Mathematical Methods for the Processing and Analysis of Chemometric Data

Guy Beavis

PhD

University of York
Mathematics

April 2022

Abstract

In the field of Chemometrics, methods are constantly being developed for both the pre-processing and analysis of data. Within this thesis, many such methods are explored in the analysis of multiple data sets with real-world applications, and flaws within common methods are discussed.

Firstly, one data set exhibits significant issues with alignment, and so a novel approach for pre-processing which uses varying wavelet levels paired with a correlation-based alignment method is presented and applied to this data set, called multi-stage feature extraction (MSFE). This method allows for poor quality NMR data to have usable features extracted accounting for issues within the spectra, and the features are graded to allow for control over the quality of the data used for analysis. Secondly, a method for identifying complementary features over any number of data sets is presented, which allows for common compounds of interest to be identified in each set using data fusion.

These methods are shown to improve on existing methods for analysis of data, as well as compound identification. For MSFE this is presented via comparison of the accuracy of models formed from data processed using a range of methods. For the data fusion approach the information gained is used to tentatively identify multiple compounds which were previously difficult to do so.

Analysis is also presented on four data sets, with models formed which categorise observations as well as identify potential markers for a variety of parameters. This includes analysis on honey observations from around to world to find how they differ between origins, as well as the floral origin to see which compounds can act as markers for fraud. Three coffee data sets are also analysed with regards to their taste intensities and origin. This analysis makes use of the proposed methods where applicable, including one NMR data set being analysed where existing methods are unable to work. Because of the novel contributions put forward in this thesis, compounds have been identified which have the potential to ask as indicators for both origin and taste intensity.

Contents

List of Tables

List of Figures

Acknowledgements

I want to thank my fiancée, Samantha, for all the help and support she has given me throughout my PhD, and life. She has stood by my side every step of the way and I could not ask for more.

My supervisor, Julie Wilson, also has my thanks for going above and beyond with the advice and supervision she has provided. I would not have been able to get to this stage without your help, and I am eternally grateful for everything you have done.

My thanks go to Adrian Charlton, Liz Dickinson, Michael Dickinson, James Donarski, Ed Bergstrom, Martyn Ward, Craig Eaton, and Mark Sykes for the guidance they have provided over the last few years with my research.

I would like to thank EPSRC for their EPSRC National Productivity Investment Fund (NPIF) studentship for providing the funding for this research (grant number EP/R512230/1), as well as Fera Science Ltd as my industrial sponsor.

Finally, I would like to thank my friends and family for always being there when needed.

Declaration

I declare that this thesis is a presentation of original work and I am the sole author. This work has not previously been presented for an award at this, or any other, University. All sources are acknowledged as references.

Chapter 2 includes contents originally published in Beavis, G., Rusilowicz, M., Donarski, J., Charlton, A., & Wilson, J. (2019). Chemometrics Applied to Nuclear Magnetic Resonance Analysis. Encyclopedia of Analytical Chemistry: Applications, Theory and Instrumentation, 1-38.v

The data and experimental information used were provided by James Donarski, Fera Science Ltd (Honey and coffee ^{1}H NMR data sets), Ed Bergstrom, University of York Chemistry Department (Coffee LC-MS data sets), and Martyn Ward, University of York Chemistry Department (Coffee GC-MS data set).

Chapter 1

Introduction

The journal ‘Chemometrics and Intelligent Laboratory Systems’ defines chemometrics as "the chemical discipline that uses mathematical and statistical methods to design or select optimal procedures and experiments, and to provide maximum chemical information by analysing chemical data" [1]. To achieve this, various statistical pattern recognition techniques are used to transform raw data into a format from which meaningful results can be obtained. Whereas inspection of single observations can uncover information regarding the structure of a particular sample, multiple observations are frequently compared to identify consistent patterns. Some of the most widely-used methods in chemometric studies are principal component analysis (PCA), projection onto latent spaces (PLS) (also called partial least squares), and linear discriminant analysis (LDA).

Chemometrics is used for many applications, such as authenticity testing [2]–[4], classification modelling [5]–[7] and identifying markers [8]–[10] that fit a set of parameters or trends within the observations. Such studies often focus on differences between observations to form discriminant rules, often in the form of thresholds for classification. For example, leptosperin is a marker for manuka honey, first identified by Kato et al. in 2014 [11]. This was found to correlate with non-peroxide antibacterial activity and the unique manuka factor (UMF), which in turn determines the quality and therefore price of manuka honey.

The methods used to analyse the chemical composition of samples are generally split into three categories: targeted, non-targeted, and semi-targeted analyses. In targeted analysis, the instruments used to analyse samples are calibrated for known compounds with individual calibration curves which result in the highest selectivity and sensitivity for the compounds of interest. By contrast, in non-targeted analysis, the number of compounds detected is maximised, at the cost of lower sensitivity. Finally, semi-targeted has more general calibration curves which limit the detection of compounds in comparison to non-targeted methods, although optimisation is aimed at multiple compounds. Using either targeted or semi-targeted analysis can improve the performance of certain compounds, but such methods are not suitable for exploratory analysis. Within this thesis, we focus only on non-targeted analysis.

One of the most common methods for data collection is nuclear magnetic resonance (NMR) which uses the molecular spin resonance frequency of certain compounds. This can only detect compounds of a certain nature, for instance, ^{1}H NMR can only detect spin relating to hydrogen atoms, and ^{13}C can only detect spin relating to carbon-13 isotopes. As a highly reproducible method, NMR is one of the most frequently used analytical chemistry methods [12]. A complementary method is mass spectrometry (MS) which measures the mass to charge ratio (m/z) of the compounds in a complex mixture. To increase the separation of the compounds within the sample this is often paired with chromatography, either liquid (LC-MS) or gas based (GC-MS). These hyphenated techniques provide an extra dimension to the data for each observation which can aid compound identification. However, there are other limitations, for instance, gas chromatography can only work with volatile compounds.

Due to the financial costs involved with data collection, great care is taken with sample handling, preparation, and machine cleanliness to minimise external effects on the observations [13]. For instance, within NMR solid particles on the outside of the sample container can cause magnetic field inhomogeneity which results in noise within the data. Differences between laboratories can also occur, i.e. different measurements achieved relating to the varying sample handling procedures. To unify the results, many labs partake in proficiency tests to measure how their results differ and minimise these effects.

Pre-processing is often required to transform the data in such a way as to minimise interference or standardise the data so that fair comparisons can be made. For example, one source of variance between observations that is often of no interest in terms of the experiment is simply when the sample was analysed. Methods that use chromatography are prone to differences between the batches that the samples were collected in and observations often need to be corrected. Common approaches for batch-correction include running quality control (QC) samples at various stages throughout the run and using these to adjust the observations [14].

Pre-processing methods can also be used to correct problems with NMR data. Unwanted shifts due to factors such as pH are a common problem with NMR and, if not properly accounted for, can adversely affect the analysis. Such shifts can often be dealt with using binning or alignment algorithms. The simplest approach for binning integrates the spectra over fixed-width regions to provide new features. However, this is now widely regarded as inappropriate and methods that dynamically calculate bin ends corresponding to peaks in the spectra, such as adaptive binning [15], are now available. This algorithm makes use of wavelet transforms to control the level of smoothing applied to the reference spectrum. Many other pre-processing methods also use wavelets to provide approximations to the signal at different resolutions or to calculate parameters for pre-processing.

Analytical chemistry methods such as NMR, LC-MS and GC-MS, can generate tens of thousands of data points for each observation leading to ‘short and fat’ data sets, rather than the more classical ‘tall and thin’ data sets used in statistics. Therefore, methods such as PCA are often used to reduce the dimensionality of the data set to a more manageable number. PCA is an unsupervised method used for visualisation and exploratory analysis as well as data reduction.

After pre-processing, data analysis methods are chosen depending on the aims, such as classification where a method such as LDA would be used, or to identify similarities between observations clustering methods such as k-means might be used (described in Chapter 2.3) . Most methods of analysis can fall into one of two categories, supervised or unsupervised. The former uses the response variable to find corresponding patterns in the data, for instance to create a discriminant function to test if a given sample is sourced from a particular country. In the latter, the response is not utilised and natural patterns within the data are explored, for instance, to see which variables cause variance within the data set. By performing unsupervised analysis it is also possible to find the key characteristics of the data, such as whether the data has any potential issues such as outliers or non-standard observations, as well as differences and similarities between observations, not necessarily relating to their group. Supervised and unsupervised methods can also be combined, for instance using the structures and components found in an unsupervised technique within a supervised method. The most common combination in chemometrics is PCA-LDA, where the PCA scores are used as variables to form the discriminant in LDA. This allows for these natural structures to be utilised in the supervised methods which can discriminate otherwise hidden latent variables to achieve greater classification. Both pre-processing and data analysis methods are continually being developed allowing for new conclusions to be drawn, as well as improving the efficiency of the analysis.

Data collected from various methods can also be combined in a process called ‘data fusion’. This allows complementary methods to be used together to find deeper patterns within the data and can aid the identification of compounds. The many different methods for data fusion can primarily be split into three tiers depending on how processed the data is before fusion. Each tier has its own positive and negative characteristics, for instance, low-level fusion, where raw data blocks are combined, has the lowest cost before analysis but obviously the large fused data block slows the speed of later analysis. Many data fusion approaches involve concatenating the data together to be used in the analysis. However, some methods such as SHY instead focus on trying to identify common compounds in complementary data sets. This is particularly useful for compound identification as databases can be cross-referenced with the additional information available, much as combined data collection methods (i.e. GC-MS) provide a second parameter for each compound.

Multiple data sets are analysed throughout this thesis, as well as being used in the novel methods presented. The first data set is ^{1}H NMR data of honey samples, with a significant number sourced from China. China is the worlds leading producer of honey, producing 24% of all honey in 2019 [16]. Due to differences between the Codex Alimentarius [17], the joint FAO/WHO foods standards, and Chinese regulations (National Standards of People's Republic of China GB 16740-2014, 2015), Chinese honey is sometimes regarded as 'fake'. This is a result of the Chinese regulation allowing changes to the contents of honey, such as the water level, which the Codex forbids. Although, there are other ways in which Chinese honey can be adulterated such as the honey being cut with sugar syrups to reduce costs, as well as other cost-saving techniques [18], differences may simply be due to differences in beekeeping practices. After nectar has been collected by bees and turned into honey it has a moisture content of approximately 50% to 70%. It is deposited in the honeycomb and capped with wax. The high-moisture honey then matures over approximately one week and its moisture content decreases to around 20%. The age of honey when harvested can vary significantly and affect the resulting honey's moisture content. In some agriculture practises the bees are used primarily for pollination, with hives transported around various fields, farms, or greenhouses [19]. In this case, the honey is considered a bi-product and is often not given the time to reach maturity resulting in greater moisture content in the final product. China produces such immature honey, in contradiction to the Codex [20], furthering the case that China's honey if often not be considered genuine honey.

The second set of data used in this thesis for demonstration is sourced from coffee samples. This data has been collected by three methods, ^{1}H NMR, GC-MS, and LC-MS. Due to the number of acids present within coffee, many of the peaks within the NMR spectra are shifted, and for the LC-MS data poor sample preparation has taken place. These samples were sourced from six different countries, and there are many ways in which coffee cherries are processed to create the beans ready to make coffee, with geographical origin often determining which is utilised. Taylors of Harrogate have three methods in which their coffee is processed [21]. The first is 'wet', where the coffee cherries are de-pulped, fermented, and then washed before roasting. This can be an expensive approach and requires a source of clean water, however, provides consistent coffee. A second approach, which is significantly easier and cheaper is 'natural' processing. Here, the cherries are left to dry in the sun with their pulp still on the bean. This however can be inconsistent as the drying time is dependent on the ripeness of the cherry. The final method is 'honey' processing, a mix between the 'wet' and 'natural' approaches, in which the cherries are de-pulped before drying in the sun. This is cheaper than the 'wet' approach and more consistent than the 'natural' approach, plus a water source is not necessary.

1.1 Scope of the thesis

This thesis aims to review and develop techniques used in chemometrics for ^{1}H NMR, GC-MS and LC-MS, primarily focusing on ^{1}H NMR.

Chapter 2 presents methods widely used in chemometrics. First, methods for the collection of chemometric data are discussed. Then many of the ways that data can be transformed before data analysis are described. Both supervised and unsupervised data analysis techniques are discussed with examples utilising Fisher's frequently-used iris data set [22]. As wavelet transforms are commonplace in chemometric analysis and are used in the methods developed within this thesis, a section covering wavelets in detail is provided, with examples.

In Chapter 3 the data sets used within this thesis are described. This covers sample preparation, the parameters used for the analytical instruments, as well as a summary of the data collected. The five data sets consists of observations from coffee and honey samples, collected using ^{1}H NMR, GC-MS and LC-MS, as well as taste intensity for several of the coffee varieties. These data sets are analysed throughout the thesis, as well as being used in the development of the novel methods presented. Each data set requires varying workflows for pre-processing and analysis, due to the differing quality between the data sets, and these are presented alongside the analysis.

Chapter 4 covers analysis of the ^{1}H NMR data sets. This begins with the straightforward analysis of the honey data, classifying and identifying potential markers for the geographical origin and flower type. Next the coffee observations are analysed and the problems due to large chemical shifts used to highlight the fact that traditional methods can not account for such issues. A novel multi-stage feature extraction method that overcomes these problems is then presented. This new approach for ^{1}H NMR data makes use of the local quality of the data to produce dynamic bins for the data, as well as assigned classes based on this quality. The shifts which occur within spectra are not constant, each compound can exhibit different magnitudes, either upfield or downfield. In order to account for this, peaks which exhibit shifts are correlated, and the shifts are corrected based on a 'driver' region, chosen based on which has the lowest interference out of the correlated regions. This approach allows for an improved adaptive binning, and the bin ends are also adjusted for each observation which aids with keeping the integrated regions consistent, rather than splitting peak intensities over multiple bins. Results for a range of different wavelet transformations are compared to calculate the most suitable level for each region in the spectra. Whereas wavelet levels often use the half-height-full-width metric for choosing the suitable wavelet level this can be either too coarse or fine for many of the peaks. By optimising the scale for each region this problem is overcome, and can further improve the quality of features that are extracted from the spectra. Features extracted in this way from the coffee data are then used in data analysis to demonstrate the power of the method in negating the effects of unwanted chemical shifts, with models created for both taste scores and coffee origin.

Chapter 5 presents the analysis of the mass spectrometry data sets as well as the methods used to overcome the problems with the data. Finally, in Chapter 6, data fusion techniques are applied, both existing methods and a new novel method. By combining complementary techniques for the same observations deeper connections can be found, as well as aiding compound identification. The new method draws elements from both STOCSY [23] and SHY [24] in order to create a method for targeted correlation of multiple data sets. Here, similar compounds are identified in three different data sets, and has no theoretical limit on the number of data sets that can be combined is this way.

The novel methods presented in this thesis are critical in the ability to analyse the data, and to identify features which can be uses as markers for both origin and taste intensity. Without the MSFE algorithm the analysis of the ^{1}H NMR data set is not possible, as existing method cannot account for the shifts in the data and identity these shifted features as markers despite them not being suitable in some cases. For the data fusion technique, multiple data sets are combined in order to find common features and this information is useful when trying to identify the compounds associated with the features.

Chapter 2

Methods

2.1 Analytical Chemical Techniques

2.1.1 Nuclear Magnetic Resonance

While all nuclei have nuclear spin, only nuclei with odd mass numbers, such as ^{1}H or ^{13}C but not ^{12}C have a net nuclear spin. When excited via an external magnetic field, these particles' spin have a resonance frequency which is dependent on the molecules involved (for example, the resonance frequency of benzene might be 400.132869 MHz) and so can provide information about chemical composition.

The observed (time domain) NMR signal generated by the oscillations in the radio-frequency detection coil is referred to as the free induction decay (FID). Various mathematical functions can be applied to the FID before Fourier transformation and can dramatically increase the quality of the spectra obtained [25]. For example, apodization convolutes the oscillations by a given function such that the noise at the so called 'foot' of the data is reduced. Apodisation or window functions weight the decay in the time domain in order to maximise the signal-to-noise ratio (SNR) in the frequency domain. The line broadening introduced by apodisation can accommodate minor peak shifts, but can introduce problems with overlapping peaks, particularly with complex mixtures. There is generally a trade-off between noise reduction and resolution, although the application of separate functions for the real and imaginary parts of the FID has been proposed as a method to improve both the SNR and the resolution [25]. An investigation of various apodisation functions in ultra-fast 2D NMR also found that sensitivity could be improved without significantly compromising resolution [26]. The study found a Gaussian function to be particularly effective at removing the distortions present in ultra-fast spectra. The choice of function can affect the results and the process of apodisation may require optimisation.

This resonance frequency data, recorded in the time domain, is Fourier transformed which converts the data into the frequency domain to provide a spectrum for each observation. By comparing the intensities to that of an added synthetic solution such as trimethylsilylpropanoic acid (TSP) which has a frequency 400.130000 MHz, the chemical shift δ, measured in ppm, is calculated via $\delta = (f_{TSP} - f)/f_{TSP}$, where f_{TSP} is the resonance frequency of the TSP and f is the frequency measured. The TSP peak is therefore calculated as 0 ppm, giving benzene a chemical shift of 7.17 ppm. A common metric used in NMR analysis is the half-height of a peak, for instance in identifying optimal parameters for pre-processing [27].

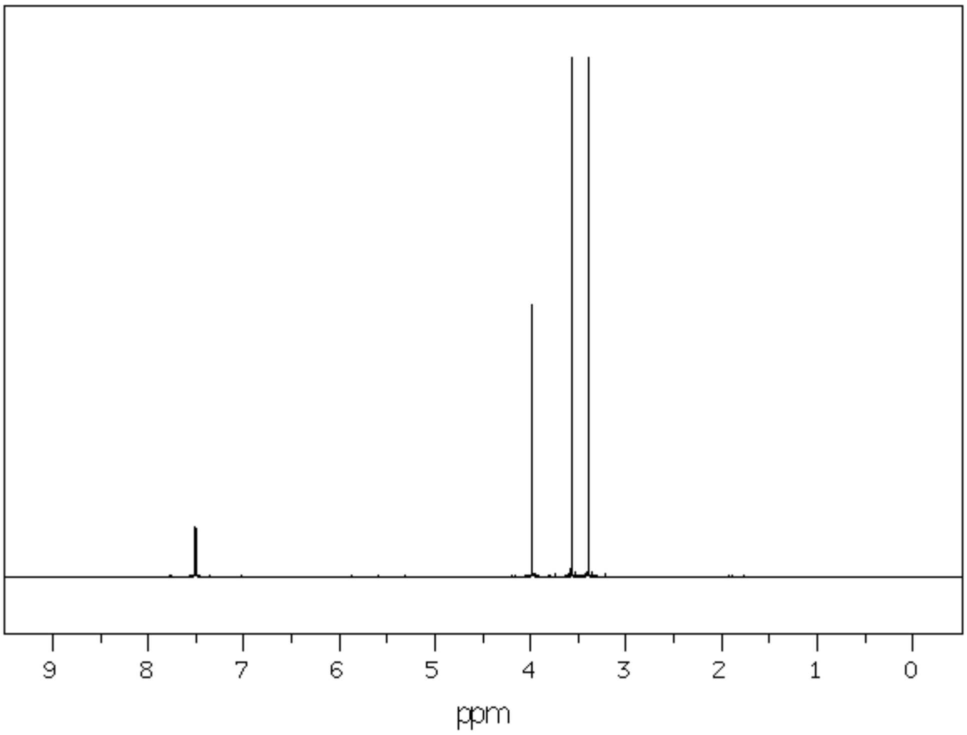


Figure 2.1: Region of an NMR spectrum for caffeine showing the intensities of various compounds and their ppm values. Image is reproduced from Chemical Book [28].

There are several means by which multiple peaks can form from a single compound. The number of hydrogen groups in the structure can determine how many peaks are formed, as each will have its own resonance frequency. Caffeine ($C_8H_{10}N_4O_2$) contains four hydrogen groups, which results in four peaks along the spectra with ppm values of 3.41, 3.59. 4.00 and 7.51 as shown in Figure 2.1. The number of hydrogen atoms in the structure also contributes to the intensity of the spin. In caffeine, three molecular groups each contain three hydrogen atoms so that the total peak area for the resulting peaks is triple that for the singular hydrogen group's peak. Figure 2.2 shows the molecular structure of caffeine with the ion that results in the peak at 3.41 ppm highlighted.

