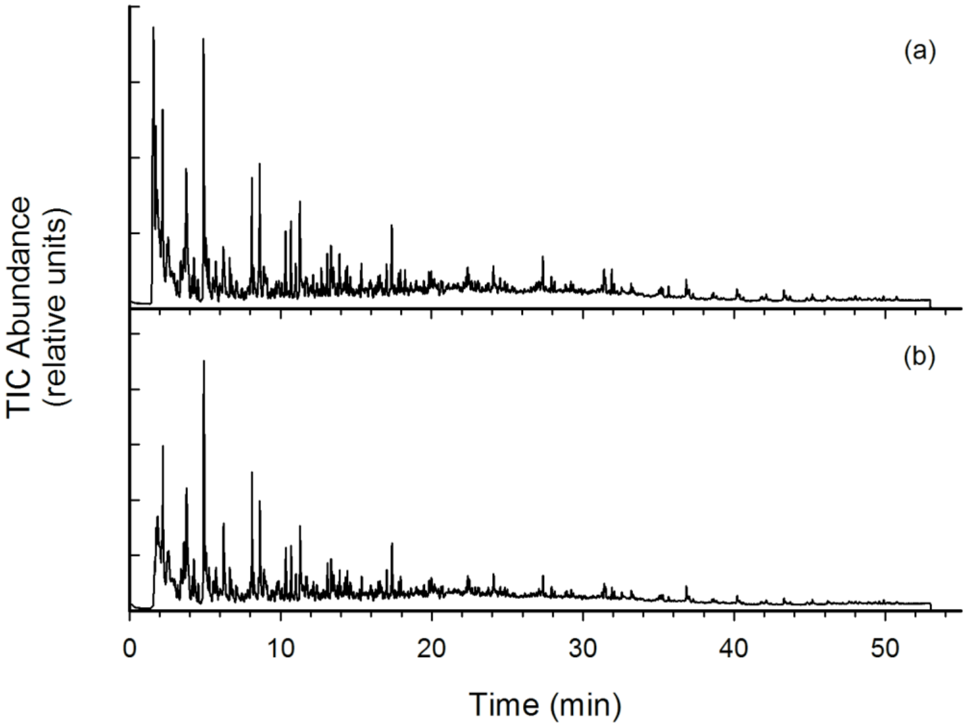

TABLE 2: NON-PARAMETRIC pMANOVA OF THE EFFECT OF BURNING ON THE PYR-GC-MS PYROLYSATE OF THE SOIL SAMPLES ANALYSED IN THIS STUDY. *R*^{2} IS THE PERCENTAGE OF VARIANCE EXPLAINED BY THE BURNING.

	Degrees of freedom	Sum of squares	Mean Squares	F (Model)	*R*^{2}	*P* (>F)
Treatment	1	7610	7610.3	3.6673	0.14288	0.003
Residuals	22	45654	2075.2		0.85712	
Total	23	53265			1.00000	

Random Forests analysis (Breiman, 2001) was used to identify compounds that classify burnt or unburnt sites. Of the 371 compounds identified by MSeasy, the variable selection procedure of Genuer *et al*. (2010), using Random Forests, produced a list of compounds ordered by their importance (Figure 2a). Applying a Classification and Regression Tree analysis (CART; Therneau *et al*., 2015) to the standard deviations of the compound's importance identified a threshold standard deviation value of approximately 0.0004 (Figure 2b). When this threshold value was applied to the importance measure of the 371 compounds, those with a standard deviation of the importance measure less than the threshold were eliminated. This reduced the number of compounds to 259. Random Forest analysis on these 259 compounds, starting with a model with the most important compound and adding compounds sequentially in the order of their ranking (nested models), produced a model of 15 compounds that had the smallest prediction (OOB) error of all the nested models (33.33%; Table 3; Figure 3). These 15 compounds were tentatively identified against the NIST08 mass spectral database (NIST, Gaithersburg, MD) and their identities are listed in Table 4.

TABLE 3: CLASSIFICATION ERROR OF BURNT AND UNBURNT SOILS BY RANDOM FOREST ANALYSIS WHEN USING FIFTEEN PREDICTOR COMPOUNDS PRODUCED BY PYR-GC-MS OF THE SOIL.