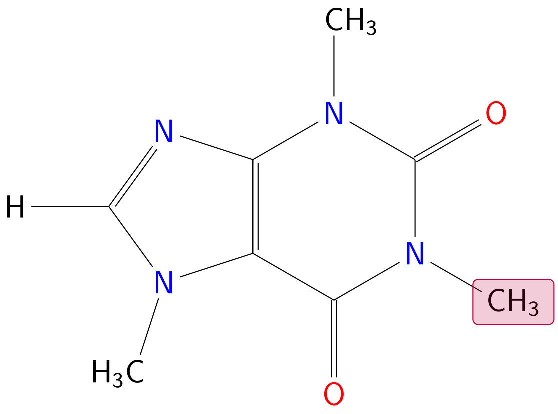


Figure 2.2: Caffeine molecular structure. The highlighted ion (CH_3) fragments in NMR and causes a peak to appear around 3.41 ppm.

Multiple peaks can also be due to peaks which have been split due to spin-spin coupling. These sets of peaks, known as multiplets, occur when neighbouring nuclei influence each other's spin. This interference both shields approximately half of the molecules and deshields the remaining half which causes two peaks to form instead of just one, one upfield and one downfield, called a doublet. Depending on the molecule this can occur multiple times, and so instead of a single peak being formed several symmetrical peaks appear in the spectra. An example of this is a triplet, which is a pattern of 3 evenly spaces peaks with a 1:2:1 relative intensities that are separated by a coupling constant J, equivalent to a doublet of doublets.

2.1.2 Mass Spectrometry

An alternative popular technique used to investigate chemical composition is Mass Spectrometry (MS). MS is used to measure the mass-to-charge ratio (m/z) as this can provide information relating to the compounds found within a observation [29]. This method works by vaporising observations by introducing an electron beam, and the vapours are converted into ions. These ions will have a positive or negative charge depending on the nature of the electron beam, and so by accelerating them through a tube they are deflected in a magnetic field towards the detector. This deflection is based on the m/z value of the ion as well as the charge, and by carefully selecting the strength of the field a limited amount of the ions will reach the detector, with the rest deflecting onto the tube walls. By adjusting the intensity of the field this deflection will change, and so over time ions with various masses can reach the detector. Depending on the charge of the beam different either a positive or negative mode data set is produced. Often the technique is repeated with the opposite charge to create both data sets, as they can contain different results. Figure 2.3 shows the intensities recorded for various masses in an example mass spectrum (reproduced from NIST Chemistry WebBooks [30]) for caffeine.

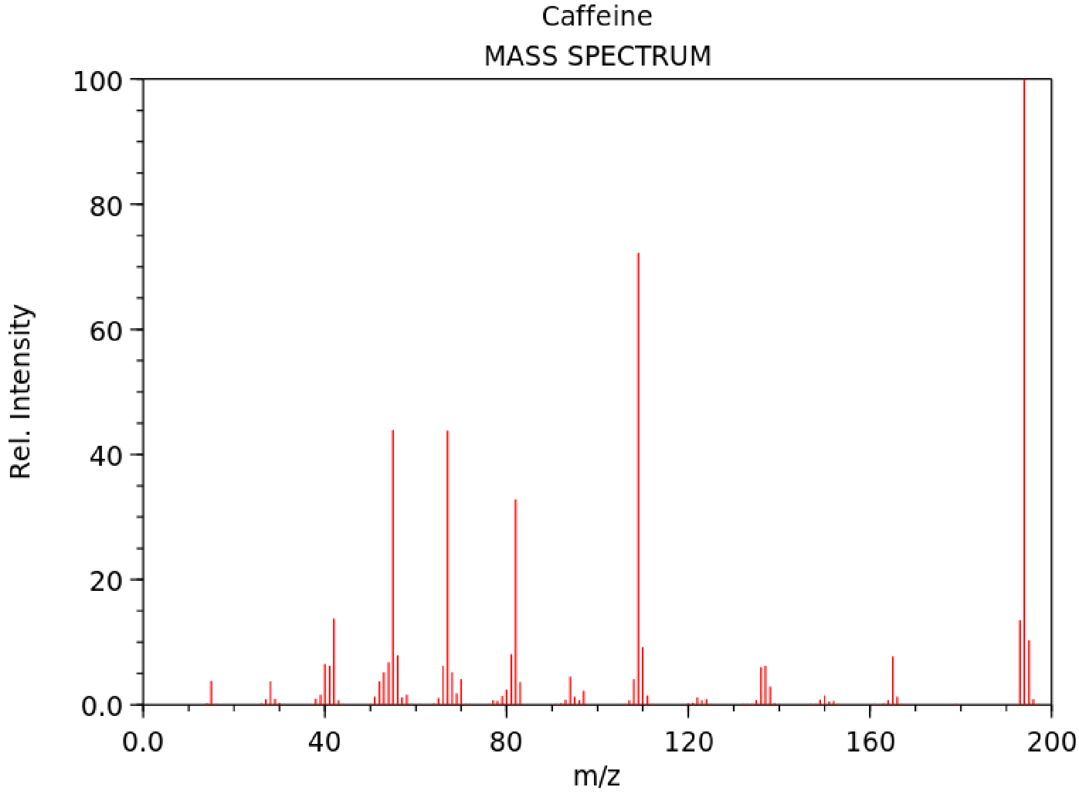


Figure 2.3: Region of an example mass spectrum, with the m/z values on the x-axis and intensity on the y-axis. This has been reproduced from NIST Chemistry WebBooks [30].

When molecules become ionized several of the ions can fragment resulting in multiple smaller molecules reaching the detector. These fragmentation patterns can be used to identify the compounds. An example of fragmenting can be seen in caffeine ($C_8H_{10}N_4O_2$). This has a molar mass of 194.19 g/mol and its mass spectrum is shown in Figure 2.3. While the peak with the greatest intensity is at m/z 194, there are several other peaks, for instance at m/zs 55, 67, 82 and 109, corresponding to fragments of the caffeine molecule [31]. Fragmentation can result in highly correlated peaks in MS, which may be either positively or negatively correlated depending on the fragments.

One method to aid mass spectrometry analysis is by prior separation of the chemical mixture using chromatography.

2.1.3 Gas Chromatography

Chromatography aims at separating compounds from each other based on their chemical composition and other properties. In gas chromatography (GC), an inert gas such as helium, is passed over a small concentration of a mixture during vaporisation (mobile phase). The two most common types of injection, i.e. injecting the sample into the column, are split and splitless. Splitless is used when trace quantities of analytes are wanted to be measured and the entire mixture is measured, whereas split injection uses a higher flow rate and a significant portion of the mixture is discarded. This is faster, however at the cost of lower concentration analytes often having their resulting concentration below the level of detection. This then passed through a heated column where volatile vapours elute at different time, and non-volatiles stick to the column (stationary phase). As the mixture elutes it passes through a detector, where a chromatograph is formed. Various other factors can affect the retention time, such as the boiling point of the component, the length of the column, and the temperature the column is at. By carefully selecting these parameters it is possible to achieve a good separation in a reasonable time.

2.1.4 Gas Chromatography - Mass Spectrometry

When coupled with mass spectrometry to form gas chromatography - mass spectrometry (GC-MS), the retention time from GC adds a third dimension, allowing complex mixtures to be further separated and individual compounds identified more easily as shown in Figure 2.4.

2.1.5 Liquid Chromatography

Instead of using a gas for the mobile phase, liquid chromatography (LC) uses a liquid, usually a solvent. This mixture is then passed through a column and the column length and diameter as well as the composition of both the solvent and column matrix determine the affinity of the components for the fixed phase, and in turn their retention time. High-performance LC (HPLC), also called high-pressure LC, is an advanced form of LC. Here instead of gravity forcing the mixture through the column a high-pressure pump is used instead to increase the pressure. This decreases the time taken, as well as allows for smaller particulate to be used which increases the separation between compounds.

2.1.6 Liquid Chromatography - Mass Spectrometry

As with GC-MS, LC and MS are often paired to provide better separation in LC-MS. However, the techniques are not directly compatible due to MS requiring a vacuum, and therefore an atmospheric pressure ionisation interface in used. Complex mixtures can be separated based on charge and two chromatographs are collected, both positive

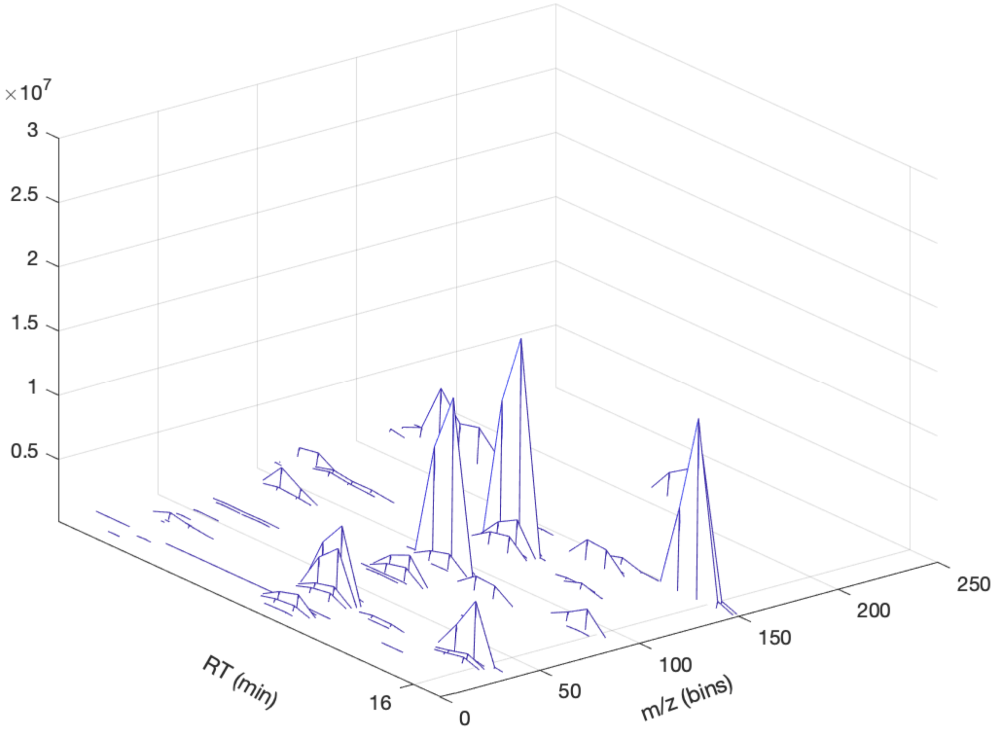


Figure 2.4: Small region of a GC-MS spectrum showing individual peaks for particular retention time and m/z combinations.

and negative modes. This again provides 2D data with m/z and retention time along with intensity.

While LC-MS and GC-MS tend to find different compounds as a result of the volatility required for GC, it is possible for a few select compounds to be found in both methods. Sugar phosphates, for example, can be identified by both GC-MS and LC-MS [32], [33], and certain nucleotides can also be detected by both [34]. When correlating between two MS methods the masses might not be equivalent, and are instead correlated fragments of the molecule.

2.2 Chemometric Methods

2.2.1 Wavelet Analysis

Wavelets are functions that oscillate with varying amplitude and are localised in time and frequency with a mean intensity of zero. The wavelet approximation of a signal allows for features to be considered at varying scales and as such are used frequently for data processing in Chemometrics. Unlike Fourier transforms whose bases are the periodic sine and cosine functions, wavelets are not periodic and instead are only non-zero within a finite domain, allowing the analysis of localised non-periodic signals. Here the construction of wavelet approximations is discussed, starting with Fourier transformations (Definition 2.2.1) for comparison.

Definition 2.2.1 (Fourier Transformation)**.** The Fourier transformation of $f(x) \in \mathbf{L}^2(\mathbb{R})$ is denoted by $\hat{f}(\omega)$ defined by

$$\hat{f}(\omega) = \int_{\mathbb{R}} f(x)e^{-i\omega x}dx. \tag{2.1}$$

Throughout this chapter, it is assumed that $f(x) \in \mathbf{L}^2(\mathbb{R})$. This transform only provides information regarding the frequency of f and does not include time-localisation. This is resolved by taking a window of the signal which is well localised prior to taking its Fourier transform, given by

$$\hat{f}_{win}(\omega, t) = \int_{\mathbb{R}} f(x)\psi(x-t)e^{-i\omega x}dx \tag{2.2}$$

where $\psi(x)$ is a well localised function. Wavelet transformations (WT) provide a similar time and frequency based approximation. The first step to approximating the signal $f(x)$ is to select an appropriate orthonormal wavelet $\psi(t)$, often called the mother function. Examples of three wavelet mother functions are shown in Figures 2.5, 2.7 and 2.10. The scaling and shifting of this mother function is used to provide a lower resolution approximation to the signal. Wavelet transformations can be applied numerous times to the same data, and are described by the number of levels. An s-level wavelet transformation means that the transformation has been applied s times, each time to the previous wavelet level approximation. Applying a high-level wavelet transform to the data provides information at different scales and resolutions, which can prove useful in noisy or large data sets. There are two types of wavelet transforms; the discrete wavelet transform (DWT) and the continuous wavelet transform (CWT) which both are useful for different applications, and are explained in the following sections.

2.2.1.1 Continuous Wavelets

Continuous wavelets are often defined solely by their mother function $\psi(t)$, and make use of wavelet transforms at every scale, with continuous shifts. This mother function $\psi(t)$ then undergoes transformation via scaling and shifting into the wavelet approximations represented by

$$\psi_{a,b}(t) = \frac{1}{\sqrt{a}}\psi\left(\frac{t-b}{a}\right), \quad a = 2^j \text{for } j \in \mathbb{Z}, b \in \mathbb{R} \tag{2.3}$$

where a is the scale and b is the shift. Scaling inversely affects the frequency of the wavelet while shifting translates the wavelet along the time axis. From here the approximation uses a similar method to that for the generation of Equation 2.2, with the resulting transform given by

$$\hat{f}_{wav}(a,b) = \int_{\mathbb{R}} f(x)\psi_{a,b}(x)dx = \langle f, \psi^{a,b}\rangle. \tag{2.4}$$

It is assumed that $\int \psi(t)\ dt = 0$, as this allows for the admissibility condition, that is to say that after a wavelet transform the original signal should be able to be reconstructed from the resolution of the identity formula

$$f = C_\psi^{-1}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} a^{-2}\langle f, \psi^{a,b}\rangle\psi^{a,b} da\ db$$

Here C_ψ is a constant dependent only on ψ, and it is assumed $C_\psi < \infty$.

One of the most popular mother functions used for CWTs is the Mexican hat wavelet (also referred to as Ricker or Marr wavelet), equivalent to the second derivative of the Gaussian function which, after normalising to have sum 2, is given by

$$\psi(t) = \frac{2}{\sqrt{3}}\pi^{-1/4}(1-t^2)e^{-t^2/2}.$$

This wavelet has been applied to NMR spectra to summarise peaks with a reduction of data size while minimising data loss [35]. Its shape, shown in Figure 2.5, resembles a typical peak (as well as a cross-plane of a Mexican hat) and thus can provide an accurate approximation to the spectrum.

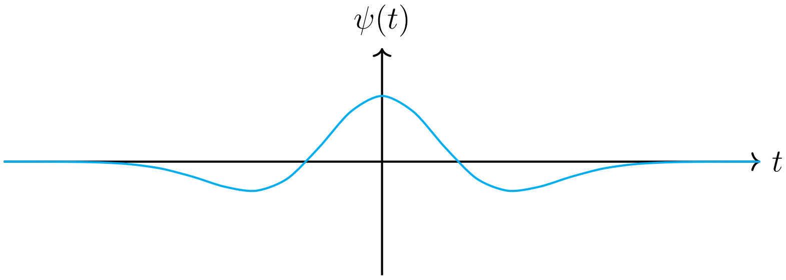


Figure 2.5: Mexican Hat Wavelet Mother Function.

Pattern matching the changes in the approximation coefficients has been used for peak identification [35], [36]. Figure 2.6 (originally published by Du et al. [36]) shows how ‘ridges’ observed when plotting the coefficients over a range of scales can be used to identify genuine peaks in the signal. The upper panel shows the signal, with peaks with their signal-to-noise ratio (SNR) > 3 circled. This ratio is calculated as the ratio of the mean intensity of a peak to the root-mean-square of the noise intensity. The next panel shows the coefficients of the wavelet transformation, obtained using a Mexican hat wavelet. Here, yellow indicates greater intensities and green lower. Ridges can clearly be seen and are highlighted in the final panel where the blue dots show high coefficients with yellow being lower. This highlights how, as the scale for the wavelet approximation increases, the peaks can become smoother, and lower intensity peaks can have their effects negated leaving only the more prominent peaks.

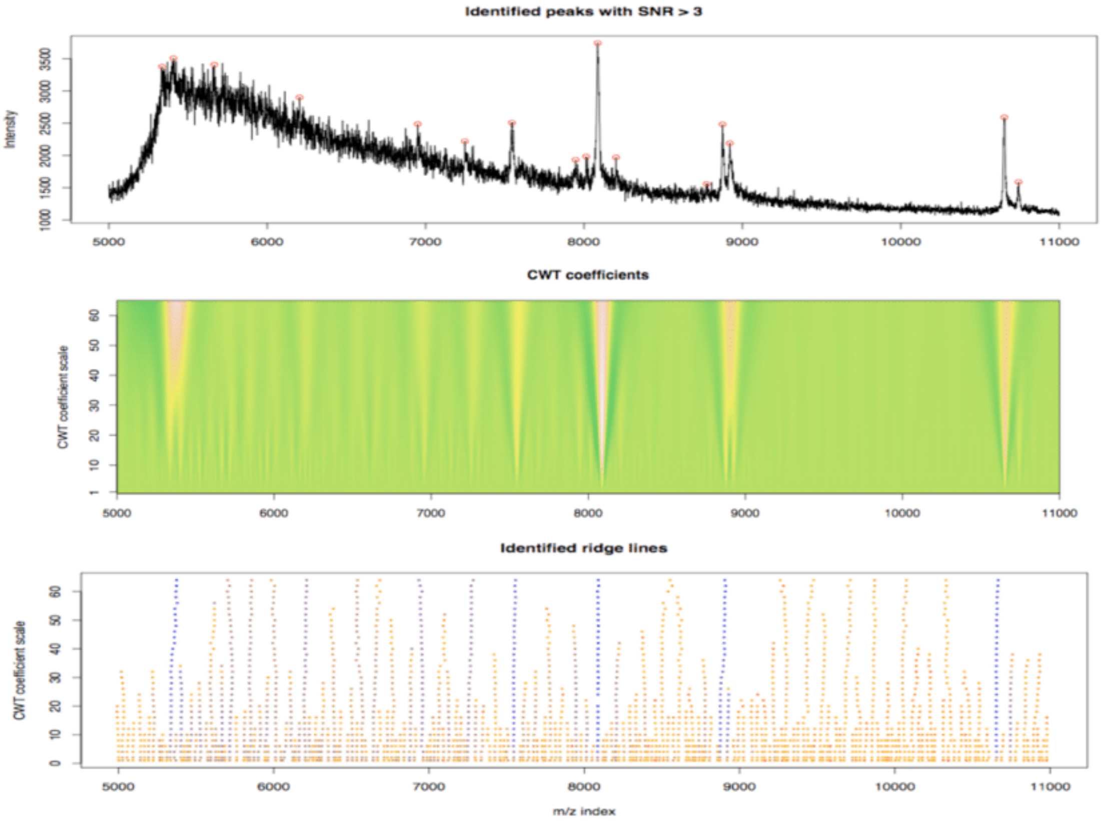


Figure 2.6: Peak identification via ridges as shown by Du et al. [36]. The first plot shows the raw MS spectrum, the second shows the CWT coefficients, and the final plot shows the identified ridge lines coloured by intensity.

2.2.1.2 Discrete Wavelets

Continuous wavelet approximations generate a 2D approximation of a 1D signal, and so to remove the potential for redundancy discrete wavelets use a restricted set of scales and shifts, with $a = a_0^j$, and $b = ka$ for $(j, k) \in \mathbb{Z}^2$. Here, as in most cases, we consider the dyadic transform where $a_0 = 2$. An increasing sequence on subspaces:

$$0 \subset \ldots \subset V_0 \subset V_1 \subset \ldots \subset L_2(\mathbb{R})$$

forms a multi-resolution analysis if the following axioms are applied:

1. $\cap_{j \in \mathbb{Z}} V_j = \{0\}, \quad \cup_{j \in \mathbb{Z}} V_j = L^2(\mathbb{R})$

2. $f(x) \in V_j \Leftrightarrow f(2x) \in V_{j+1} \quad \forall j \in \mathbb{Z}$

3. $\exists\ \varphi \in V_0$ such that $\{\varphi(x-k)\}_{k \in \mathbb{Z}}$ is an orthonormal basis for V_0

as described by Daubechies [37].