Soil category	Predicted count		Classification error (%)
	Burnt	Unburnt	
Burnt	9	3	0.25
Unburnt	5	7	0.42

FIGURE 2: THE IMPORTANCE (A) AND STANDARD DEVIATION OF THE IMPORTANCE (B) OF THE COMPOUNDS IDENTIFIED BY MSEASY (NICOLE *ET AL*., 2012) AVERAGED OVER 999 RANDOM FOREST PERMUTATIONS. THE RED LINE IN PANEL (B) IS THE THRESHOLD VALUE DETERMINED BY CLASSIFICATION AND REGRESSION TREE ANALYSIS (THERNEAU *ET AL*., 2015) ABOVE WHICH VARIABLES ARE RETAINED FOR FURTHER ANALYSIS.

FIGURE 2: THE IMPORTANCE (A) AND STANDARD DEVIATION OF THE IMPORTANCE (B) OF THE COMPOUNDS IDENTIFIED BY MSEASY (NICOLE *ET AL.*, 2012) AVERAGED OVER 999 RANDOM FOREST PERMUTATIONS. THE RED LINE IN PANEL (B) IS THE THRESHOLD VALUE DETERMINED BY CLASSIFICATION AND REGRESSION TREE ANALYSIS (THERNEAU *ET AL.*, 2015) ABOVE WHICH VARIABLES ARE RETAINED FOR FURTHER ANALYSIS.

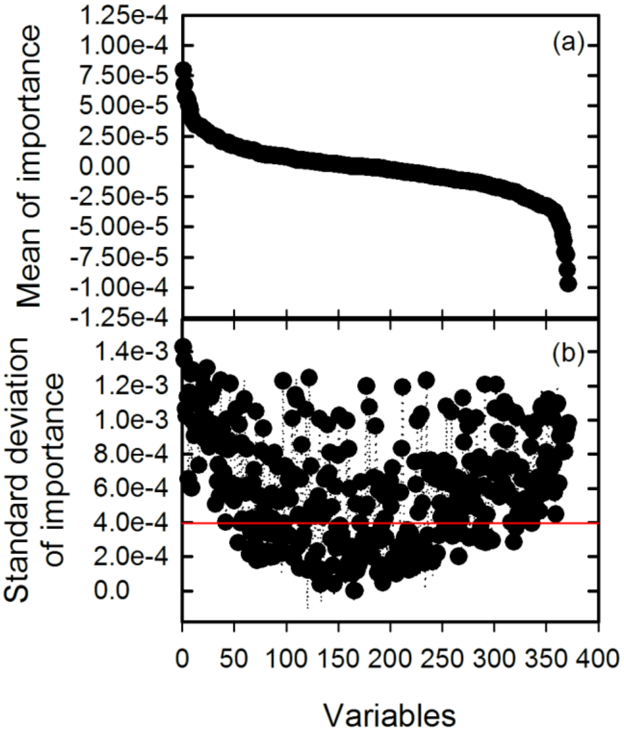

FIGURE 3: OUT-OF-BAG (OOB) ERROR RATE (PREDICTION ERROR) FOR THE NESTED RANDOM FOREST ANALYSIS AFTER ELIMINATING VARIABLES (COMPOUNDS) USING CLASSIFICATION AND REGRESSION TREE ANALYSIS (THERNEAU *ET AL.*, 2015).

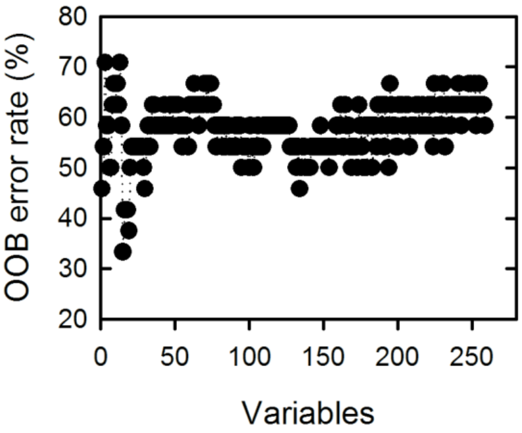

TABLE 4: TENTATIVE IDENTIFICATION OF 15 CLUSTERS (COMPOUNDS) CLASSED AS IMPORTANT IN DISTINGUISHING BETWEEN SOIL FROM BURNT AND UNBURNT PLOTS WHEN USING VARIABLE SELECTION BY RANDOM FORESTS. EXAMINATION OF THE MASS SPECTRA WAS PERFORMED USING NIST08 MASS SPECTRAL LIBRARY IN NIST MS SEARCH (NIST MS SEARCH *V*.2.0F; NIST, GAITHERSBURG, MD). RETENTION TIME (RT) ± ONE STANDARD DEVIATION (S.D.). MATCH FACTOR IS THE SCORE (OUT OF 1000) THAT THE MASS SPECTRA IDENTIFIED FOR THE CLUSTER IS HOMOLOGOUS TO THE ONE CONTAINED IN THE NIST08 MASS SPECTRAL LIBRARY.