It can be shown that A_{2^j}, the approximation to a signal at a resolution 2^j, is an orthogonal projection onto the vector space V_{2^j}. A requirement for the ability to numerically characterise A_{2^j} is to have an orthonormal basis of V_{2^j}. The construction of this basis is defined in Theorem 2.2.1, originally presented by Mallet [38].

Theorem 2.2.1 (Scaling Function)**.** Let $(V_{2^j})_{j \in \mathbb{Z}}$ be a multi-resolution approximation of $\mathbf{L}^2(\mathbb{R})$. There exists a unique function $\varphi(x) \in \mathbf{L}^2(\mathbb{R})$, called a *scaling function*, such that if we set $\varphi_{2^j}(x) = 2^j \varphi(2^j x)$ for $j \in \mathbb{Z}$, then $\left(\sqrt{2^{-j}}\varphi_{2^j}(x - 2^{-j}n)\right)_{n \in \mathbb{Z}}$ is an orthonormal basis of V_{2^j}. A scaling function $\varphi(x)$ must also be continuously differentiable.

The factors of 2^j are included as they normalise the function to the $\mathbf{L}^2(\mathbb{R})$ norm. Via decomposition of the signal, given by

$$\forall f(x) \in \mathbf{L}^2(\mathbb{R}), A_{2^j} f(x) = 2^{-j} \sum_{n=-\infty}^{\infty} \langle f(u), \phi_{2^j}(u - 2^{-j}n) \rangle \phi_{2^j}(x - 2^{-j}n)$$

the orthogonal projection of $f(x)$ on V_{2^j} can now be calculated.

From the definition for the multi-resolution analysis we have $V_{-1} \subset V_0$, and so $1/\sqrt{2}\varphi(x/2)$ can be expressed as a linear combination of the basis functions for V_0, i.e.:

$$1/\sqrt{2}\varphi(x/2) = \sum_{k=0}^{M-1} c_k \varphi(x-k) \tag{2.5}$$

for some coefficients c_k, and this is called the scaling equation. This provides the equations to be able to go from V_0 to V_{-1}, and can be generalised to go from any space V_j to V_{j-1} as

$$\sqrt{2^{j-1}}\varphi(2^{j-1}x) = \sum_{k=0}^{M-1} c_k \sqrt{2^j} \varphi(2^j x - k)$$

Let W_{j-1} be the orthogonal complement of V_{j-1} in V_j, and so as $W_{-1} \subset V_0$, $1/\sqrt{(2)}\varphi(x/2) \subset W_{-1}$ can be expressed as a linear combination of the basis functions for V_0, i.e.:

$$1/\sqrt{(}2)\varphi(x/2) = \sum_{k=0}^{M-1} d_k\varphi(x-k)$$

for some coefficients d_k, and generalised in the same way V_{-1} can be, referred to as the mother function. It can be shown also that

$$d_k = (-1)^k c_{M-1-k} \tag{2.6}$$

where M indicates the support of the wavelet.

The simplest and first discrete wavelet was proposed in 1909 by Haar [39] which has been used in image coding, edge extraction and feature selection [40]. Here, the mother function and scaling function are described by

$$\psi(t) = \begin{cases} 1, & \text{for } 0 \le t < \frac{1}{2} \\ -1, & \text{for } \frac{1}{2} \le t < 1 \\ 0, & \text{otherwise} \end{cases} \qquad \varphi(t) = \begin{cases} 1 & 0 \le t < 1 \\ 0 & \text{otherwise} \end{cases}$$

plotted in Figure 2.7.

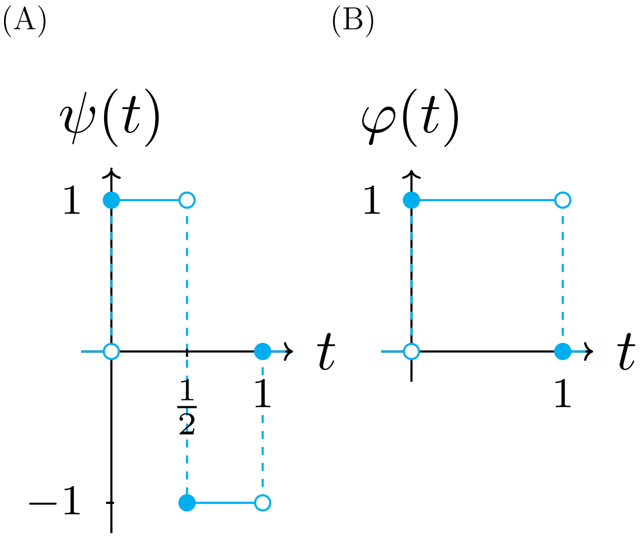


Figure 2.7: Visualisation of Haar wavelets. (A) shows the mother function $\psi(t)$, and (B) the scaling function $\varphi(t)$.

The signal $f(x)$ is equivalent to

$$f(x) = \sum_{k=0}^{N-1} A_{0,k}\varphi(x/2-k)$$

with $A_{0,k} = f(k)$. Via dilation of the scaling function $\varphi(x)$ by a factor of 2 the first level smoothed approximation to $f(x)$, S_1 is calculated as

$$f(x) \sim S_1 = \sum_{k=0}^{N/2-1} A_{1,k}\varphi(x/2 - k)$$

with $A_{1,k} = \frac{A_{0,k}+A_{0,k+1}}{2}$. S_1 therefore only has $N/2$ coefficients. The mother function $\psi(x)$ is used to calculated the differences between the approximation S_1 and the original signal $f(x)$, creating the detail coefficients $D_{1,k}$ which also has magnitude $N/2$. At the next scale the approximation S_2 is derived from S_1 and provides a smoother approximation due to its increased span, calculated as

$$S_1 = S_2 + \sum_{k=0}^{N/4-1} D_{2,k}\varphi(x/4 - k).$$

When the details are added to the smooth approximation there is no loss of information as the scale increases. This is expressed in Figure 2.8, and the formulas shown in Equation 2.7. The figure shows how at each subsequent scale the approximation is split into the following approximation and the details, with the details remaining at every following scale.

$$\begin{aligned}
f(x) &= S_1 + \sum_{k=0}^{N/2-1} D_{1,k}\psi\left(\frac{x}{2} - k\right) \\
&= S_2 + \sum_{k=0}^{N/4-1} D_{2,k}\psi\left(\frac{x}{4} - k\right) + \sum_{k=0}^{N/2-1} D_{1,k}\psi\left(\frac{x}{2} - k\right) \\
&= \sum_{k=0}^{N/4-1} A_{2,k}\varphi\left(\frac{x}{4} - k\right) + \sum_{k=0}^{N/4-1} D_{2,k}\psi\left(\frac{x}{4} - k\right) + \sum_{k=0}^{N/2-1} D_{1,k}\psi\left(\frac{x}{2} - k\right) \\
&= \sum_{k=0}^{N/2^j-1} A_{j,k}\varphi\left(\frac{x}{2^j} - k\right) + \sum_{j=0}^{J}\sum_{k=0}^{N/2^j-1} D_{j,k}\psi\left(\frac{x}{2^j} - k\right)
\end{aligned} \tag{2.7}$$

where $A_{j,k}$ is the k^{th} coefficient for the approximation at scale j, likewise for the details $D_{j,k}$ and N is the number of coefficients in the original signal.

The filter coefficients can easily be calculated for the Haar wavelet using Equation 2.5, shown here. The support for Haar wavelet at the first scale is over the range $[0, 1]$, and so M is 2.

$$\begin{aligned}
1/\sqrt{2}\varphi(x/2) &= \sum_{k=0}^{M-1} c_k\varphi(x - k) \\
&= \sum_{k=0}^{1} c_k\varphi(x - k) \\
&= c_0\varphi(x) + c_1\varphi(x - 1)
\end{aligned}$$

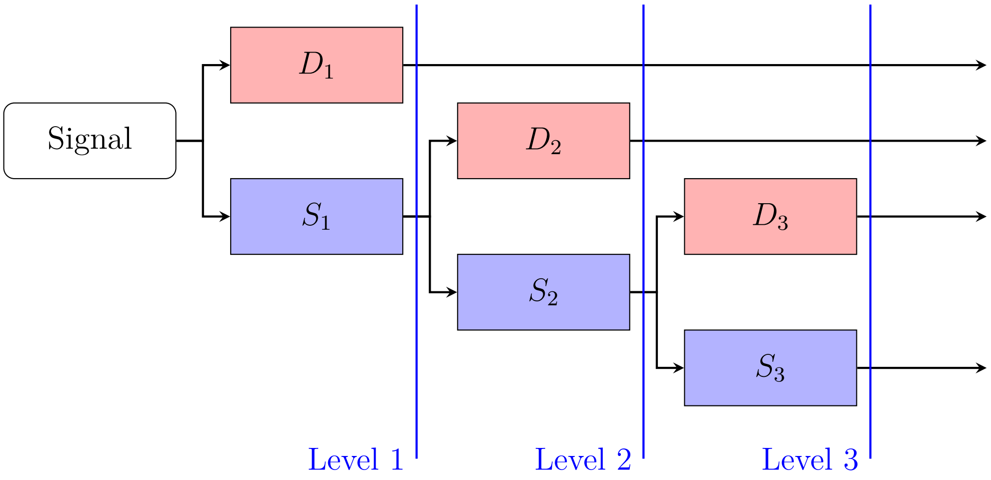


Figure 2.8: Diagram displaying the discrete wavelet transformation at various levels. D_j here represents the detail coefficients, and S_j is the approximation.

At $x = 1$,

$$\begin{aligned}1/\sqrt{2}\varphi(1/2) &= c_0\varphi(1) + c_1\varphi\\ 1/\sqrt{2} * 1 &= c_0 * 0 + c_1 * 1\\ 1/\sqrt{2} &= c_1\end{aligned}$$

and at $x = 0$

$$\begin{aligned}1/\sqrt{2}\varphi(0) &= c_0\varphi(0) + c_1\varphi(-1)\\ 1/\sqrt{2} * 1 &= c_0 * 1 + c_1 * 0\\ 1/\sqrt{2} &= c_2\end{aligned}$$

and using these coefficients in Equation 2.6, $d_0 = (-1)^0 c_{2-1-0} = c_1 = 1/\sqrt{2}$, and $d_1 = (-1)^1 c_{2-1-1} = -c_0 = -1/\sqrt{2}$.

An example of the Haar wavelet transform is shown in Figure 2.9. Here 16 data points have been generated (black), and the resultant Haar wavelet approximations from 1^{st} to 4^{th} level have been applied, expanded to allow comparison. The 1^{st} level is shown by the red lines here, which has 8 data points. At this level the peak location approximately halfway along is still present, and at the 2^{nd} level (orange) it is smoothed out. The final level (blue) has a span of 16, and so the original signal is now compressed to just a single value.

The detail coefficients D can provide valuable insight into the structure of the signal $f(x)$ and are used, for example, in noise detection. Donoho and Johnstone [41] make use of this by thresholding under the assumption that the highest frequencies correspond to noise where the threshold is given by $\sigma\sqrt{2\log N}$. Here N is the number

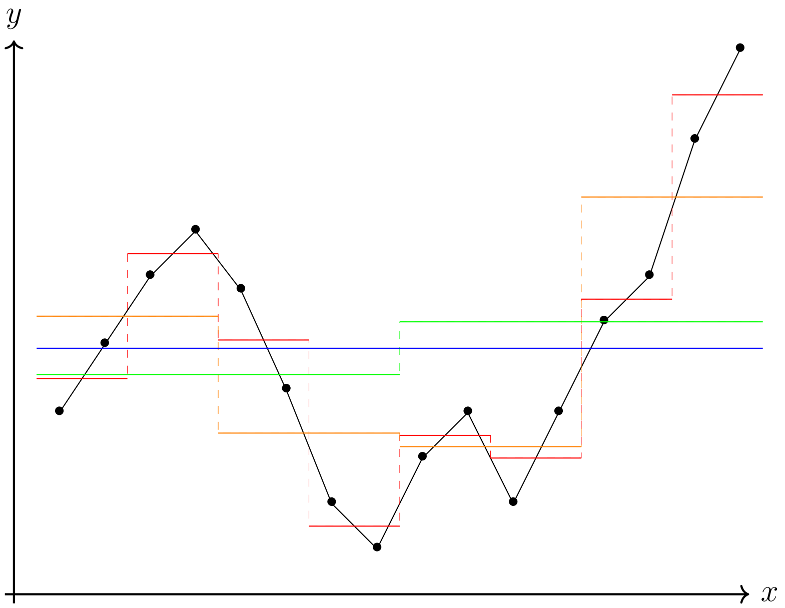


Figure 2.9: Increasing scales for the Haar wavelet of a set of example data. The black dots represent the original data, then red is the 1^{st} level approximation, followed by orange, then green, and finally blue at the 4^{th} level. Each wavelet has been stretched in the x-axis to be able to be overlaid when in reality the resolution is half at each subsequent level.

of data points and σ is the variance of the noise estimated by

$$\hat{\sigma} = \frac{median\{|w_{1,1}|, \dots, |w_{1,N/2}|\}}{0.6745} \tag{2.8}$$

where $w_{1,i}$ is the i^{th} wavelet coefficient for the first level of detail.

Due to discontinuity, the Haar wavelet basis is often not an appropriate basis for approximating smooth functions. Other issues can also arise with discrete wavelets, such as border issues [42] where the approximation is not possible for $f(x)$ for $x = 0$ and $x = N - 1$, however there exist several trivial solutions to this such as padding or performing circular convolutions.

Other examples of discrete wavelet mother functions include one of Daubechies' wavelet families known as coiflets [37]. These were constructed with compact support $|\text{supp } \phi| = |\text{supp } \psi| = 2M - 1$ such that, for a coiflet of order L,

$$\begin{aligned} \int \phi(x)dx &= 1 \\ \int x^l \phi(x)dx &= 0 \qquad \text{for } l = 1, \dots, L-1 \\ \int x^l \psi(x)dx &= 0 \qquad \text{for } l = 1, \dots, L-1 \end{aligned}$$

The solution to these equations generates the filter coefficients for the coiflet wavelet family which can be found in [37], with values for L between 1 and 5 generating the C6, C12, C18, C24 and C30 wavelets, as shown in Appendix A.1 Table A.1. This naming convention relates to the number of filter coefficients each level L generates, which is 6 per level. Figure 2.10 shows the coefficients of three coiflet wavelets: C6, C18 and C30, scaled such that each has the same total sum (2). These wavelets, which have a large peak at $k = 0$, resemble peaks and so the coiflet wavelet family is often used in NMR analysis.

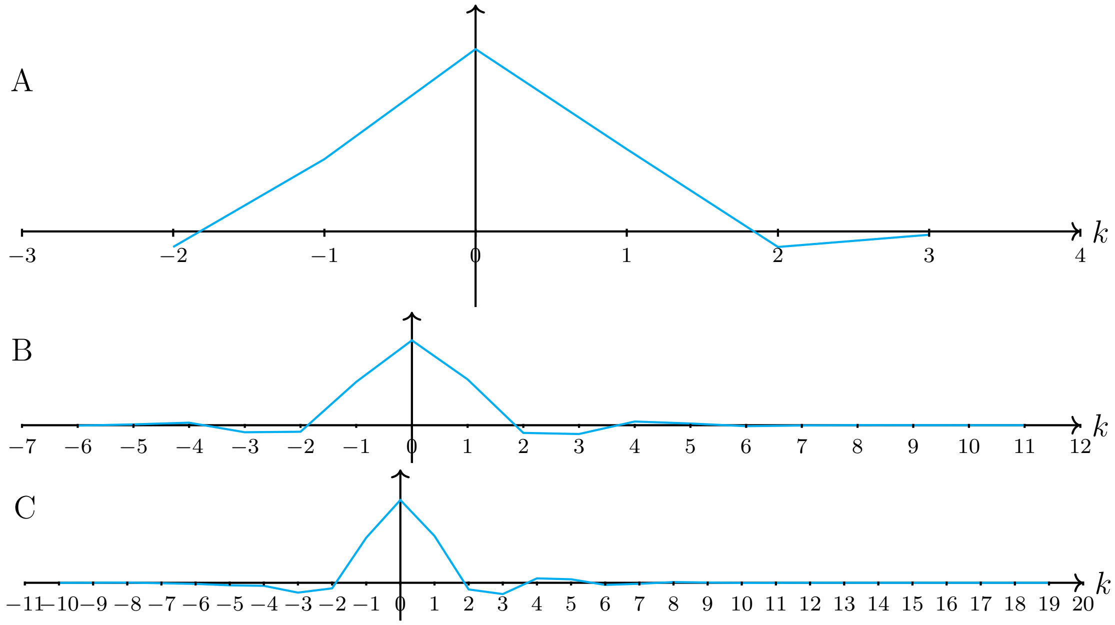


Figure 2.10: Filter coefficients for the coiflet family of wavelets. (A) shows the C6 wavelet, (B) is C18, and (C) represents C30, generated from different coiflet wavelet orders L. As the number of coefficients increase the smoothness of the approximation increases.

Non-decimated Discrete Wavelets Discrete wavelet transforms in which the number of coefficients reduces by a factor each time are referred to as decimating wavelets. Several methods have been developed for non-decimated discrete wavelet transforms. As the fast, or decimated, wavelet transform is not shift invariant, Donoho and Johnstone [41] developed the non-decimated wavelet transform. For each level, the non-decimated transform interleaves the decimated approximation with a second approximation obtained obtained by shifting the spectrum by one data point. This transformed approximation therefore has N data points in comparison with the decimated approximation having just $N/2$, which aids with the comparison between various scales and removes the dependence on initial position.

A 1-level non-decimated Haar wavelet approximation is represented in Figure 2.11, with (A) showing the two approximations (red starting at 1, blue starting at 2), with them being meshed together to form the magenta approximation in (B). The issue with decimating transforms can be seen in (A), where the two Haar approximations are shown to produce different results. The blue approximation is already able to smooth out the second small maximum, whereas the red approximation does not. The magenta transform is still a smoother version of the original signal, as well as retaining the same resolution. A natural result of this method is the approximation appears to be shifted half a data point left, however this is inconsequential due to its limited size.

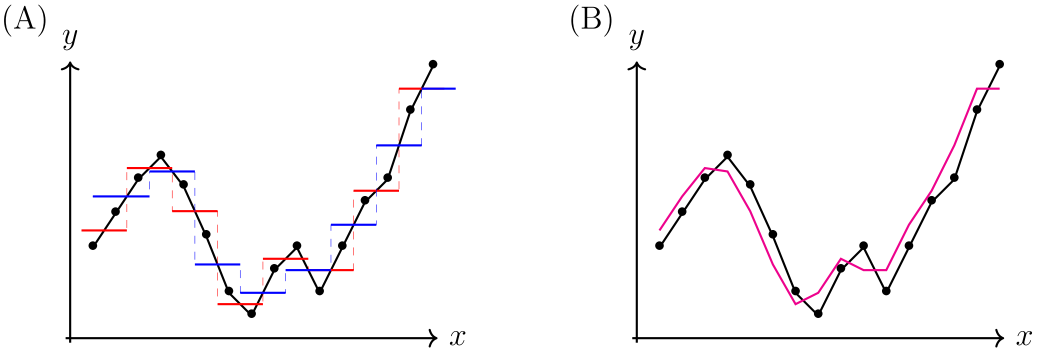


Figure 2.11: Example of the non-decimated discrete wavelet transformation. (A) shows the raw signal (same as in Figure 2.7, with two Haar wavelet approximations. The red approximation starts at the first data point, and the blue approximation starts at the second. (B) shows the resulting non-decimated approximation in purple, as well as the original signal.

2.2.2 Data pre-processing

Some of the methods used for data pre-processing to prepare them ready for analysis are described here. Where methods and algorithms are used later in this thesis a more detailed explanation is given.

2.2.2.1 Denoising

Very small random fluctuations have been shown to cause significant variation in the clustering of otherwise identical spectra [43]. Thermal interference, magnetic field variations, instrumental instability, electrical interference and the digitisation of analogue signals can all contribute to the presence of noise in NMR spectra. Cryogenically cooled probes have helped to reduce electronic noise and apodisation can remove the asymmetric “sinc wiggles” that can appear due to truncation of the free induction delay (FID) [44]. As well as processing of the FID, the SNR can be improved by increasing the number of scans, which should in theory incrementally reduce the noise to zero. In order to reduce noise in targeted MS one popular approach is to select reagents which are known to only ionize the molecules of interest. In practice, random noise can be reduced, but not eliminated due to limits on the number of scans possible. As such sampling processing methods are combined with data pre-processing methods in order to minimise the effects noise can have.

The simplest data pre-processing method used is to select a careful choice of threshold, below which the signal is set to zero. This method has been shown to alleviate the problem and produce almost perfect clustering in principal components analysis (PCA). The choice of threshold necessarily changes the variance between observations and this simple form of noise reduction requires judgement to optimally reduce noise whilst retaining real information.

Other methods for denoising include using wavelets to smooth the data [45], [46], as described in Section 2.2.1. The spectra are represented by wavelet functions, and the wavelet decomposition allows features to be considered at different scales.

Singular value decomposition (SVD) is also used in a method proposed by Cadzow [47] which can separate noise from signals, and also can reveal smaller peaks obscured by noise. This is achieved by truncating the SVD of X and applying a function f (for instance to obtain a minimum variance estimate), and then reconstructing the signal using the anti-diagonals of the SVD. As an alternative, Li et al. [48] describe a time-fractional diffusion method, which extends the classical diffusion model to use an n^{th} order differential, and show that it outperforms the classical method for smoothing NMR spectra.

SVD is usually applied to the frequency domain, however a method proposed by Laurent et al. [49] involves applying SVD to the time domain before the Fourier transformation. When applied to example data this alternative method was shown to detect signals which are as small double the maximum noise, regardless of the peak width. This can result in an improvement for acquisition time; something which is incredibly useful for low sensitivity experiments.

2.2.2.2 Baseline Correction

Poor sampling techniques, such as observation inhomogeneity, or inappropriate phase correction can contribute towards baseline distortion, and can be a large source of variance in analysis [50]. There are numerous methods to correct for this, such as approximation and subsequent removal of the baseline, albeit computationally difficult. These methods take place in either the time-domain where the offset can be resolved before it becomes to difficult to correct [51], or the frequency domain where the baseline is estimated along the full domain passing through the noise regions, but not following peaks.

Dietrich et al. used a standard numeric derivative for baseline recognition after smoothing with a moving average filter [52]. This can lead to small peaks being smoothed out, but more advanced smoothing algorithms significantly increase computer time and can be prohibitively slow with 2D spectra. However, the continuous wavelet transform (CWT) provides good baseline smoothing without the computational expense of other algorithms [53].

Simultaneous correction of both the phase and baseline using multi-objective optimisation has been shown to give improved results in comparison to consecutive correction [54]. In an alternative method, Sokolenko et al. [55] use both dimensions of the complex-valued data created from the Fourier transformation to incorporate apodization as well as phase and baseline correction in the frequency domain. The use of both real and imaginary data to describe peaks allows overlapping peaks to be fitted, something not possible in traditional 1D ^{1}H NMR. Figure 2.12, from Sokolenkos' article, shows the result of their new algorithm for baseline correction. The dashed line in the top two plots shows the estimated baseline over the data, and the lower plots show the overall fit broken down into individual peaks.

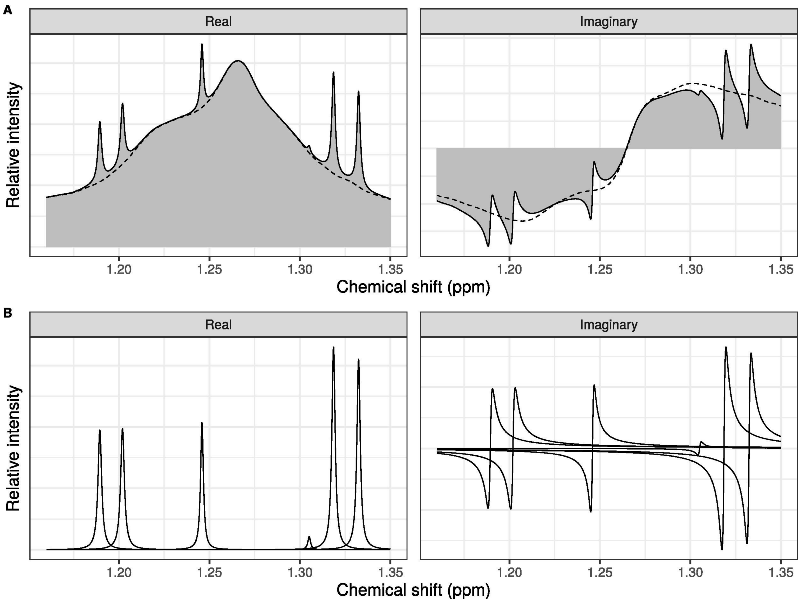


Figure 2.12: Sokolenkos' proposed algorithm applied to a human plasma observation.

2.2.2.3 Binning

There are many conditions which can cause shifts in the data, such as changes in temperature, variable pH levels of the observations, and different ionic strengths [56]. These shifts create differences between spectra that are not related to differences of interest, for example between groups [57]. The shifts are often not consistent along the spectra and can move in either direction, so manually shifting every spectrum along is not a viable solution.

One of the most common methods to overcome the issue of unwanted chemical shifts is to filter the data into bins, also called buckets. The spectrum is split into segments, the simplest consisting of equally spaced bins, normally 0.04 ppm in width. All of the data-points in these bins are then collated, usually by summing the intensities [58]. This reduces the resolution of the spectra such that the unwanted shift is no longer noticeable. A major problem with this method is that peaks can be split across multiple bins, or several peaks can be combined in one bin. A simple solution to this issue is to set individual bin boundaries manually, however this would be a very time consuming task requiring a degree of knowledge about the peaks in the spectra [59], as well as have a low reproducibility due to the human interaction.

In order to improve on uniform binning, methods that allow non-equidistant bin widths have been developed. Adaptive intelligent [60] binning uses variable bin widths, without the need of a reference spectrum. Beginning with one bin covering the full spectra, this method uses a metric incorporating the maximum intensity of each bin paired with the intensities of the bin ends to compare splitting each bin at optimal points, or to leave the bin as is. This process is then repeated until the metric no longer improves after splitting bins. In order to generalise adaptive intelligent binning to multi-dimensional data Worley et al. [61] redefine the metric such that it allows n dimensions, providing a significant improvement from uniform binning.

An interesting comparison between spectral binning and wavelet denoising showed that PCA performed directly on the wavelet coefficients gave better results than PCA on binned data when applied to a series of heteronuclear single quantum coherence spectra of proteins with different ligands present [62]. The wavelet-PCA scheme is also found to detect outliers better, although this may be due to shifts in frequencies bridging bin boundaries rather than the effects of noise as standard equal length bins were used. Wavelet transformation can reduce the resolution of the data and an alternative approach to de-noising and compression is proposed by Puig-Castellví et al. in which genuine peaks are determined by considering the number of neighbouring values after an initial thresholding step [63]. This two-stage de-noising method was tested with various 2D and 3D NMR spectral data sets, giving data reduction up to 2000-fold without affecting the resolution of the peaks remaining.

Adaptive Binning Adaptive Binning (AB) was introduced in 2006 by Davis et al [27]. As apposed to AI binning, this applies a wavelet transformation (usually the Haar wavelet) to a reference spectrum, and under the assumption that each peak on the reference spectrum (after denoising and at a suitable level wavelet-transformation) contains only one peak from each observation the method finds local minima in the spectrum which are used as the bin boundaries. In order to account for noise their method uses hard-thresholding using the Donoho-Johnstone threshold, shown in

Equation 2.8.

In the case where the region has unwanted chemical shift, the reference spectrum can appear asymmetrical or multi-modal, as illustrated in Figure 2.13. In this example the reference spectrum generated from the maximum of each observation spectra will split the centre peak into 3 separate bins, and the final peak into approximately 5 bins. The median reference will create 2 bins in the second set of peaks, and around 4 bins for the third. With sufficient smoothing the reference will lose the small fluctuations and better represent the spectra. These methods are not without flaw, the maximum is more susceptible to changes in the baseline (for instance, when the baseline of one spectrum has greater intensity than peaks in another), while the median can sometimes exclude peaks if they are only found in a minority of the observations. As such smoothing the reference spectrum is recommended, and here Davis et al. suggest using the maximum.

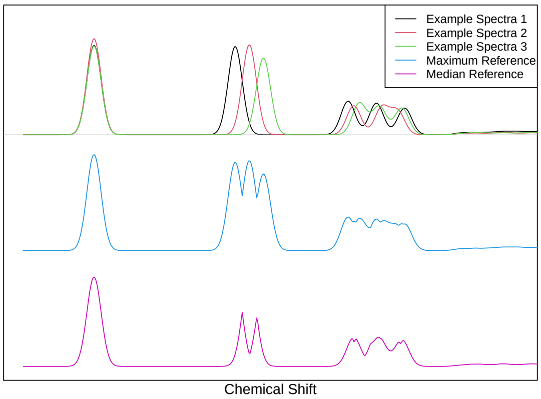


Figure 2.13: An example of how an increase of unwanted shift can decrease the quality of the reference spectrum. Two reference spectra are shown, the median (pink) and the maximum (blue).

In order to smooth the reference spectrum in this algorithm Davis et al. use the non-decimated discrete wavelet transformation. This will remove small humps in the spectrum, however at a too-high level it could lead to over-smoothing, grouping together several peaks which should be separate. By combining the careful choice of reference spectrum, the wavelet transformation of this reference, and then finding the local minima of this transform this method creates bins. This method is designed to be less susceptible to peak splitting than uniform binning, and with careful choice of the level of the transforms can account for most unwanted chemical shifts.

Examples of adaptive binning regions are illustrated in Figure 2.14, using 3-level wavelet transformations, with the level determined from the half-height of a peak. Here, the regions to be binned are highlighted in alternating colours, with black vertical lines at the bin ends. As seen by the bin ends these all are located at minima in the spectra, which minimises effects of neighbouring peaks. Due to the smoothing the minima found at approximately 5.345 ppm and 5.335 ppm have been smoothed, and so these are not be split. These peaks are likely from doublets, which are known to occasionally cause issues with binning and similar methods.

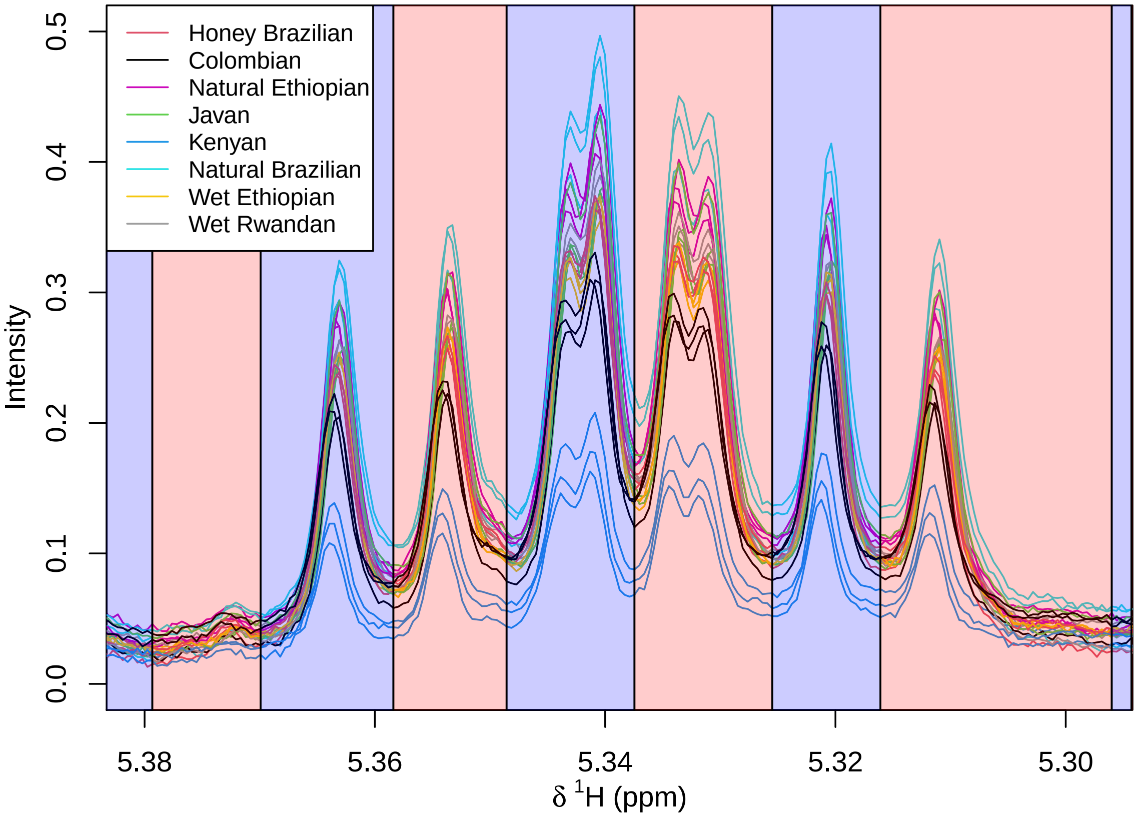


Figure 2.14: Example region showing the bins found in the adaptive binning algorithm after 3-level decimated wavelet transforms. Different coloured regions show the different bins, and black lines the boundary.

2.2.2.4 Alignment

An alternative to binning the data is alignment, where the spectra are warped in such a way to mitigate the issues caused by the shifting.

Two common procedures for automated peak alignment include dynamic time warping (DTW) [64], originally used in speech processing, and correlation optimised warping (COW) [65]. Both are pairwise alignment methods and require a target spectrum to be chosen as the reference to which spectra are to be matched. The DTW uses distance as a similarity measure between two signals and is sensitive to differences in peak intensities [66], which led to the use of the correlation coefficient in COW. In order to align the signal to a reference COW splits them into segments. Each of these is then warped to match the reference segments with a parameter called the slack which limits how much warping can occur. However, COW is computationally expensive and can be sensitive to baseline distortion. The alignment of peaks selected in a prior peak-picking algorithm can provide an effective and computationally less intensive alternative [67]. Kim et al. have shown that a Bayesian model for the estimation of alignment parameters, that simultaneously corrects the baseline, outperforms both DTW and COW in terms of correlation between spectra and execution times [68].

Interval-correlation-shifting (icoShift) [69] is one of the most commonly used methods for alignment. The method segments the full spectra using a recursive segmentwise peak alignment method, and then these segments are aligned individually using a fast-Fourier-transform engine. This does however require a reference spectrum (discussed in the previous section), which can be either one observation spectrum, or calculated from the spectra. Common methods for calculating an example spectrum for this are to use the maximum or median at each point. As the spectra may also be approximated by Gaussian functions, it is possible to generate a reference spectrum using several average peak positions, intensities, and width [70]. By using this Gaussian reference spectrum instead of a randomly chosen observation or similar it removes the ambiguity over which spectra to choose, which can lead to more repeatable results.

In algorithms using automated selection of segments, differentiation between regions containing peaks and those with only noise is vital. Wang et al. developed an algorithm that uses the frequency of intensity fluctuations to allow small peaks to be distinguished from the effects of baseline distortion and noise and used icoShift to align the regions containing peaks [71]. They applied their IFFD-icoShift method to the alignment of NMR spectra from human urine and demonstrated increased interpretability in PCA.

Dynamic Time Warping methods (DTW) Originally used in the 70's for speech processing, DTW has in the past 2 decades began to be used for purposes other than chemometrics, such as cropland mapping [72], [73], or various uses is the medical field [74], [75]. Despite *time* being in the name, there are many different domains this can be used in, such as with chemical shift.

The algorithm discussed in Giorgino's 2009 paper [76] is presented here.

Given two series, a *test* $X = (x_1, \ldots, x_N)$ and a reference $Y = (y_1, \ldots, y_M)$. Assuming a non-negative, *local dissimilarity* function f is defined between any pair of elements

$$d(i, j) = f(x_i, y_j) \geq 0,$$

with f usually the Euclidean distance.

At the core of the technique lies the warping curve $\phi(k), k = 1 : T$:

$$\begin{aligned} \phi(k) &= (\phi_x(k), \phi_y(k)) \text{ with} \\ \phi_x(k) &\in \{1 : N\} \\ \phi_y(k) &\in \{1 : M\} \end{aligned}$$

The warping functions $\phi_x(k)$ and $\phi_y(k)$ remap the domain indices of X and Y. Given ϕ, we can compute the average accumulated distortion between the warped time series X and Y:

$$d_\phi(X, Y) = \sum_{k=1}^{T} d\left(\phi_x(k), \phi_y(k)\right) m_\phi(k)/M_\phi$$

where $m_\phi(k)$ is the weighting coefficient at each step k, and M_ϕ is the normalisation constant. This ensures that the distortions are comparable along different paths. Constraints are imposed on ϕ, such as monotonicity to avoid peak ordering changes, a staple to any alignment algorithm.

The idea for DTW is to find the optimal alignment ϕ such that

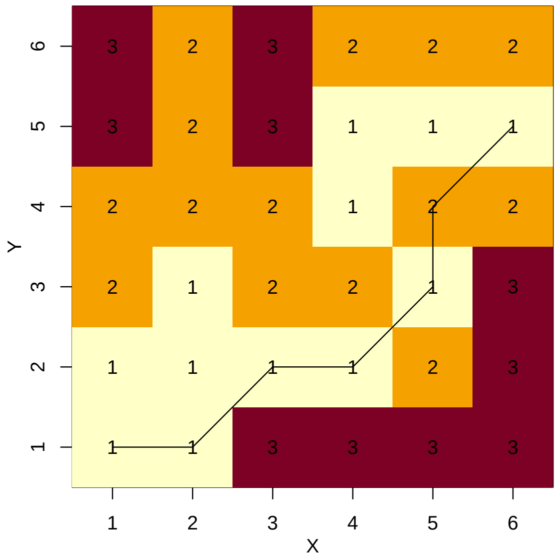
$$D(X, Y) = \min_\phi d_\phi(X, Y)$$

One way to visualise this is found in Figure 2.15. This shows the optimal alignment path on the local distance matrix of a test set of data. The path in this example is weighted, so each step left or upwards starting from (1,1) has a costing equal to the distance travelled, with diagonal steps costing double (i.e. equal to stepping both left and up). It is also not able to move further than 1 space away at any step. Other step patterns, such as the Rabiner-Juang [77] method exist, which are able to step in greater increments.

2.2.2.5 Feature Extraction

As well as binning and aligning data one common data manipulation step is to extract features. Non-equidistant binning can be considered a method of feature extraction. As bins should relate to peaks, some advanced methods can identify the metabolites involved by comparing them to a reference database. Peak extraction can also considerably increase the efficiency of analysis by reducing the number of variables.

One method of matching the peaks to known metabolites is proposed by Xi et al, [78] where each of the peaks are matched and scored to reference compounds in a database, and this also allows for a level of displacement. When tested with complex



Y \ X	1	2	3	4	5	6
6	3	2	3	2	2	2
5	3	2	3	1	1	1
4	2	2	2	1	2	2
3	2	1	2	2	1	3
2	1	1	1	1	2	3
1	1	1	3	3	3	3

Figure 2.15: Optimal solution for an example set of data with DTW. Here each coordinate represents a distance, which it is also coloured by.

biological observations Xi shows that it is an efficient and fast method of metabolite identification.

For ‘Bayesian automated metabolite analyser for NMR spectra’ (BATMAN) [79], a Markov chain Monte Carlo algorithm is used to de-convolute peaks. Automated metabolite identification and quantification using chemical shifts, J-couplings and relative peak intensities is then applied to the peaks. Although computationally expensive, the complementary information in BATMAN, introduced via user-controlled templates, has been shown to produce clinically relevant results when applied to blood plasma observations from lung cancer patients [80].

Speaq 2.0 Speaq 2.0, developed as a complete workflow for NMR spectra processing and quantification, developed by Beirnaert et al in 2018 [35], makes use of wavelets for peak identification and clustering for alignment.

The first step in the Speaq 2.0 algorithm is to find the peaks in the spectra. Instead of making use of a reference spectrum like other methods do, this omits this step and applies a CWT to each individual observation. The wavelet mother function chosen here is the Mexican hat function discussed in Section 2.2.1.1, as this has one major peak and is near-symmetrical. By comparing the local maxima of the CWT coefficients at different scales ‘ridges’ are identified, which relate to peaks positions in the original spectra and so their location can be extracted. An example of these ridges is shown in Figure 2.16, which shows the Mexican hat coefficients for a region of an NMR spectra, with scales from 1 to 10. Here the orange/white areas represent higher coefficients resulting from peaks, and greens and yellows lower coefficients from noise.