Cluster ID	Importance (× 10^{-5})	Compound name	Synonyms (if applicable)	Mean RT (± s.d.)	Match Factor
117	7.9578	1-methyl-4-(1-methylethyl)-2,3-Dioxabicyclo[2.2.2]oct-5-ene	Ascaridol	18.15 ± 0.05	616
26	6.7576	Ethylbenzene		5.540 ± 0.03	768
29	5.7148	Methylenecyclooctane		6.000 ± 0.03	674
43	5.6922	(Z,Z,Z)-9,12,15-Octadecatrienoic acid	Linolenic acid	7.670 ± 0.07	716
76	5.5032	4-Aminobenzyl cyanide		12.04 ± 0.03	603
535	5.0336	4-(2,6,6-trimethyl-2-cyclohexen-1-yl)-3-Buten-2-ol		15.41 ± 0.02	624
97	4.9275	4,7-dimethyl-benzofuran		15.22 ± 0.06	640
155	4.6378	Butanoic acid, 3-methyl-, 3,7-dimethyl-2,6-octadienyl ester, (E)-	Geranyl isovalerate	22.66 ± 0.01	681
54	4.0222	2-methyl-cyclohexanone,		9.480 ± 0.06	636
157	3.8161	1β-(3-methyl-1,3-butadienyl)-2α,6β-dimethyl-3β-acetoxy-Bicyclo[4.1.0]heptan-2-ol		22.89 ± 0.06	672
165	3.6373	4-methoxy-2-phenyl-cycloheptene		23.73 ± 0.01	628
15	3.5774	3-Heptyn-1-ol		3.600 ± 0.07	676
304	3.3937	6,6-dimethyl-bicyclo[3.1.1]hept-2-ene-2-ethanol	Homomyrtenol	10.55 ± 0.02	676
114	3.3858	2-methyl-naphthalene		17.81 ± 0.01	660
87	3.3727	4-(2,5-Dihydro-3-methoxyphenyl)butylamine		13.91 ± 0.07	692

DISCUSSION

The nature of electron-impact mass-spectrometric techniques, such as pyr-GC-MS, produces large, multi-dimensional datasets of retention times, mass-to-charge-ratios and fragment intensities. This requires either time-consuming examination of individual chromatograms to extract mass-spectra of individual compounds or the use of software to deconvolute chromatograms using automated algorithms. The methodology employed in this study was designed to use rapid automated, unsupervised data mining of pyrolysis GC-MS chromatograms to identify compounds that are useful in characterising soil from burnt or unburnt plots.

Unsupervised data mining using the clustering method of MSeasy (Nicole *et al.*, 2012) identified nearly 400 useful compounds which, when compared as a whole between soils from burnt and unburnt plots using pMANOVA, showed that there were significant differences in the pyrolysate generated from those soils. Prescribed burning accounted for about one-seventh of the difference between the pyrolysates of the soil from burnt and unburnt plots. This highlights the complexity of organic materials within soil (i.e. indicated by the high number of compounds identified), even from sites putatively similar composition. However, by being able to distinguish between the burnt and unburnt soils without time-consuming or complex sample preparation (e.g. solvent extraction or acid digestion), it shows that pyr-GC-MS is an appropriately sensitive and relatively quick technique for this characterisation process.