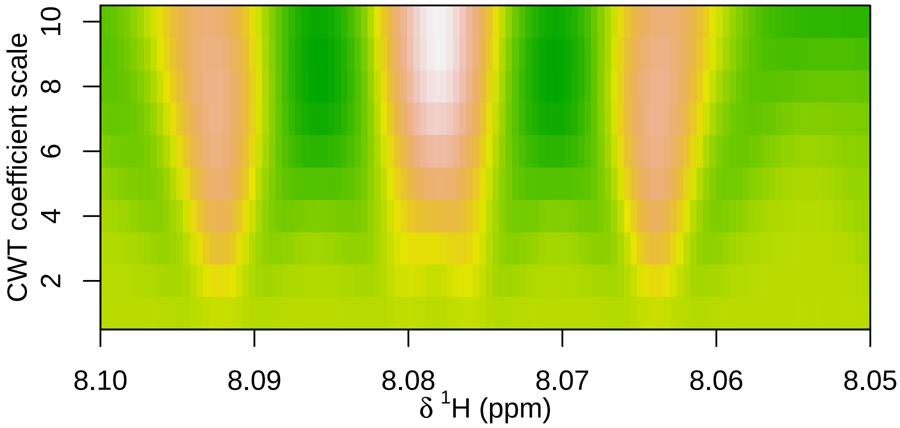


Figure 2.16: Mexican hat coefficients for an example region of an NMR spectra. Three ridges can be identified (red stripes) and become more prominent as the scale increases.

After identifying the peaks, they need to be grouped and Speaq 2.0 achieves this via hierarchical clustering applied to peaks within windows. After a window width is selected it is applied to each point in the data, and the region just outside the window is inspected for peaks which might have been missed, in which case the window width is increased. In order to minimise the effects small peaks can have on alignment it is possible to ignore these nearby neighbouring peaks, using Du et al's alternative algorithm [36] which ignores these when estimating the peak maxima position. There is no metric to determine which is optimal, and the choice is left to the user of Speaq 2.0. Within each window a distance matrix is formed from the peak ppms, and a dendrogram formed from this. Starting at the greatest split the nodes are checked for the number of repeated observations; if this is high the following branch point is inspected until a low proportion of duplicate observations is present. At this point the branch is cut and considered a peak. For these duplicate peaks the Gower distance (Definition 2.2.2) is measured, and the duplicates with the greatest distance removed.

Definition 2.2.2 (The Gower Distance)**.** The Gower distance is defined as the average of partial dissimilarities between two points in feature f, calculated as

$$d(i,j) = \frac{1}{p}\sum_{f=1}^{p} d_{ij}^{(f)}.$$

The partial dissimilarity, $d_{ij}^{(f)}$, is the ratio between the absolute difference of the observations in feature f, and the range of all observations calculated as

$$d_{ij}^{(f)} = \frac{|x_{if} - x_{jf}|}{R_f}$$

where R_f is the range of feature f.

The window analysis continues until all of the peaks have been considered in each window, at which point each cluster should have one peak from each observation. Speaq 2.0 also applies two verification steps during clustering. The first is to merge neighbouring groups together and regroup as this allows for large shifts in the spectrum. After this the Jaccard index for each pair of neighbouring groups is calculated.

Definition 2.2.3 (The Jaccard Index)**.** Given two sets, X and Y, the Jaccard index $J(X, Y)$ is given by

$$J(X, Y) = |X \cap Y| \ / \ |X \cup Y|$$

Where this index is low the neighbouring groups are considered as one group. Otherwise they are kept separate.

Silhouette Scores A measure for the quality of the clustering, silhouette scores (Definition 2.2.4) are calculated. These scores provide a measure of how well the clustering is for each peak within observations, by comparing cohesion with separation, and provide scores between -1 and 1. A score of 1 means that the peak is well matched to the other peaks in its cluster, and poorly matched to the other clusters. A score of -1 means the opposite.

The mean SSs for each group can be calculated, and then by comparing this to the SNR it is possible to identify regions which might not be well aligned.

Definition 2.2.4 (Silhouette Scores)**.** Within each region the mean distance between point p_j and all other points in the region C_k, $a(p_j)$, and the dissimilarity between neighbouring regions $b(p_j)$ are calculated by

$$a(p_j) = \frac{1}{|C_J| - 1} \sum_{p_k \in C_J, p_j \neq p_k} d(p_j, p_k)$$

and

$$b(p_j) = min_{k \neq j} \frac{1}{|C_k|} \sum_{p_k \in C_k} d(p_j, p_k)$$

where $d(p_i, p_j)$ is the distance between points p_i and p_j.

From these two scores the silhouette score for each point is calculated by

$$s(p_j) = \frac{b(p_j) - a(p_j)}{max\{a(p_j), b(p_j)\}}$$

providing scores between 1 and -1.

2.2.2.6 Normalisation

Normalisation is used to correct for differences in overall concentration in order to make observations comparable with each other. In addition to variation due to the amount of material in the observations, differences in factors such as pulse calibration can cause inter-observation variation. Normalisation scales each spectrum by a derived constant, and there are several methods to do this, such as scaling to an internal standard or scaling so that the total sum is the same for each spectrum [81]. Alternative methods also exist, such as histogram matching. In histogram matching each spectrum is transformed into a histogram, and then each histogram is compared to a reference spectrum histogram and scaled to minimise the dilution factor.

Vu et al. compared normalisation methods in terms of their ability to recover true peak intensities from experimental NMR data with added random noise and varying dilution factors [82]. They found that normalising to a constant sum performed best, whilst the worst results were obtained for histogram matching.

2.2.2.7 Batch Correction

For large-scale chemometric assessments data may be collected in separate batches. With methods such as LC-MS, a large proportion of variance can sometimes be a direct result of these batches and therefore batch correction is often performed, usually involving quality control (QC) observations [14]. NMR however has been shown to have minimal intra-batch variation, with Dumas [83] showing that a large proportion of reproducibility errors are a result of specimen handling in-homogeneity.

It is a safe assumption that the intensities should be consistent across each analyte within QCs so that, where drift is present, a linear correction can be determined from the QCs and applied to each observation. A popular method proposed by Hendriks et al. [84] uses Analysis of Covariance to achieve this. In this method, the i^{th} observation with uncorrected intensity x_u is corrected using the formula

$$x_{c,i} = x_{u,i} - \overline{x_u} + \hat{x_u}$$

where $x_{c,i}$ is the corrected intensity for the i^{th} observation, $\overline{x_u}$ is the mean intensity for this analyte over all observations, and $\hat{x_u}$ is the intensity predicted by the linear regression model obtained from the QCs.

When QCs are not available batch correction is still possible. In 2016 Rusilowicz [85] showed that the removal of the background trend, estimated as a smooth function over all observations, can be effective. Multiple trend functions were compared for the background correction, and a simple moving average was found to be the most effective in terms of reducing the relative standard deviations between replicates. This method can also account for large within batch differences which are not represented by the QCs and which can be made worse by QC batch correction. In order to see which method is the most suitable, the results from both methods can be compared to find which has the lowest variance from batch differences.

2.2.2.8 Variable Scaling

As larger peaks naturally have greater variance than small peaks [86], variance-based methods such as PCA can be dominated by high abundance metabolites. Similarly, large variables will contribute more to distance metrics in cluster analyses, whereas low intensity variables might provide clearer differences between groups. Mean-centering is a column operation in which the mean value is subtracted to give each variable a zero mean. This removes the offset between high and low abundance metabolites, but does not change the variance, and is used in combination with variable scaling techniques [87]. For example, autoscaling / UV-scaling, widely used in metabolomics, involves dividing the mean-centred variable by the standard deviation, giving each variable unit variance (UV-scaling). This method of scaling allows all variables to have equal influence in multivariate analysis, but can also scale up unwanted noise peaks and so it is often used in parallel with feature extraction techniques. Pareto Scaling, in which the standard deviation is replaced by its square root, is frequently used as a less intensive scaling method as larger variables are reduced more than smaller variables. VAST (variable stability) scaling weights down the influence of variables with greater variance, giving greater influence on variables that change less [88]. Other scaling techniques include range scaling, which uses the range over which a variable is observed as a scaling factor, and level scaling which uses the mean.

In a comparison of scaling techniques, range and autoscaling were found to give the most biologically sensible results in PCA [87]. Other methods were found to be too dependent on the mean or the fold change, and VAST scaling led to results that were able to interpreted with relation to changes within the spectrum unlike UV-scaling, which is not interpretable. Purohit et al. found that a generalised log transformation of the variables produced more normally distributed data that was more suitable for multivariate analysis [88]. An additional parameter introduced into this so-called Glog transformation to reduce the scaling of noise was shown to improve classification results in comparison to auto-scaling and Pareto scaling with both 1D and 2D NMR data-sets [89]. However, results vary between studies and there is no one-size-fits-all scaling technique.

2.2.2.9 Data Fusion

Data sets can be complemented with data from other techniques to produce a single model or decision, in a process called data fusion [90], with data blocks usually scaled independently to ensure that neither data set dominate. Data fusion is generally done in three different tiers of data fusion [91]. Low-level fusion involves all the data from each source being concatenated into a singular block. This has the benefit of having the least amount of set up required, however the resulting data sets can often be extremely large, as well as many of the variables not being of any interest. Mid-tier fusion involves procedures such as feature extraction occurring before concatenation occurs, and high-tier fusion is where results from individual models are compared and combined together. While these methods contain fewer variables than the low-level fusion, they both require suitable pre-processing, and for high-level it also requires analysis to be performed on the individual data sets. Of the three tiers mid level is the most popular with NMR data due to the large number of variables this method produced and is used in a wide range of methods such as identifying tomato growth methods and milk identification [92], [93].

Examples of approaches for data fusion include time-of-flight mass spectrometry and Orbitrap-MS data were fused with ^{1}H NMR observations using mid-level data fusion by Spiteri et al [94]. To negate block bias, scaled eigenvalues and eigenvalues were used in place of scaling. This works by multiplying standardised eigenvalues by the percentage of variance each component accounts for (calculated from the eigenvalues), prior to combining the data sources. This fusion was shown to provide perfect classification for classification of honey origin, which was not possible with the raw data. When using several different data blocks it is possible to do different fusions in the workflow, as is the case with Ríos-Reina et al. [95]. For example, mid-infrared and near-infrared spectroscopy (NIR) data sets have been combined via low level fusion, and subsequently combined with peak areas from ^{1}H-NMR data as well as excitation-emission matrices. This combination was used to test wine vinegar origins, and was shown to be an improvement on models based on just one data set. One advancement on data fusion is called Fused Adjacency Matrix approach [96], where several distance matrices are created on extracted data, and these are then fused to give a matrix weighted on how frequently pairs of observations are similar. When applied to three data sources: visual, NIR, and NMR data, it is shown to find groupings which are more evident than just standard mid-level fusion.

Statistical Total Correlation Spectroscopy (STOCSY) is a method devised by Cloarec et al. [23] in 2005 which makes use of a correlation matrix C. Correlations between variables in two data sets are calculated to create a pseudo-two-dimensional dataset, and the patterns within this can be used to aid identification. In recent years STOCSY has been used in a variety of metabolomic analyses, such as in 2020 it was used by Beteinakis [97] for identifying biomarkers in olives to aid with classification of origin and treatments. After performing initial analysis on their data they select variables with were found as important with regards to their aims, and use these as the driving variables for STOCSY. This was able to locate peaks which lead them to identifying Hydroxytyrosol and Tyrosol which are both known to be present in olives. One limitation STOCSY has is its inability to work across various data sets. This was however remedied in 2006 by Crockford et al. [24] in a method developed from STOCSY, called Statistical Heterospectrosopy (SHY). The Pearson correlation coefficient is calculated for each pair of variables between two data sets. For each of these pairs the null hypothesis that these features have no relationship is tested, and a p-value calculated. Where either the correlation coefficient or the p-value is below a threshold (which the authors suggest 0.001), the correlation is discarded. The resulting correlations are then assumed to be from either the same or related compounds, and so this can be used as a tool for finding chemical information about the compounds and aids with molecular identification.

2.3 Data Analysis Methods

2.3.1 Supervised Vs Unsupervised Data Analysis

Most methods for machine learning can be categorised into one of two approaches, which are unsupervised and supervised learning. In supervised learning the method utilises the response vector, usually in the form of labels, in order to be able to estimate the response from the data, often inspecting the structure of the data in relation to this. Unsupervised methods do not use the response vector, but can be used to reveal patterns and clusters in the data.

2.3.2 Unsupervised Methods of Data Analysis

Unsupervised techniques are used for exploratory data analysis, and can be used to investigate patterns in the data, or identify potential outliers. Clustering methods classify observations into a number of groups relating to the data structure, or dimensionality reduction, where the number of variables are reduced in order to improve efficiency of further calculations. Unsupervised techniques are also often used in data visualisation, as these underlying latent structures can often provide insight into the relationships between variables. A small number of these structures can help explain significant proportions of the data, and so plotting them help with identifying the trends.

2.3.2.1 Principal Components Analysis

Principal Components Analysis (PCA) is an unsupervised technique, and is one of the most widely used multivariate techniques [98]. This is a result of its key property, which is that it reduces the dimensionality of a given data set into a few principal components which can account for a large proportion of the variance via changing the basis of the data. The components can then be used to quickly visualise the data as the variations which contribute to the first few PCs can highlight key features such as large separations between classes, or unwanted features such as shifted data or analytical issues such as batch separation.

First described by Karl Pearson in 1901 [99], and then in 1933 Harold Hotelling developed Pearsons' work into the widely used algorithm [100]. The theory behind this statistical method is, within a observation data set by considering m variables, $(x_j)_{j=1}^{m}$, these variables might have a more generalised set of independent variables behind them. These independent variables are calculated by finding the linear combination of the form $P = \sum_{j=1}^{m} \alpha_j x_j$ with the greatest spread of data, where the α_i's are the loadings of the variables, and the variables are mean centred. For n observations, the variance of these linear combinations can be given in the form

$$\begin{aligned} var(P) =& \frac{1}{n}\sum_{i=1}^{n}(p_i - \overline{p})^2 = \frac{1}{n}\sum_{i=1}^{n}(p_i)^2 \\ =& \frac{1}{n}\sum_{i=1}^{n}(\sum_{j=1}^{m}\alpha_j x_{i,j})^2 = \frac{1}{n}\sum_{i=1}^{n}(\sum_{j=1}^{m}\alpha_j x_{i,j})(\sum_{k=1}^{m}\alpha_k x_{i,k}) \\ =& \sum_{j=1}^{m}\sum_{k=1}^{m}\alpha_j\alpha_k\frac{1}{n}\sum_{i=1}^{n}x_{i,j}x_{i,k} = \sum_{j=1}^{m}\sum_{k=1}^{m}\alpha_j\alpha_k cov(x_j, x_k) \\ =& \boldsymbol{\alpha}^T\boldsymbol{\Sigma}\boldsymbol{\alpha} \end{aligned}$$

which is what is needed to be maximised, subject to $\sum_{j=1}^{m} \alpha_j^2 = 1$. To find the maximum of this variance Lagrange multipliers are used, subject to

$$F = \alpha^T\boldsymbol{\Sigma}\alpha + \lambda(1 - \alpha^T\alpha)$$

The maximum of this occurs at

$$\begin{aligned} \frac{\partial F}{\partial \boldsymbol{\alpha}} = \frac{\partial}{\partial \boldsymbol{\alpha}}\left(\boldsymbol{\alpha}^T\boldsymbol{\Sigma}\boldsymbol{\alpha} + \lambda(1 - \boldsymbol{\alpha}^T\boldsymbol{\alpha})\right) &= 0 \\ 2\boldsymbol{\Sigma}\boldsymbol{\alpha} - 2\lambda\boldsymbol{\alpha} &= 0 \\ (\boldsymbol{\Sigma} - \lambda\mathbf{I})\boldsymbol{\alpha} &= 0 \end{aligned}$$

and so the problem is reduced to finding the eigenvectors and eigenvalues of the covariance matrix. Here λ_i will be the eigenvalues and α_i will be the associated normalised eigenvectors. The first principal component will give the best one-dimensional approximation, and by adding more linear combinations this approximation improves. Each component is chosen such that each successive component describes the most amount of variance which has not already been described, and this is achieved by each subsequent component being orthogonal to the previous components. Often, to keep the number of dimensions down, only the first few components will be used. Depending on the analysis the number of components chosen will vary. If trying to find differences within the observations the first few components can be used for visualisation, and if the aim of the analysis is something more complex (such as classification) the number of components used is chosen such that a certain proportion of the variance is accounted for.

PCA is an unsupervised method that can be used to identify patterns in the data and the coefficients, or loadings, for the relevant principal component can be used to determine the original variables contributing most to any patterns or trends. PCA can also be used to identify potential outliers that may skew the results in further analysis. If a variable contains a potential outlier the variance within this will be greater, and therefore the contribution to the principle components will be increased. Depending on the severity of the outlier, when the principle components are plotted this variable might stand out due to a higher loading, and the observation with the potential outlier might also appear separated from the others.

When PCA reveals patterns in the data, the loadings for the relevant principal components (PC) can be used to determine the spectral features responsible. However, the loadings often show that very many features contribute similarly to the variance so that the results sometimes are difficult to interpret in terms of metabolomic changes.

Example PCA calculations In order to provide an example of how principal component analysis can work the widely known Fisher's Iris data set [22] is analysed here. This data set contains 150 observations of 3 iris flowers types, *Setosa, Virginica* and *Versicolor*. For these observations 4 measurements were taken, which are the lengths and widths for both the petals and the sepals, recorded in cm. This data is shown in Figure 2.17, with each variable plotted against each other, coloured by flower type. While there is minimal correlation between the sepal measurements, there is a strong positive correlation between the two petal measurements. The setosa observations also display separation from the other two varieties in the petal measurements.

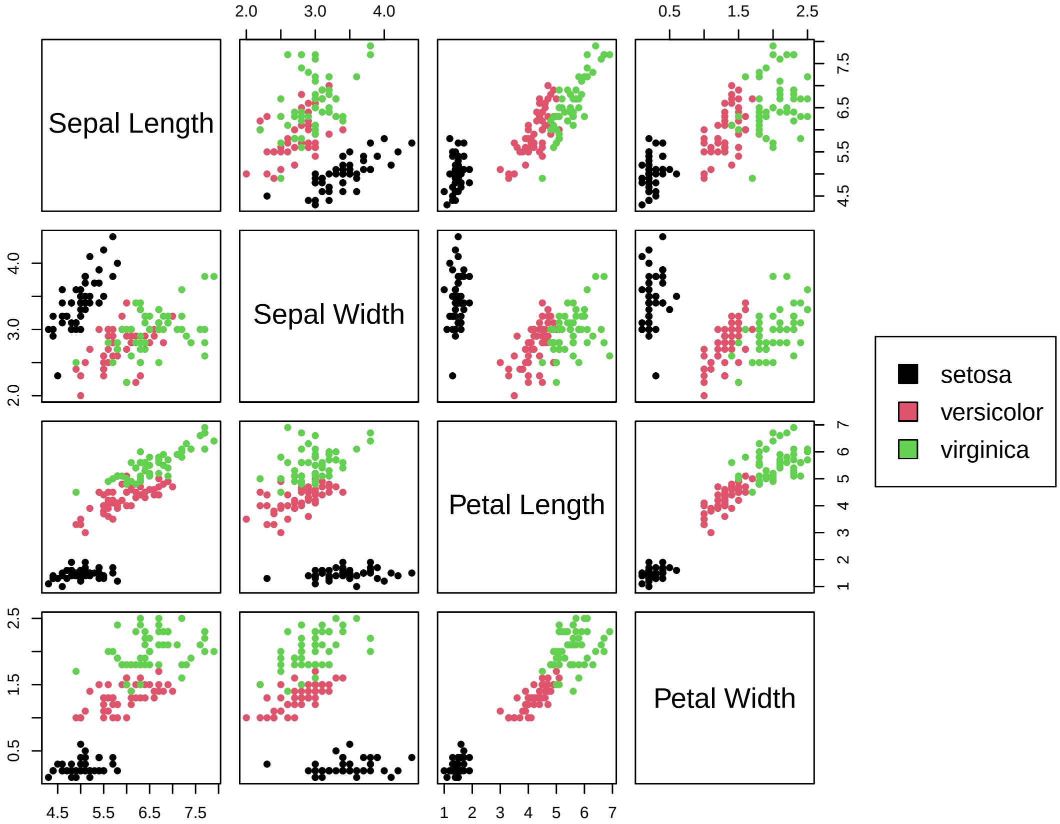


Figure 2.17: Scatter plots showing the four variables for the iris data set, coloured by flower species. The plots show some separation with the petal variables, and minimal separation with the sepal variables.

With 150 observations the initial step is to calculate the covariance matrix from the 4 variables. This is shown in Table 2.1.

	Sepal Length	Sepal Width	Petal Length	Petal Width
Sepal Length	0.686	-0.042	1.274	0.516
Sepal Width	-0.042	0.190	-0.330	-0.122
Petal Length	1.274	-0.330	3.116	1.296
Petal Width	0.516	-0.122	1.296	0.581

Table 2.1: Covariance Matrix of the iris data set, round to 3dp.

From here, the eigenvectors and eigenvalues need to be calculated. Here the eigenvalues are calculated as λ as 4.228, 0.243, 0.078, 0.024 with corresponding eigenvectors of

$$\alpha_1 = \begin{bmatrix} 0.361 \\ -0.085 \\ -0.857 \\ 0.358 \end{bmatrix}, \alpha_2 = \begin{bmatrix} -0.657 \\ -0.730 \\ -0.173 \\ 0.075 \end{bmatrix}, \alpha_3 = \begin{bmatrix} 0.582 \\ -0.598 \\ -0.076 \\ -0.546 \end{bmatrix}, \alpha_4 = \begin{bmatrix} 0.315 \\ -0.320 \\ -0.480 \\ 0.754 \end{bmatrix}$$

Using these eigenvectors as coefficients in a linear model the principal component scores S can be calculated via $\alpha^T X^T = S^T$. The first of these components accounts for 92.5% of the variance, with the second adding just 5.3% showing that a large proportion of the variation in the iris data set can be achieved with only two principal components. This allows for clear informative visualisation (Figure 2.18), and can increase the efficiency of further calculations. Figure 2.18 shows a clear separation between the flower species predominantly in the first component, and inspection of the loadings (in blue) show that both the petal variables are significant contributors to this. Comparing this result to Figure 2.17 it is clear to see how these variables contribute to the variance relating to species.

2.3.2.2 Clustering

k-means Clustering In order to perform clustering based on the patterns found within the data one of the common methods is to apply k-means clustering. Here k clusters are formed from the data, and each observation is assigned into one of these. This method is centroid-based, meaning that it aims at minimising the distance between each data point in the observations, and the centroids describing the clusters. After selecting a suitable value for k, k observations are randomly selected to act as the centroids. For all the remaining observations these are assigned to their closest centroid. For each of the clusters the centroids are re-calculated from all of the observations within. This process is then repeated, starting we re-assigning each observation to the closest centroid until the convergence criteria are met. These criteria usually involve a number of acceptable changes (often 0) or a maximum number of iterations.

Example k-means clustering on iris data To illustrate the k-means algorithm the iris data set is once again used. Here k is chosen to be 3, as this matches the number of species in the data (although, as this is unsupervised this is technically not known). Three observations are randomly selected as the initial centroids, and after one iteration the clusters illustrated in Figure 2.19 are formed. Due to the fact this data consists of four variables, the data is displayed using the first two principle components. As seen the setosa variables account for 50/53 observations assigned to the first cluster after just this first step. There is still considerable overlap between

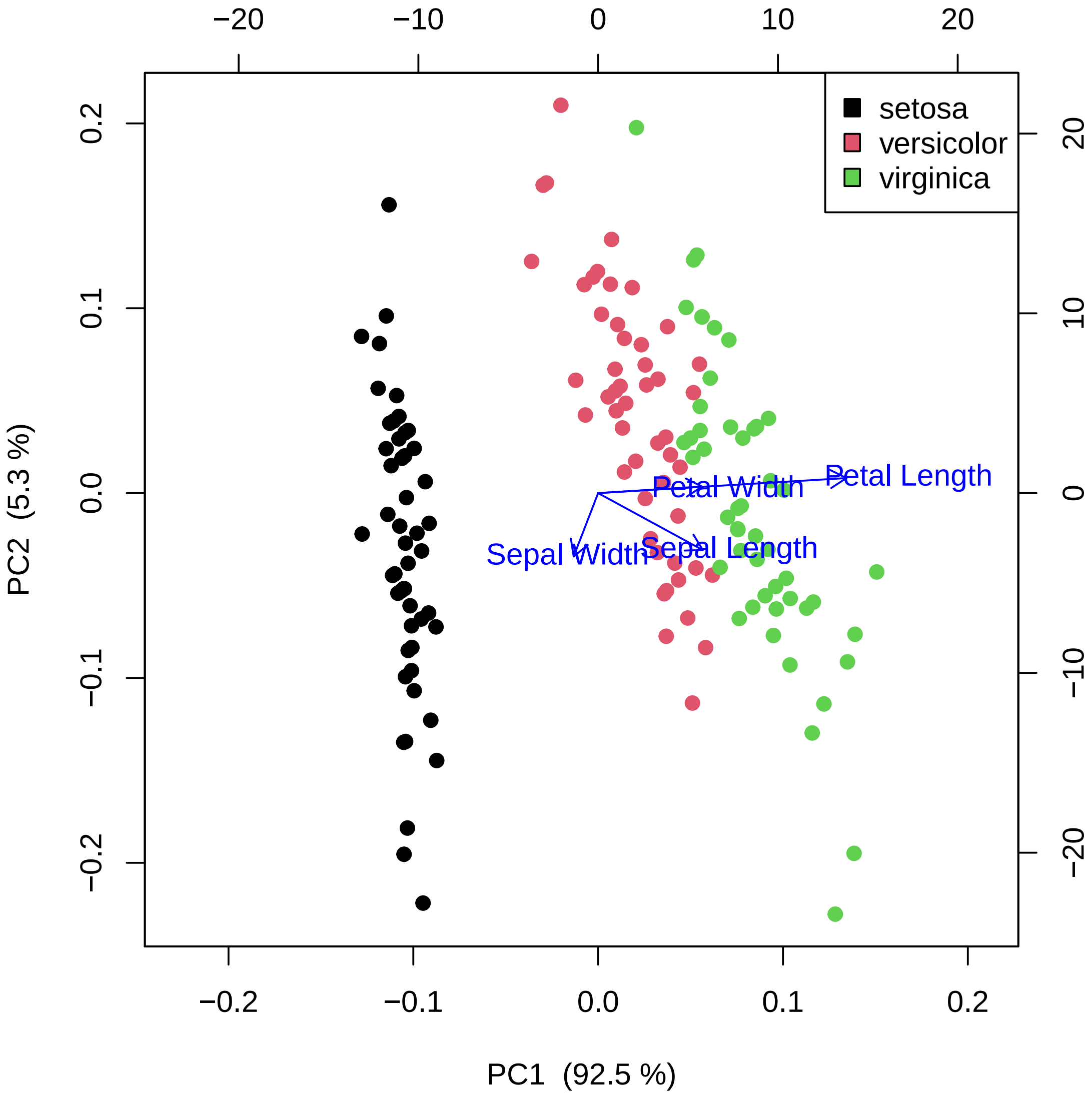


Figure 2.18: Biplot of the Iris data set representing the scores and loadings. The points represent the scores of each observation in the relevant PC coloured by flower species, and the blue lines represent the loadings with their direction indicating how influential they are within the corresponding principal components.

the remaining observations, with no clear separation formed. Continuing this the algorithm converges after just 2 iterations, with the results of clustering are shown in Table 2.2. For each of the three clusters the modal class (the class with the greatest frequency) can be considered the assigned class. As expected the setosa observations are all classified well into Cluster 1, and there is still confusion between the remaining two species.

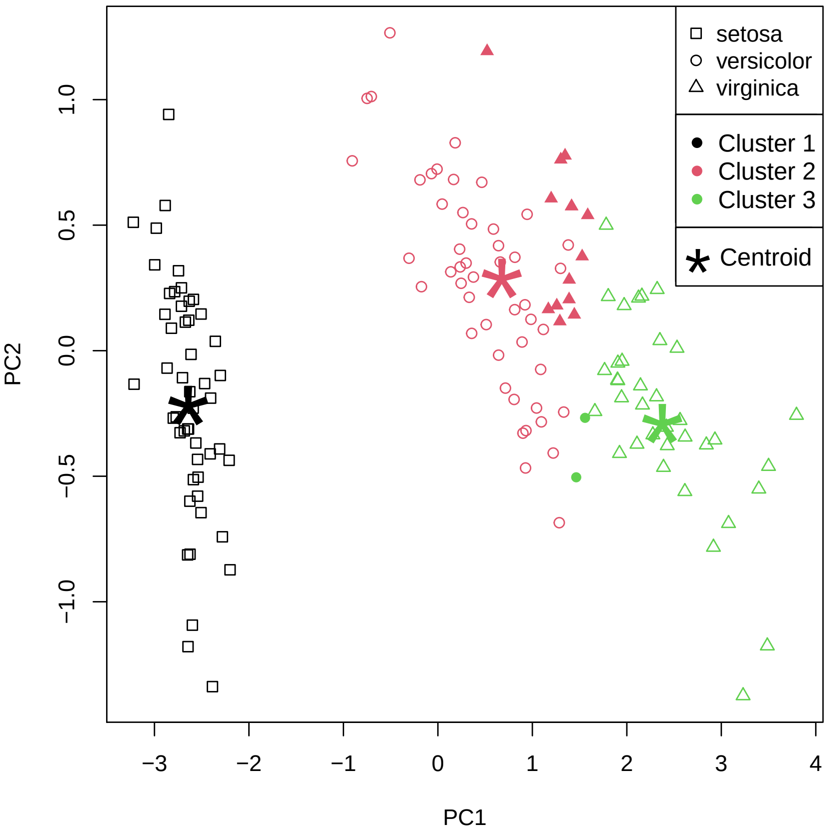


Figure 2.19: PCA scores for the iris data set, with plotting symbol representing species. The colour used for each observation is based on the cluster they belong to after just one iteration of k-means clustering, showing how powerful and efficient this method can be. The centroids for each of the three clusters is also shown by a star, coloured based on which cluster they represent. Observations belonging to the modal class are not filled, with the others within the cluster filled.

		Reference		
		Setosa	Versicolor	Virginica
Prediction	Cluster 1	50	0	0
	Cluster 2	0	48	14
	Cluster 3	0	2	36

Table 2.2: Confusion matrix showing the number of observations assigned to the k-means clusters after the algorithm terminates, taking 2 iterations.

2.3.3 Supervised methods of Data Analysis

2.3.3.1 Methods of Model Validation

Once models are created they need to be validated to ensure that they are able to appropriately model the data and are able to predict other observations accurately. The basic approach for this is a method called hold-out, where the data is split into two distinct groups, a training set where the model is estimated from, and a test set where the model is independently checked against. By using this test set it is possible to check for over- or under-fitting. Over-fitting is when the model can correctly predict a large proportion of the training data, but when applied to other data it performs poorly, often dropping by a significant accuracy. This is a result of the model considering noise as a valid feature of each observation, and so does not generalise to other data sets. Under-fitting is when the model is too simple to be able to predict the response for both sets of data. Using the hold-out method Shao et al. [101] show that 100% accuracy on a training set can easily drop to 85% for the test set, showing over-fitting. The hold-out method is not always appropriate, for instance if there is a low number of observations, and there are many other methods available.

The most common alternative method is k-fold validation. Here, data is split into k equal sets, and as with the hold-out method one of these is removed to act as a test set, and the remaining $k - 1$ sets are used to generate a model. The error rate is recorded, and then the process is repeated removing a different set k times. An average error rates is calculated, and this is used as the error rate of the model. As the number of folds k becomes smaller, the training sets used become larger. This means that the model can theoretically have an improved fit for the data. This can lead to over-fitting, and so careful consideration is needed when choosing k, and the choice will depend on the number of observations. The extreme version of this method occurs when $k = n$, called leave-one-out (LOO) validation, which is often used when the number of observations is small. This should only be used when there is a low quantity of observations available, as it can sometimes produce exaggerated accuracies.

2.3.3.2 Partial Least Squares

The non-linear iterative partial least squares (NIPALS) algorithm developed by Wold [102] was originally used to calculate principal components without the need for the covariance matrix. It was later adapted as an alternative to principal components regression (PCR) for over-determined regression problems [103].

Partial Least Squares (PLS) works similarly to PCA, however the response is also accounted for when decomposing the data into the components, positively weighting variables which correlate with it. This supervised technique creates X-scores, $\mathbf{t}_a, (a = 1, 2, \ldots, A)$ where A is the number of components in the model, and X-loadings given by $\mathbf{p}_a$, similar to how PCA works, but it also calculates X-weights denoted by $\mathbf{w}_a$, Y-scores given by $\mathbf{u}_a$, Y-loadings, and Y-weights $\mathbf{c}_a$. Due to the response being included this method is often used when the analysis in question is to find separation between the observations, and to predict the outcome. Similar to PCA, good approximations to the predictor variables can often be obtained from just a few linear combinations of the predictor variables with coefficients known as X-loadings. Similarly, good approximations can be obtained for multivariate response variables, in terms of Y-scores with associated Y-loadings and the relationship between $\mathbf{X}$ and $\mathbf{Y}$ is ensured by a regression equation. The loadings show the relationship between the scores and the original variables and can be used to identify the spectral features most related to the response variables.

Wold's algorithm, originally published in Wold's 2001 paper [104] is presented here. Wold states that the X-scores are orthogonal, and can be used as predictors of $\mathbf{Y}$ and also model $\mathbf{X}$. This is as a result of the assumption that they can both partially be modelled by the same latent variables (LV).

The X-scores are estimated by linear combinations of the original variables $\mathbf{x}_k$ with the weights w^*_{ka}. As with Wold's paper, formulas are shown in both element and matrix form.

$$t_{ia} = \sum_k W^*_{ka} X_{ik}; \qquad \mathbf{T} = \mathbf{X}\mathbf{W}^* \tag{2.9}$$

These X-scores have two interesting properties:

(A) When multiplied by the loadings p_{ak}, the scores provide good approximations of $\mathbf{X}$, such that the X-residuals, e_{ik} are small, where

$$X_{ik} = \sum_a t_{ia} p_{ak} + e_{ik}; \qquad \mathbf{X} = \mathbf{T}\mathbf{P}' + \mathbf{E}$$

When $\mathbf{Y}$ is multivariate, the Y-scores $\mathbf{u}_a$ after being multiplied by the weights c_{am} are good approximations of $\mathbf{Y}$ such that the residuals g_{im} are small, calculated such that

$$Y_{im} = \sum_a u_{ia} c_{am} + g_{ik}; \qquad \mathbf{Y} = \mathbf{U}\mathbf{C}' + \mathbf{G}$$

(B) A good predictor of $\mathbf{Y}$ is provided by the X-scores. I.e.

$$Y_{im} = \sum_a c_{ma} t_{ia} + g_{ik}; \qquad \mathbf{Y} = \mathbf{TC}' + \mathbf{F} \tag{2.10}$$

where $\mathbf{F}$ constitutes of the Y-residuals, f_{im}.

From Equation (2.9), Equation (2.10) can be rewritten in the form of a multiple regression model:

$$Y_{im} = \sum_a c_{ma} \sum_k w^*_{ka} x_{ik} = \sum_k b_{mk} x_{ik} + f_{im} \qquad \mathbf{Y} = \mathbf{XW}^*\mathbf{C}' + \mathbf{F} = \mathbf{XB} + \mathbf{F}$$

where b_m ($\mathbf{B}$) denotes the PLS-regression (PLS-R) coefficients.

PLS-R Algorithm There are several algorithms for PLS based on the shape of the data. The method presented by Rännar et al. in 1995 [105] is appropriate for data with a low number of observations whereas the method of Lindgren et al. published in 1993 [106] is applicable to data with a large quantity of observations and a low number of variables. Here the original NIPALS algorithm is reproduced from Wold's 1984 paper [107]. After normalising the data matrices $\mathbf{X}$ and $\mathbf{Y}$, the following steps are performed:

A) Obtain an initialisation vector $\mathbf{u}$, which is usually chosen as the column of $\mathbf{Y}$ with the largest variance.

B) Calculate the X-weights $\mathbf{w} = \mathbf{X}'\mathbf{u}/\mathbf{u}'\mathbf{u}$, norm $\mathbf{w}$ to $||\mathbf{w}|| = 1$

C) Calculate the X-scores, $\mathbf{t}$, as $\mathbf{t} = \mathbf{Xw}$

D) Calculate the Y-weights, $\mathbf{c}$, as $\mathbf{c} = \mathbf{Y}'\mathbf{t}/\mathbf{t}'\mathbf{t}$

E) Calculate the updated Y-scores, $\mathbf{u}$, as $\mathbf{u} = \mathbf{Yc}/\mathbf{c}'\mathbf{c}$

F) From here, convergence is tested on the change in $\mathbf{t}$. Until convergence is reached repeat from B. If convergence is reached, continue with (G).

G) Remove the current component from $\mathbf{X}$ and $\mathbf{Y}$, and use these reduced matrices in the next component. This is done via
$\mathbf{p} = \mathbf{X}'\mathbf{t}/\mathbf{t}'\mathbf{t}$
$\mathbf{X} = \mathbf{X} - \mathbf{tp}'$
$\mathbf{Y} = \mathbf{Y} - \mathbf{tc}'$

H) Repeat from (A) until there is no more significant information in $\mathbf{X}$ about $\mathbf{Y}$.

Due to the nature of the PLS algorithm, if categorical data is to be used it must first undergo a transformation. The response vector is expanded into an $n \times l$ matrix, where n is the number of observations and l is the number of groups with exist within Y. Within each column the data is represented in a binary format, where 1 symbolises the observation belongs to that group, and 0 otherwise. This transformed data can then be used as a numeric response for each category without assuming a relationship between them, for instance the difference between category 1 and category 2 is now equal to the difference between category 1 and category 3. The predicted Y values from the PLS-DA model can range in value. The most common approach to assigning a category to the predicted value is to use the category with the greatest value. This algorithm is called Partial Least-Squares Discriminant Analysis (PLS-DA).

Orthogonal PLS Like PCA, the PLS model is obtained by maximising variance and can be difficult to interpret due to the fact they contain the linear combinations of the variables, found with respect Y. A modification of PLS, termed orthogonal PLS (O-PLS) has been proposed, which separates the variance in the model into two parts [108]. The first is the variation that is common to both the data matrix and the response matrix and is therefore of most interest in classification and prediction. The other part, the so-called "structured noise", is the variance specific to the data matrix and not related to the response matrix. Filtering out the uncorrelated noise leads to a model that is easier to interpret and allows the structure of the noise to be analysed separately, for example using PCA. In classification, with a response matrix consisting of zeros and ones indicating class membership, O-PLS separates the within-groups variance and the between groups variance. The method can also be applied to time series with time as the response to extract the variance related to time [108].

VIP Scores Variable Importance in Projection (VIP) scores are a metric which is used to measure how important each variable is within a number of components. It combines the scores and loadings via

$$\mathrm{VIP}_j = \sqrt{p \sum_{k=1}^{h} \left(SS\left(c_k \boldsymbol{t}_k\right) \left(\boldsymbol{w}_{jk} / \left\|\boldsymbol{w}_k\right\|\right)^2 \right) / \sum_{k=1}^{h} \mathrm{SS}\left(c_k \boldsymbol{t}_k\right)} \tag{2.11}$$

where $SS\left(c_k \boldsymbol{t}_k\right) = c_k^2 \boldsymbol{t}_k' \boldsymbol{t}_k$. A standard threshold of 1 is applied to determine the importance of a variable, as this is the mean squared VIP score. This therefore selects all of the above average variables, however in the majority of analyses only the largest VIP scores are selected.

Example PLS-DA calculations For demonstration, PLS-DA is applied to a training set formed from 67% (n = 100) of the iris data observations. After convergence is achieved the root mean squared error of prediction (RMSEP) can be calculated for different numbers of components to determine the appropriate number to use in analysis. For the RMSEP the number of components is usually chosen as the lowest number that for any number of components added after, there is a negligible improvement in the RMSEP. For both the Setosa and Virginica iris species the RMSEP plateaus, where as for Versicolor there is no notable decrease, the greatest being at 2 components as can be seen in Figure 2.20. It is not necessary to add the third component which would create a more complex model for minimal return.

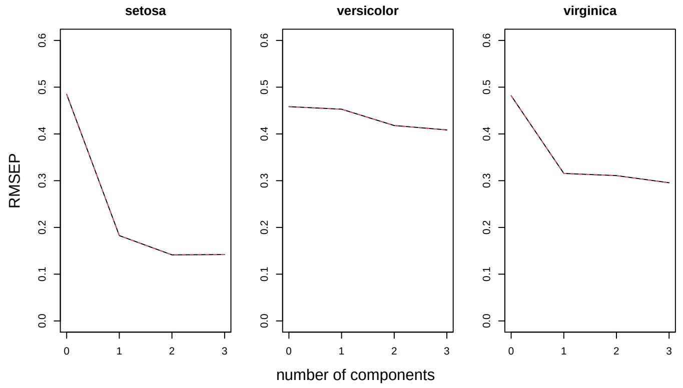


Figure 2.20: Root mean squared error of prediction (RMSEP) for PLS-DA components in the iris analysis for each of the species.