Permutational MANOVA of the chemical dissimilarity matrix data generated from the chromatograms of the soil from burnt and unburnt plots is useful when comparing the resulting pyrolysates as a whole but does not identify which compounds are the main causes of the dissimilarity. Random Forest analysis is an ensemble learning method that can be used for classification and regression and it utilises variable importance selection (where, in this study, we define variables as compounds) to do this (Breiman, 2001). In this study, we used the variable selection procedure proposed by Genuer *et al*. (2010). This approach was chosen because attempts to use a selection strategy based on the recursive elimination of compounds, as described by Diaz-Uriarte *et al*. (2006), consistently produced prediction error rates greater than ones where the variables were selected at random (data not shown). This may be a function of how the method of Diaz-Uriarte *et al*. (2006) works by eliminating 20% of the compounds having the smallest importance and building a new forest with the remaining compounds. The proportion of compounds to be eliminated is an arbitrary parameter and does not depend on the data (Genuer *et al*., 2010). The method of Genuer *et al*. (2010) resulted in the best classification (smallest prediction error) of chromatograms from the soil from burnt and unburnt plots by using just 15 predictor compounds. However, the prediction error indicated that, overall, there was a one-in-three chance of misclassifying the soil category (burnt or unburnt) when using these predictor compounds (Figure 3). Misclassification was greater for soil from unburnt plots compared to soil from burnt plots (Table 3). This could be a consequence of: (a) the limited number of samples (soil from 12 burnt plots and 12 unburnt plots) from which to try and identify important variables, hence a single misclassification can cause a disproportionally large change in the OOB error rate; (b) trying to classify between two categories (burnt or unburnt) when burning only accounted for a small proportion of the difference in combustion products (pyrolysates); and (c) soil from unburnt plots containing pyrogenic material from previous fires. Of the four sites, two were previously burnt in 2003 and two have no recorded fire history (Table 1). Presumably this 'old' pyrogenic material would still contribute a signal, albeit smaller than for soil from recently burnt area. However, in this study, the signal may have still been strong enough to get a misclassification by Random Forest analysis. Therefore, there is a need to understand how marker compounds change over time after fire.

When attempting to classify between soils from burnt and unburnt areas based upon their pyrolysates, a diverse suite of compounds were identified as being important. Cluster IDs 54, 117, 155, 157, 304 and 535, as defined by MSeasy, were tentatively identified as being constituent compounds of essential oils. For example, ascaridol is a bicyclic monoterpene with a peroxide bridging group. These compounds would reflect origins from biomass incorporated into the soil or from root exudations that have adsorbed to soil particles (Lin *et al*., 2007). As demonstrated by de la Rosa *et al*. (2012), the presence of organic compounds of plant origin in soil is significantly reduced by fire. Therefore, a strong presence of these organic compounds in soil would lead to an unburnt classification while soil with very few of these compounds would be classified as burnt. The remaining compounds identified as important in the classification are known products of pyrolysis. For instance, ethylbenzene, as a monocyclic aromatic hydrocarbon, is a pyrolysis product from many biomass components (Kaal *et al*., 2014). Polyaromatic hydrocarbons such as 2-methyl-napthalene can originate from black carbon and have the potential to differentiate between soil from burnt and unburnt plots (Kaal *et al*., 2009). The presence of dimethyl benzofuran in soil is indicative of thermally-altered soil organic matter (Kaal *et al*., 2014). Similarly, the presence of N-containing compounds (e.g. 4-aminobenzyl cyanide and 4-(2,5-Dihydro-3-methoxyphenyl)butylamine) may be indicative of pyrolysis of amino-acids or proteinaceous material (Chiavari and Galletti, 1992) but is also consistent with the enrichment in heterocyclic nitrogen compounds with the heating of soil organic matter (de la Rosa *et al*., 2012). Although we would expect prior to the analyses to have compounds of plant origin and those known to be affected by fire in the list of important classification compounds, the clustering algorithms of MSeasy and the Random Forests procedure do not make that *a priori* assumption. This demonstrates the potential of the methodology described in this study to rapidly and rationally identify differences and the features that cause them in complex matrices such as SOM.

In this study, we demonstrate that it is possible to use an unsupervised data mining approach to successfully identify several statistically important compounds from pyr-GC-MS chromatograms that can be used to classify between soils that have been collected from recently burnt or unburnt sites. This methodology was able to select these compounds even when prescribed burning contributed to a small proportion of the overall difference between the soils. Furthermore, the compounds identified by this approach were compounds known to be affected by or produced by burning. However, further work is required to answer the following questions:

a) Can the results of this classification be used to make accurate predictions of whether soil has been burnt using a larger dataset of soil pyr-GC-MS chromatograms, including those from high intensity bushfire or different soil or vegetation types?

b) Does burning change the concentration of marker compounds in soil in a predictable way, for example in relation to fire intensity or severity, so that they can be used as a quantitative measure of fire impact?

c) How does time after fire affect the concentration of marker compounds?