An accuracy of 84% is achieved from the test set consisting of the 50 observations not used to calculate the model, with the confusion matrix shown in Table 2.3. As seen in the PLS and k-means models there is still overlap between the Versicolor and Virginica observations, and the Setosa variables are all able to be correctly classified.

		Reference		
		Setosa	Versicolor	Virginica
Prediction	Setosa	16	0	0
	Versicolor	0	9	5
	Virginica	0	3	17

Table 2.3: Confusion matrix of the test set of data for PLS-DA on the iris data set.

Biplots can also be used to show how the X and Y loadings and scores interact with each other. Figure 2.21 shows both the X scores and X loadings for the iris data. Component 1 unsurprisingly has a clear separation based on species as seen in PCA. The X loadings show that petal length is the key cause of variance relating to species, as this has the greatest absolute loading within the first component, which is where the separation between species lies. The sepal variables are significant in the second component which shows no separation due to species.

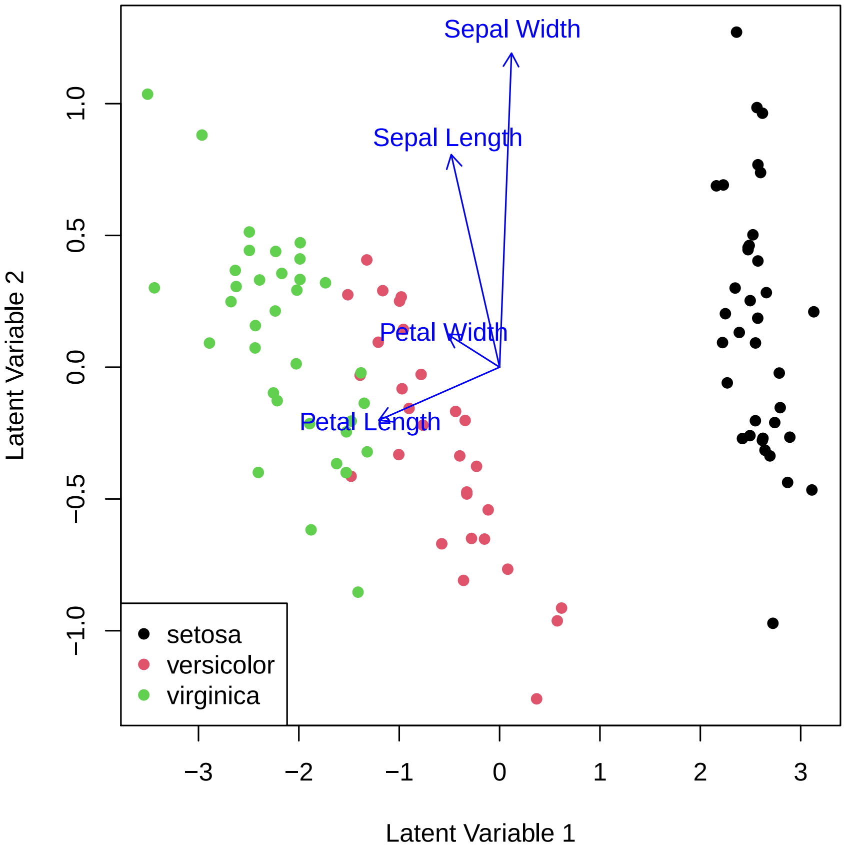


Figure 2.21: Biplot showing X scores and X loadings from PLS-DA of the iris data with data points coloured according to species. The X scores show a clear separation within the first component.

VIP scores can also be calculated for each of the variables as shown in Figure 2.22 for the iris data. The variable with the largest VIP score (and the only one above the threshold 1) is petal length, showing that this is a significant variable in terms of discriminating the species.

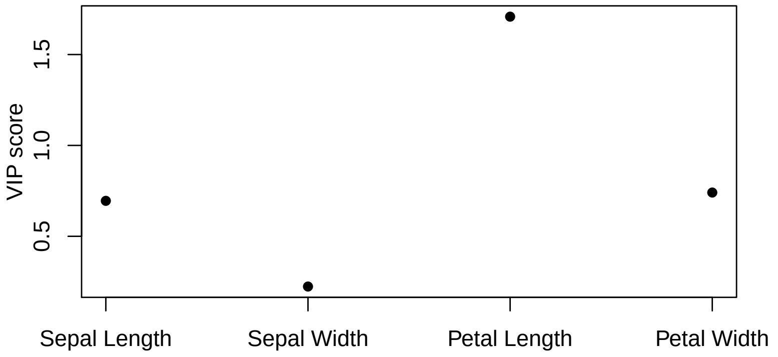


Figure 2.22: VIP scores for each of the 4 iris variables. Petal Length has a noticeably larger score than the other 3 variables, suggesting the variance within this variable is similar to that in the response and as such is useful for discrimination.

2.3.3.3 Linear Discriminant Analysis

Another supervised method which can be used in classifying data is Linear Discriminant Analysis (LDA). As opposed to PCA and PLS which analyse the variance in the data, this technique looks for the greatest separation between groups. The most common method of LDA is Fisher's technique [22] which maximises the between group variance whilst minimising the within group variance. This is achieved by maximising the function

$$\mathcal{F}(\alpha) = \frac{\alpha^T S_B \alpha}{\alpha^T S_W \alpha}$$

where S_B is the between-group scatter matrix defined by

$$S_B = \sum_{i=1}^{C} (\mu_i - \overline{\mathbf{x}})(\mu_i - \overline{\mathbf{x}})^T$$

with C being the number of classes, $\overline{x}$ the grand mean, and μ_i is the mean of class i. S_W is then defined as

$$S_W = \sum_{j=1}^{C} \sum_{x \in c_j} (\mathbf{x} - \mu_j)(\mathbf{x} - \mu_j)^T.$$

$\mathcal{F}$ is invariant under scaling of the vector $\alpha \to \beta\alpha$, and so α is chosen such that $\alpha^T S_W \alpha = 1$. As such in order to maximise $\mathcal{F}$ it is equivalent to solving

$$\min_\alpha \quad -\frac{1}{2}\alpha^T S_B \alpha$$

subject to $\alpha^T S_W \alpha = 1$. Using Lagrangian multipliers this can be shown to be solved via

$$\mathcal{L}_P = -\frac{1}{2}\alpha^T S_B \alpha + \frac{1}{2}\lambda(\alpha^T S_W \alpha - 1).$$

Differentiating with respect to α results in

$$\frac{\partial \mathcal{L}_P}{\partial \alpha} = -S_B \alpha + \lambda S_W \alpha = 0,$$

and for non-singular S_W a solution lies at $S_W^{-1} S_B \alpha = \lambda \alpha$ which is a generalised eigen-value problem, the solution of which provides α.

Both PCA and PLS-R are commonly used with linear discriminant analysis, where the scores for the first few components are used in place of variables. While PCA is commonly used to quickly identify separation between groups, it has been shown in a number of studies that PLS-DA is a superior method for classification based on accuracies. Barker and Rayens [109] used simulated data to show that as

	Sepal Length	Sepal Width	Petal Length	Petal Width
Setosa	5.01	3.42	1.46	0.25
Versicolor	5.94	2.77	4.26	1.33
Virginica	6.59	2.97	5.55	2.03

Table 2.4: Group means for the iris data, shown to 2dp.

S_W	Sepal Length	Sepal Width	Petal Length	Petal Width
Sepal Length	38.96	13.63	24.62	5.65
Sepal Width	13.63	16.96	8.12	4.81
Petal Length	24.62	8.12	27.22	6.27
Petal Width	5.65	4.81	6.27	6.16

S_B	Sepal Length	Sepal Width	Petal Length	Petal Width
Sepal Length	63.21	-19.95	165.25	71.28
Sepal Width	-19.95	11.34	-57.24	-22.93
Petal Length	165.25	-57.24	437.10	186.77
Petal Width	71.28	-22.93	186.77	80.41

Table 2.5: Scatter matrices for the iris data, shown to 2dp.

the noise within variables increases the accuracy greatly decreases for PCA, however PLS was only slightly affected. Shao and Li [101] achieved similar results in the analysis of fruit and vegetable firmness.

Example LDA calculations The iris data is used again for this example, split into a training set consisting of 100 out of the 150 observations, and a test set with the remaining 50. The first step is to calculate the group means for each of the three flower species, shown in Table 2.4.

From here S_W and S_B scatter matrices can be calculated, shown in Tables 2.5.

Using the scatter matrices and the formula $S_W^{-1} S_B \alpha = \lambda \alpha$ the eigenvectors can be calculated, shown in Table 2.6. For the last two components the eigenvalue is 0, and so only the first two eigenvectors are shown.

	LD1	LD2
Sepal Length	-0.21	-0.01
Sepal Width	-0.39	-0.59
Petal Length	0.55	0.25
Petal Width	0.71	-0.77

Table 2.6: First two eigenvectors for iris data set, rounded to 2dp.

Using these coefficients it is possible to predict the classes for each of the observations. The confusion matrices for the training and test set predictions are shown in Tables 2.7 and 2.8 respectively. For the training set this achieves 98% accuracy (n = 100) with just one Veriscolor classified as Virginica and one Virginica classified as Versicolor. The separation of these two species is far less than for Setosa, as seen in scores plots for both PCA and PLS-R. The test set accuracy is also 98% (n = 50) with one Versicolor observation misclassified as Virginica, showing that over-fitting is not occurring.

			Reference	
		Setosa	Versicolor	Virginica
	Setosa	32	0	0
Prediction	Versicolor	0	35	1
	Virginica	0	1	31

Table 2.7: Confusion Matrix showing the training results of LDA modelling species, with 100 of the 150 iris observations as a training set.

			Reference	
		Setosa	Versicolor	Virginica
	Setosa	18	0	0
Prediction	Versicolor	0	13	0
	Virginica	0	1	18

Table 2.8: Confusion matrix of the test set of data for LDA on the iris data set.

Comparing the results from the LDA model to the PLS-DA model it suggests that LDA is superior to PLS-DA with regards to accuracy. PLS can find latent variables within the data which can help explain patterns within the data, and so in some cases can achieve similar or greater accuracies than LDA.

2.3.3.4 Tree-based Methods

A decision tree is a machine-learning algorithm, used for classification or regression, with branches reflecting the possible outcome of tests on the variables [110]. The variables tested are chosen to maximise the information gain, based on the reduction of uncertainly via a measure such as entropy. Paths through the tree, from the root node with the first test to leaf nodes giving the class or response, model the rules between the explanatory variables and the associated output. To reduce the possibility of over-fitting the data, decision trees can be *pruned* to remove branches having the smallest decrease in classification error.

In order to form a decision tree first the data is split at various points and the entropy is chosen for each of these splits S using the formula

$$E(S) = \sum_{i=1}^{c} -p_i \mathrm{log}_2 p_i$$

where p_i is the probability of event i occuring. In the case of multiple groups the entropy is combined via

$$E(S, X) = \sum_{c \in X} P(c)\ E(c)$$

where $P(c)$ is the probability that the observation belongs to group c. This entropy is then subtracted from the original entropy of the data set to calculate the information gain, and the split with the highest gain is chosen. This is then repeated for each branch until there is no further information gain, and the leaf nodes are formed.

Ensemble classifiers aim to improve the robustness of classification by averaging the results from multiple models. In order to attempt to improve on the robustness of decision trees, numerous trees formed from random subsets of the original data are collated together to form a random forest (RF) [110]. This way, if there exist only a few variables which classify the data well these will be selected in the majority of trees they exist in, and will stand out more than variables which do not contribute to classification.

Random forest do not produce a single final tree, and so it can be difficult to identify how much each variable contributes to the model. Multiple methods do however exist in order to calculate important variables from the random forests. The two common methods are to inspect the mean decrease in accuracy (MDA) for each variable, or the mean decrease in Gini index (MDG). Gini index is calculated via

$$Gini = 1 - \sum_{i=1}^{C} (p_i)^2$$

where p_i is the probability that a observation belongs to class i, and C is the number of classes. If a observation classifies well the gini index will be low, and so a high MDG implies the feature can be useful for classification.

Example tree-based calculation Despite the low number of levels in the response variable and low number of variables the iris data set is once again examined as an example of tree based methods. Initially a decision tree is made on the training set, and this is plotted in Figure 2.23, with no pruning applied. This shows that 3 leaf nodes are created, one for each of the flower species and only one variable, petal length, is used to discriminate between them. As seen previously with the loadings in PCA (Figure 2.17) a low entropy is to be expected for this variable, which is what is desired when creating a tree. For the training set this again achieves 98% accuracy (n=100), with the same observations misclassified as with LDA.

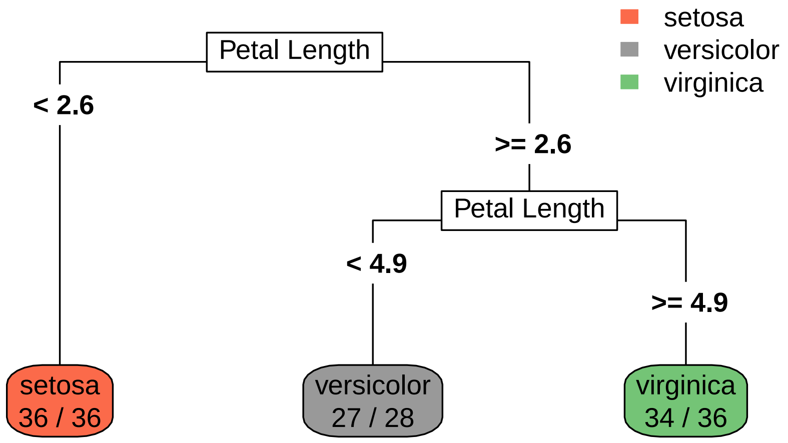


Figure 2.23: Decision tree of the iris data set showing 97% accuracy with three leaf nodes and two branch points.

Applying this tree to the test data gives a noticeable drop in accuracy, down to 90% (Table 2.9) which suggests that under or over-fitting could be occurring. This is potentially a result of only using a single variable in the decision tree, and so will more than likely be under-fitting. As such, this model is inappropriate to be used as a model to distinguish floral species in this iris data set.

		Reference		
		Setosa	Versicolor	Virginica
	Setosa	18	0	0
Prediction	Versicolor	0	12	3
	Virginica	0	2	15

Table 2.9: Confusion matrix of the test set of data for decision trees on the iris data set.

A random forest consisting of 500 trees was also used, with each tree consisting of 3 out of the 4 variables. As 25% of these will not be including petal length the apparent under-fitting that occurred when using just the one tree will hopefully be avoided. An accuracy of 100% is now achieved for the training set, and 96% for the test set suggesting that the issue has largely been resolved. This accuracy is similar to that achieved via LDA, and higher than PLS-DA. This shows that each method, while different, can achieve similar acceptable predictions while also being able to highlight different patterns and trends within the data. The MDA and MDG can be calculated from the forest, and are shown in Table 2.10. As seen the variable with the largest MDA and MDG is petal length, supporting previous results. MDA and MDG for petal width are similar, suggesting that this variable is also important for discrimination.

	Mean Decrease Accuracy	Mean Decrease Gini
Sepal Length	11.5	6.6
Sepal Width	6.1	2.0
Petal Length	34.1	30.7
Petal Width	30.6	26.6

Table 2.10: Mean decrease in accuracy and gini index for the iris variables calculated via a random forest.

Chapter 3

Data

Two data sets are analysed in this thesis and used to provide examples for the methods developed and discussed. These data sets and the corresponding meta data are described here. For each of the data sets there are similar research aims, and this is primarily to discover any trends and patterns within the data. This focuses on finding relations to the categories each observation has, such as the origins. After these trends are found, the compounds which are responsible for these patterns a

3.1 Honey Data

The first set of data comes from honey samples. The geographical origin and floral composition of honey is often studied because of the adulteration than can occur in the market [4], [18], [111], as well as to find markers for different types of honey [8], [112], [113].

3.1.1 ^{1}H NMR data acquisition

Data was collected using a Bruker Avance 500 MHz NMR spectrometer equipped with a TCI cryoprobe. A central frequency of 500.1323531 MHz was used to acquire the spectra, and the Bruker pulse sequence noesyphpr with on-resonance presaturation was used to suppress the intensity of the water signal. A 90° observation pulse length of 9.0 µs and a delay between transients of 3 s were used. 65536 complex data points were acquired with a spectral width of 20.8278 ppm, giving an acquisition time of 3.1457 s. One-dimensional (1D) ^{1}H NMR spectroscopic data were processed using FELIX software (Accelrys, San Diego, CA, USA). A sine-bell shaped window function phase shifted by 90° was applied over all data points before Fourier transformation, phase and baseline correction. The chemical shift of all data was referenced to the TSP resonance at 0 ppm. The area of this resonance was set to unity for all spectra acquired.

3.1.2 ^{1}H NMR data set

^{1}H NMR data were collected for 325 honey samples from across the globe. In addition to different geographical origins, samples originated from different flower types, as seen in Figure 3.1. However the exact floral origin is not known for all observations with many being different honey samples blended together (blends).

In fact, most of the samples are polyfloral, meaning that many flowers are within range of the hives and no singular floral type accounts for the required percentage of the nectar from a single floral origin [17]. For the monofloral honeys many of the types are limited to just one or two geographical origins, with the most different origins for a floral type found being just three for Acacia.

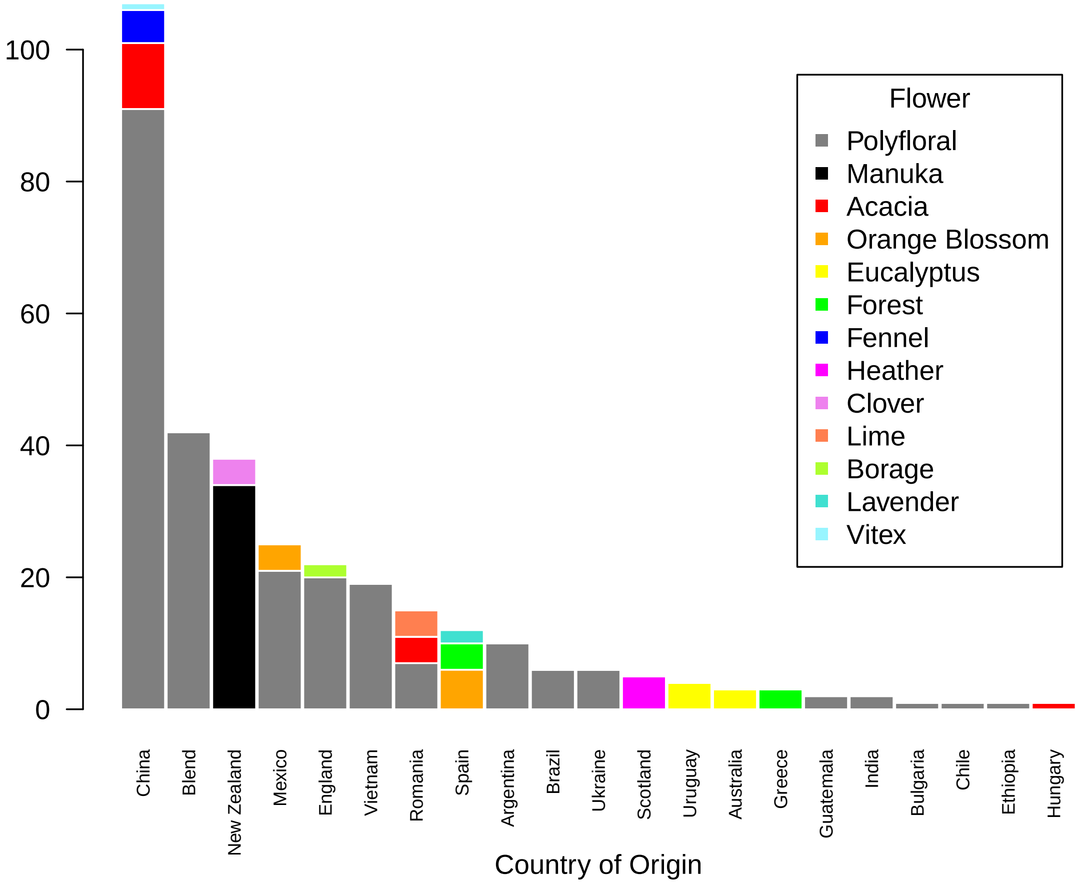


Figure 3.1: Honey samples by flower and country ordered by the number of observations in the data set.

3.2 Coffee Data

Data from six different coffee origins were collected using three techniques, ^{1}H nuclear magnetic resonance (^{1}H NMR), liquid chromatography-mass spectrometry (LC-MS) and gas chromatography-mass spectrometry (GC-MS). The Rwandan coffee was not measured via GC-MS.