The rationale for this study and our future attempts to answer questions such as these is to develop a reliable, quantitative method for post-fire assessment of fire severity and intensity on soil that can be used and interpreted by land managers. Soil burn assessment protocols are available (Keeley, 2009) but these generally include categories or indices of fire intensity (e.g. loss of surface litter or soil organic layer, deposition of ash and charred organic matter) and are largely subjective and open to interpretation by the assessor.

ACKNOWLEDGEMENTS

This project was funded by the Bushfire and Natural Hazards Cooperative Research Centre. We thank Adam Leavesley and his staff from Parks and Conservation ACT for help in site selection and field work. We thank Ariana Iaconis, Marco Harbusch and Bonnie Cannings from the University of Sydney for assistance in the field.

REFERENCES

1. Alexis MA, Rasse DP, Knicker H, Anquetil C, Rumpel C (2012) Evolution of soil organic matter after prescribed fire: a 20-year chronosequence. *Geoderma*, **189**, 98–107.
2. Anderson MJ (2001) A new method for non-parametric multivariate analysis of variance. *Austral Ecology*, **26**, 32–46.
3. Anderson MJ (2006) Distance-based tests for homogeneity of multivariate dispersions. *Biometrics*, **62**, 245–253.
4. Bassilakis R, Carangelo RM, Wojtowicz MA (2001) TG-FTIR analysis of biomass pyrolysis. *Fuel*, **80**, 1765–1786.
5. Breiman L (2001) Random Forests. Machine Learning. **45**, 5–32.
6. Britt PF, Buchanan AC, Owens CV, Skeen JT (2004) Does glucose enhance the formation of nitrogen containing polycyclic aromatic compounds and polycyclic aromatic hydrocarbons in the pyrolysis of proline? *Fuel*, **83**, 1417–1432.
7. Buurman P, Roscoe R (2011) Different chemical composition of free light, occluded light and extractable SOM fractions in soils of Cerrado and tilled and untilled fields, Minas Gerais, Brazil: a pyrolysis-GC/MS study. *European Journal of Soil Science*, **62**, 253–266.
8. Buurman P, Peterse F, Martin GA (2007) Soil organic matter chemistry in allophanic soils: a pyrolysis-GC/MS study of a Costa Rican Andosol catena. *European Journal of Soil Science*, **58**, 1330–1347.
9. Chiavari G, Galletti GC (1992) Pyrolysis-gas chromatography mass-spectrometry of amino-acids. *Journal of Analytical and Applied Pyrolysis*, **24**, 123–137.
10. de la Rosa JM, Faria SR, Varela ME, Knicker H, Gonzalez-Vila FJ, Gonzalez-Perez JA, Keizer J (2012) Characterization of wildfire effects on soil organic matter using analytical pyrolysis. *Geoderma*, **191**, 24–30.
11. de la Rosa JM, Gonzalez-Perez JA, Gonzalez-Vazquez R, Knicker H, Lopez-Capel E, Manning DAC, Gonzalez-Vila FJ (2008) Use of pyrolysis/GC-MS combined with thermal analysis to monitor C and N changes in soil organic matter from a Mediterranean fire affected forest. *Catena*, **74**, 296–303.
12. Derenne S, Quenea K (2015) Analytical pyrolysis as a tool to probe soil organic matter. *Journal of Analytical and Applied Pyrolysis*, **111**, 108–120.
13. Derenne S, Tu TTN (2014) Characterizing the molecular structure of organic matter from natural environments: an analytical challenge. *Comptes Rendus Geoscience*, **346**, 53–63.
14. Diaz-Uriarte R, Alvarez de Andrés S (2006) Gene selection and classification of microarray data using random forest. *BMC Bioinformatics*, **7**, 3.
15. Dunn JC (1974) Well separated clusters and fuzzy partitions. *Journal of Cybernetics*, **4**, 95–104.
16. Eglin T, Ciais P, Piao SL, Barre P, Bellassen V, Cadule P, Chenu C, Gasser T, Koven C, Reichstein M, Smith P (2010) Historical and future perspectives of global soil carbon response to climate and land-use changes. *Tellus Series B-Chemical and Physical Meteorology*, **62**, 700–718.
17. Forbes MS, Raison RJ, Skjemstad JO (2006) Formation, transformation and transport of black carbon (charcoal) in terrestrial and aquatic ecosystems. *Science of the Total Environment*, **370**, 190–206.
18. Genuer R, Poggi JM, Tuleau-Malot C (2010) Variable selection using random forests. *Pattern Recognition Letters*, **31**, 2225–2236.
19. Hammes K, Schmidt MWI, Smernik RJ, Currie LA, Ball WP, Nguyen TH, Louchouarn P, Houel S, Gustafsson O, Elmquist M, Cornelissen G, Skjemstad JO, Masiello CA, Song J, Peng P, Mitra S, Dunn J C, Hatcher PG, Hockaday, WC, Smith DM, Hartkopf-Froeder C, Boehmer A, Luer B, Huebert BJ, Amelung W, Brodowski S, Huang L, Zhang W, Gschwend PM, Flores-Cervantes DX, Largeau C, Rouzaud JN, Rumpel C, Guggenberger G, Kaiser K, Rodionov A, Gonzalez-Vila FJ, Gonzalez-Perez JA, de la Rosa JM, Manning DAC, Lopez-Capel E, Ding L (2007) Comparison of quantification methods to measure fire-derived (black/elemental) carbon in soils and sediments using reference materials from soil, water, sediment and the atmosphere. *Global Biogeochemical Cycles*, **21**, GB3016.
20. Hawthorne SB, Miller DJ, Barkley RM, Krieger MS (1988) Identification of methoxylated phenols as candidate tracers for atmospheric wood smoke pollution. Environmental Science and Technology 22: 1191–1196.
21. Isbell RF (2016) The Australian Soil Classification (2nd edition). CSIRO Publishing, Clayton South, VIC, Australia, 152 pp.
22. Jenkins ME, Bell TL, Norris J, Adams MA (2014) Pyrogenic carbon: the influence of particle size and chemical composition on soil carbon release. *International Journal of Wildland Fire*, **23**, 1027–1033.
23. Kaal J, Martinez-Cortizas A, Eckmeier E, Costa Casais M., Santos Estevez M, Criado Boado F (2008a) Holocene fire history of black colluvial soils revealed by pyrolysis-GC/MS: a case study from Campo Lameiro (NW Spain). *Journal of Archaeological Science*, **35**, 2133–2143.
24. Kaal J, Martinez-Cortizas A, Buurman P, Boado FC (2008b) 8000 yr of black carbon accumulation in a colluvial soil from NW Spain. *Quaternary Research*, **69**, 56–61.
25. Kaal J, Martinez-Cortizas A, Nierop KGJ, Buurman P (2008c) A detailed pyrolysis-GC/MS analysis of a black carbon-rich acidic colluvial soil (Atlantic ranker) from NW Spain. *Applied Geochemistry*, **23**, 2395–2405.
26. Kaal J, Martinez-Cortizas A, Nierop KGJ (2009) Characterisation of aged charcoal using a coil probe pyrolysis-GC/MS method optimised for black carbon. *Journal of Analytical and Applied Pyrolysis*, **85**, 408–416.
27. Kaal J, Rumpel C (2009) Can pyrolysis-GC/MS be used to estimate the degree of thermal alteration of black carbon? *Organic Geochemistry*, **40**, 1179–1187.
28. Kaal J, Lantes-Suarez O, Cortizas AM, Prieto B, Martinez MPP (2014) How useful is pyrolysis-GC/MS for the assessment of molecular properties of organic matter in archaeological pottery matrix? An exploratory case study from north-west Spain. *Archaeometry*, **56**, 187–207.

29. Keeley JE (2009) Fire intensity, fire intensity and burn severity: a brief review and suggested usage. *International Journal of Wildland Fire*, **18**, 116–126.
30. Knicker H (2007) How does fire affect the nature and stability of soil organic nitrogen and carbon? A review. *Biogeochemistry*, **85**, 91–118.
31. Liaw A, Wiener M (2002) Classification and regression by randomForest. *R News*, **2**, 18–22.
32. Lin C, Owen SM, Penuelas J (2007) Volatile organic compounds in the roots and rhizosphere of *Pinus* spp. *Soil Biology and Biochemistry*, **39**, 951–960.
33. Nicole F, Guitton Y, Courtois E, Moja S, Legendre L, Hossaert-McKey M (2012) MSeasy: unsupervised and untargeted GC-MS data processing. *Bioinformatics*, **28**, 2278–2280.
34. NIST (2008) NIST standard reference database 1A: NIST/EPA/NIH mass spectral library (NIST 08) and NIST mass spectral search program (Version 2.0f) manual. US Department of Commerce, National Institute of Standards and Technology, Gaithersburg, MD, USA.
35. NSW Land and Property (2016) NSW Spatial Data Catalogue: Fire History – Wildfire and Prescribed Burns. https://sdi.nsw.gov.au/nswsdi/catalog/main/home.page. Last accessed 22/06/2016.
36. Oksanen J, Blanchet GF, Kindt R, Legendre P, Minchin PR, O'Hara RB, Simpson GL, Solymos P, Stevens MHH, Wagner H (2014) Vegan: Community Ecology Package. R package version 2.2-1.
37. Possell M, Jenkins M, Bell TL, Adams MA (2015) Emissions from prescribed fires in temperate forest in south-east Australia: implications for carbon accounting. *Biogeosciences*, **12**, 257–268.
38. R Core Team (2015) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.
39. Ralph J, Hatfield RD (1991) Pyrolysis-GC-MS characterization of forage materials. *Journal of Agricultural and Food Chemistry*, **39**, 1426–1437.
40. Rousseeuw PJ (1987) Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*. **20**, 53–65.
41. Sáiz-Jiménez C (1994) Analytical pyrolysis of humic substances – pitfalls, limitations, and possible solutions. *Environmental Science and Technology*, **28**, 1773–1780.
42. Schellekens J, Buurman P, Fraga I, Martinez-Cortizas A (2011) Holocene vegetation and hydrologic changes inferred from molecular vegetation markers in peat, Penido Vello (Galicia, Spain). *Palaeogeography Palaeoclimatology Palaeoecology*, **299**, 56–69.
43. Schellekens J, Buurman P, Pontevedra-Pombal X (2009) Selecting parameters for the environmental interpretation of peat molecular chemistry – a pyrolysis-GC/MS study. *Organic Geochemistry*, **40**, 678–691.
44. Schlesinger WH (1986) Changes in soil carbon storage and associated properties with disturbance and recovery. In: The Changing Carbon Cycle: A Global Analysis, Trabalka JR, Reichle DE (Eds.), Springer-Verlag, New York, 194–220.
45. Schmidt MWI, Noack AG (2000) Black carbon in soils and sediments: analysis, distribution, implications, and current challenges. *Global Biogeochemical Cycles*, **14**, 777–793.
46. Sharma RK, Chan WG, Seeman JI, Hajaligol MR (2003) Formation of low molecular weight heterocycles and polycyclic aromatic compounds (PACs) in the pyrolysis of alpha-amino acids. *Journal of Analytical and Applied Pyrolysis*, **66**, 97–121.
47. Therneau T, Atkinson B, Ripley B (2015) rpart: Recursive Partitioning and Regression Trees. R package version 4.1-10.
48. Tolu J, Gerber L, Boily JF, Bindler R (2015) High-throughput characterization of sediment organic matter by pyrolysis-gas chromatography/mass spectrometry and multivariate curve resolution: a promising analytical tool in (paleo)limnology. *Analytica Chimica Acta*, **880**, 93–102.
49. Urbanski SP, Hao WM, Baker S (2009) Chemical composition of wildland fire emissions. In: Wildland Fires and Air Pollution, Bytnerowicz A, Arbaugh MJ, Riebau AR, Andersen C (Eds.) Developments in Environmental Science Volume 8, Elsevier BV, Amsterdam, 79–107.
50. Vancampenhout K, Wouters K, De Vos B, Buurman P, Swennen R, Deckers J (2009) Differences in chemical composition of soil organic matter in natural ecosystems from different climatic regions – a pyrolysis-GC/MS study. *Soil Biology and Biochemistry*, **41**, 568–579.
51. Vancampenhout K, Wouters K, Caus A, Buurman P, Swennen R, Deckers J (2008) Fingerprinting of soil organic matter as a proxy for assessing climate and vegetation changes in last interglacial palaeosols (Veldwezelt, Belgium). *Quaternary Research*, **69**, 145–162.
52. Volkova L, Weston C (2013) Redistribution and emission of forest carbon by planned burning in *Eucalyptus obliqua* (L. Herit.) forest of south-eastern Australia. *Forest Ecology and Management*, **304**, 383–390.
53. Wenig P, Odermatt J (2010) Efficient analysis of Py-GC/MS data by a large scale automatic database approach: an illustration of white pitch identification in pulp and paper industry. *Journal of Analytical and Applied Pyrolysis*, **87**, 85–92.