Flavour profiles for coffees from five different countries were provided by expert tasters (Rwandan coffee was also not taste tested). The profiles were summarised by scores for six taste variables; citrus, floral, berry, chocolate, nuts and spice. These scores range from 1 to 5 with higher values representing stronger flavours, displayed in Figure 3.2. It can be seen that Kenyan observations have a strong citrus flavour, with low scores for all other flavours. Ethiopian coffee is characterised by floral and berry flavours, whilst Brazilian coffee has strong chocolate and nut flavours. Javan coffee has just slightly lower scores for chocolate and nut flavours but with added spice and Colombian coffee also has a mix of flavours, mostly floral, berry and chocolate.

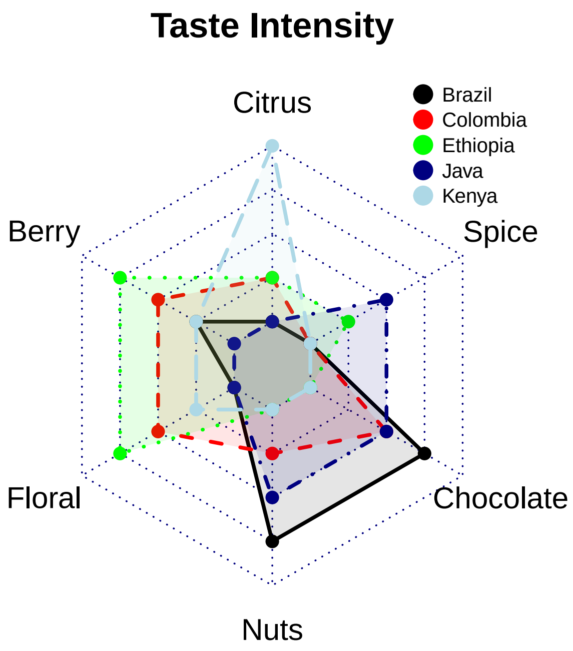


Figure 3.2: Radar plot showing the intensities of the flavours assigned by expert tasters for each of the coffee origins. Flavour strength is measured from 1 to 5, with 5 being the strongest. In the plot a score of 1 is indicated by a point on the inner hexagon, and a score of 5 by a point on the outer hexagon.

3.2.1 Coffee sample preparation

In order to prepare the coffee beans 10g of ground coffee beans were added to 250ml of just off the boil water in a cup. It was brewed for 4 minutes, and then using 2 spoons the ‘crust’ was removed. The coffee was then filtered with a syringe and passed through a 0.4 μm filter. This first step was uniform for all three techniques.

3.2.2 ^{1}H NMR coffee preparation

After filtering samples were snap frozen in liquid nitrogen for 30 minutes then lyophilised for a minimum of 72 hours. 75 mg of the lyophilised sample was extracted in 1.5ml of 250mM D_2O phosphate buffer (K_2HPO_4/KH_2PO_4, pH = 7.0) with 1mM sodium azide and 1mM trimethylsilylpropionic acid- d4 sodium salt (TSP), and vortexed for 30min. Samples were then centrifuged at 21,000g for 10min at 20°C. The supernatant was filtered using 13mm PTFE 0.45μm Klarity syringe filters, and 600μl added to 5mm Wilmad NMR tubes.

3.2.3 ^{1}H NMR data acquisition

^{1}H NMR data was acquired on 500 MHZ Bruker spectrometer analysing Hydrogen nuclei. 1D Spectra were acquired at a central frequency of 500.1323546 MHz data into 32768 data points over a spectral width of 7042.25 Hz (equating to 14.08 ppm). A 3.5s relaxation delay was found to be sufficient for the acquisition of quantitative data for all resonances. A sine bell-shaped window function phase shifted by 90°was applied over all data points before Fourier transformation, phase, and baseline correction. All spectra were acquired at 300 K. The chemical shift of all data was referenced to the TSP resonance at 0 ppm.

3.2.4 ^{1}H NMR data set

This data set comprises 3 analytical replicates of each 8 coffee samples from different origins, giving 24 observations in total. The samples originated from Brazil, Colombia, Ethiopia, Rwanda, Kenya and Java with two different processing methods (as described in Chapter 1) used for both Ethiopian and Brazilian coffee. Experiment number 16 failed the quality assurance, and so was repeated as experiment 116. Table 3.1 shows the origin of each observation with details of the method by which the coffee beans were processed.

3.2.5 LC-MS coffee preparation

The coffee samples were filtered again through a 0.2 μm filter and put into LC-MS vials and frozen at -80°C. The coffee was defrosted the day before analysis.

3.2.6 LC-MS data acquisition

LC–HRMS analysis was conducted on samples in a random order, with QC samples run between every 8 experimental samples (a “batch”), resulting in four batches. A reverse phase aqueous chromatography column, Waters Cortecs T3 150x3mm,

Experiment Number	**Replicate Number**	**Passed Quality Assurance?**	**Processing Method**	**Sample Origin**
1	1	yes	Natural	Brazil
2	1	yes	Wet	Rwanda
3	1	yes	Wet	Ethiopia
4	1	yes	Wet	Kenya
5	1	yes	Wet	Java
6	1	yes	Natural	Ethiopia
7	1	yes	Natural	Colombia
8	1	yes	Honey	Brazil
9	2	yes	Natural	Brazil
10	2	yes	Wet	Rwanda
11	2	yes	Wet	Ethiopia
12	2	yes	Wet	Kenya
13	2	yes	Wet	Java
14	2	yes	Natural	Ethiopia
15	2	yes	Natural	Colombia
16	2	no	Honey	Brazil
17	3	yes	Natural	Brazil
18	3	yes	Wet	Rwanda
19	3	yes	Wet	Ethiopia
20	3	yes	Wet	Kenya
21	3	yes	Wet	Java
22	3	yes	Natural	Ethiopia
23	3	yes	Natural	Colombia
24	3	yes	Honey	Brazil
116	2	yes	Honey	Brazil

Table 3.1: Metadata for the ^{1}H NMR analysis of coffee samples.

(Waters, Wilmslow, UK), was used. Mobile phases were 0.1% formic acid in water (mobile phase A, MPA) and 0.1% formic acid in acetonitrile (mobile phase B, MPB). Gradient applied was 100% MPA for 3 minutes before increasing to 60% MPA/ 40% MPB over 17 minutes. It was then increased again to 10% MPA/ 90% MPB over 5 minutes and held for 2 minutes before reverting to 100% MPA and held for 5 minutes. Injection volume was 10 μl, flow rate was 0.4 ml/min and column temperature was 30°C. The MS used was a Thermo Exactive Orbitrap (Thermo Fisher Scientific, MA, USA.) set at 120,000 resolution with the full-width at half-maximum at 200 m/z with an acquisition speed of 2Hz. Data were acquired between 100 and 1000 m/z for 30 minutes per sample. The column was conditioned before sample analysis using six QC injections.

3.2.7 LC-MS data set

LC-MS data were collected for 3 replicates each of the three samples from five different origins, as used in GC-MS, giving 45 observations in total as shown in Table 3.2. Data alignment and peak picking were performed using Progenesis QI (Nonlinear Dynamics, Waters Corporation, Newcastle Upon Tyne, UK), after normalisation.

3.2.8 GC-MS coffee preparation

In order to remove water 5g of sodium sulphate (NaSO4) was added, and GC-MS analysis was performed on the same day the coffee was made.

3.2.9 GC-MS data acquisition

A gas chromatograph/mass selective detector (GC/MSD) (Agilent, Santa Clara, CA, USA, Agilent 7890 with Agilent 7200 Q-TOF) was used. Separation was performed on a BPX5 column (50 m, 0.32 mm i.d., 1 µm film thickness) (SGE). An agilent masshunter workstation was used with the GC/MS system. The flow rate (He) was 2 ml min-1 under splitless mode. The injector temperature was 290°C. The column temperature program was: 40°C held for 5 minutes ramped at 10°C/min to 300°C and held for 3 min. Data were acquired in the electron impact (EI) mode scanning from m/z 35 to 500 at 5 Hz. Headspace samples were equilibrated for 5 minutes at 50°C with shaking. A heated syringe at 70°C was then used to inject a volume of 1000 µL.

3.2.10 GC-MS data set

GC-MS data were collected for 6 or 7 replicates of samples from 5 coffee origins as shown in Table 3.3, giving 33 observations in total. Only one of each type of processed coffee was used here. For Ethiopia the naturally processed coffee was chosen, and for Brazil the ‘honey’ processed coffee was used. As such, the samples are only labelled by their origin and the processing method is ignored.

Experiment Number	**Origin**	**Biological Replicate Number**	**Analytical Replicate Number**	**Batch Number**	**Run Order**
1	Brazil	1	1	1	66
2	Brazil	1	2	6	103
3	Brazil	1	3	8	117
4	Brazil	2	1	2	77
5	Brazil	2	2	4	91
6	Brazil	2	3	9	125
7	Brazil	3	1	3	83
8	Brazil	3	2	5	94
9	Brazil	3	3	7	112
10	Colombia	1	1	1	68
11	Colombia	1	2	6	101
12	Colombia	1	3	8	116
13	Colombia	2	1	2	74
14	Colombia	2	2	4	89
15	Colombia	2	3	9	126
16	Colombia	3	1	3	80
17	Colombia	3	2	5	98
18	Colombia	3	3	7	110
19	Ethiopia	1	1	1	67
20	Ethiopia	1	2	6	105
21	Ethiopia	1	3	8	115
22	Ethiopia	2	1	2	76
23	Ethiopia	2	2	4	88
24	Ethiopia	2	3	9	124
25	Ethiopia	3	1	3	84
26	Ethiopia	3	2	5	97
27	Ethiopia	3	3	7	111
28	Java	1	1	1	69
29	Java	1	2	6	104
30	Java	1	3	8	118
31	Java	2	1	2	73
32	Java	2	2	4	87
33	Java	2	3	9	122
34	Java	3	1	3	81
35	Java	3	2	5	96
36	Java	3	3	7	109
37	Kenya	1	1	1	70
38	Kenya	1	2	6	102
39	Kenya	1	3	8	119
40	Kenya	2	1	2	75
41	Kenya	2	2	4	90
42	Kenya	2	3	9	123
43	Kenya	3	1	3	82
44	Kenya	3	2	5	95
45	Kenya	3	3	7	108
47	QC	0	0	2	71
48	QC	0	0	3	78
49	QC	0	0	4	85
50	QC	0	0	5	92
51	QC	0	0	6	99
52	QC	0	0	7	106
53	QC	0	0	8	113
54	QC	0	0	9	120

Table 3.2: Metadata for the LC-MS analysis of coffee samples.

Experiment Number	**Origin**	**Biological Replicate Number**	**Analytical Replicate Number**
1	Colombia	1	1
2	Colombia	2	1
3	Colombia	3	1
4	Colombia	4	1
5	Colombia	5	1
6	Colombia	5	2
7	Colombia	5	3
8	Brazil	1	1
9	Brazil	1	2
10	Brazil	2	1
11	Brazil	3	1
12	Brazil	4	1
13	Brazil	5	1
14	Ethiopia	1	1
15	Ethiopia	2	1
16	Ethiopia	3	1
17	Ethiopia	3	2
18	Ethiopia	3	3
19	Ethiopia	4	1
20	Ethiopia	5	1
21	Java	1	1
22	Java	1	2
23	Java	2	1
24	Java	3	1
25	Java	4	1
26	Java	5	1
27	Kenya	1	1
28	Kenya	2	1
29	Kenya	3	1
30	Kenya	3	2
31	Kenya	3	3
32	Kenya	4	1
33	Kenya	5	1

Table 3.3: Metadata for the GC-MS analysis for coffee samples.

Chapter 4

^{1}H NMR data analysis

4.1 Analysis of Honey Data

A considerable proportion of observations in the honey data set (107/325) originate from China, most of these being polyfloral. As discussed in the Introduction, there are claims that Chinese honey does not fit the widely accepted standards for honey because of differences with how it is processed. The second highest origin is New Zealand, with many being manuka honey. This is one of the most expensive types of honey on the market selling approximating 4000% the price of regular honey [114]. Due to this mark up extensive research has been carried out on authenticity trials to minimise the risk of adulterated manuka honey [11], [115].

Many of the geographical origins have a low number of observations, limiting the analyses that can be performed. Some flower types within the data set also have very low sample numbers and therefore most of the analysis focuses on origins, geographical and floral, for which a suitable number of observations are available.

Exploratory analysis of the ^{1}H NMR data was carried out using PCA with unscaled data. The loadings show that the difference between neighbouring data points in some parts of the spectra are the greatest source of variance (Figure 4.1(A)). The spacing between each data point is calculated as 0.0006 ppm (5dp), and within these loadings there are a number of pairings which have this difference in ppm. For example two loadings relating to 3.8047 and 3.8053 can both be seen having large negative PC1 values. Closer inspection shows this to be due to shifts between the spectra (Figure 4.1(B)). For example, four of the most important loadings are related to the peak located around 3.805 ppm. These chemical shifts are not limited to the small region shown in 4.1(B) but occur over a broad area of the spectra, as shown in 4.1(C). It is clear that these shifts would cause issues for analysis and therefore need to be dealt with. As the shifts are small, binning or bucketing is appropriate.

Adaptive binning was applied to the data using a third level wavelet transform to provide the smooth reference spectrum. This reduces the number of data points from 32768 to just 371 features and removes the negative effects caused by the unwanted chemical shifts. Each of these features relates to peaks found within the spectrum, as each peak can span multiple data points.

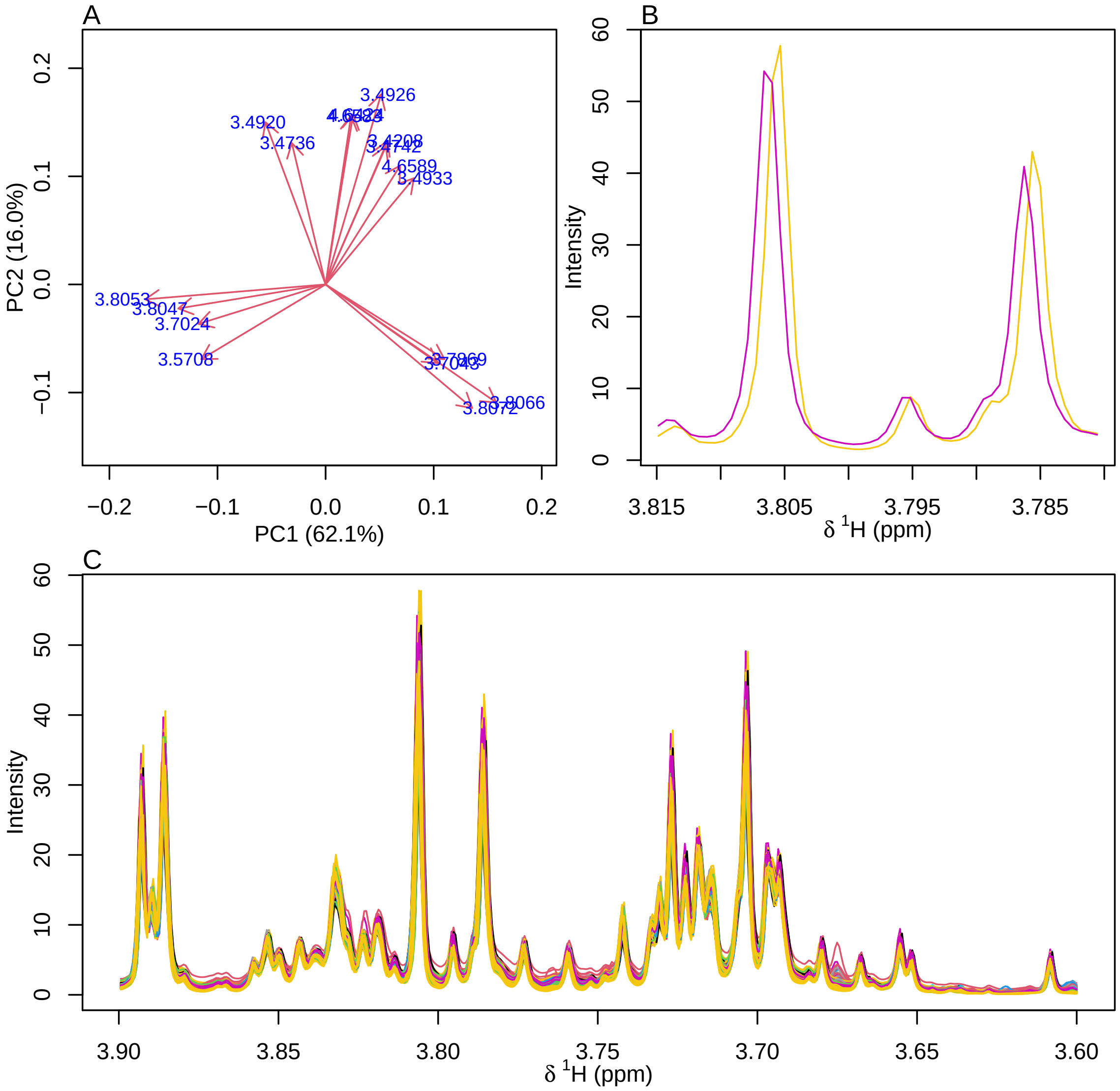


Figure 4.1: (A) The loadings for PC1 and PC2 with the greatest magnitudes for the unscaled honey data. (B) A region of two example honey spectra with the shifts clearly visible. (C) A broader region showing the shifts across all observations.

PCA loadings for PC1 and PC2 with the greatest magnitudes shown in Figure 4.2, and the scores plots for the first four components obtained from the binned data are shown in Figure 4.3. It is clear that binning has resolved the issue caused by the shifts.

In Figure 4.3, the scores are shown coloured by geographical origin in (A) and (C) and floral origin in (B) and (D). Some separation related to floral origin can

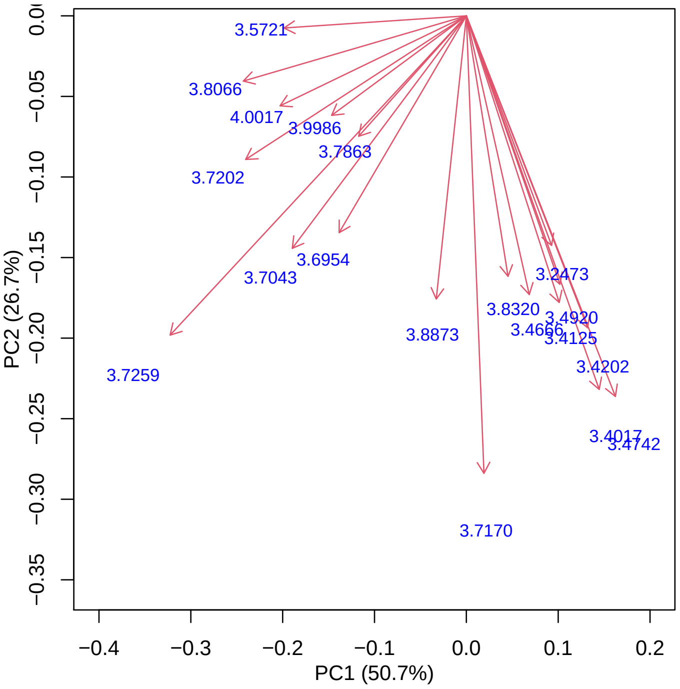


Figure 4.2: The loadings for PC1 and PC2 with the greatest magnitudes for the binned honey data (unscaled).

be seen along PC1, accounting for half of the total variance. Manuka honeys are tightly clustered, whilst polyfloral honeys have predominately positive PC1 scores. While some clustering is present, there does not appear to be any significant separation based on geographical origin in these first four components (95% of the total variance) obtained from unscaled data, which suggests that any differences related to geographical origin would be due to low intensity peaks.

After scaling the data (plots shown in Figure 4.4) an Australian Eucalyptus honey and a Chinese polyfloral honey appear as potential outliers based on their PC2 scores. However, these were not excluded from the analysis as this is not sufficient evidence to suggest they will negatively effect analysis. In PC3 and PC4 there is improved clustering based on the origin, for instance the New Zealand observations within PC4, and the Vietnamese in both PC3 and PC4. As a result of this clustering, scaled data will be used for all analysis.

4.1.1 Differences due to geographical origin

The initial analysis is to investigate if there are any clear and consistent differences between the geographical origins of the honeys. Obviously for the origins with a low number of observations the limited training sets will likely result in low accuracy. Rather than exclude these at this stage, analysis is conducted including them in order to see what these classify as and what honeys they appear to be similar to. A

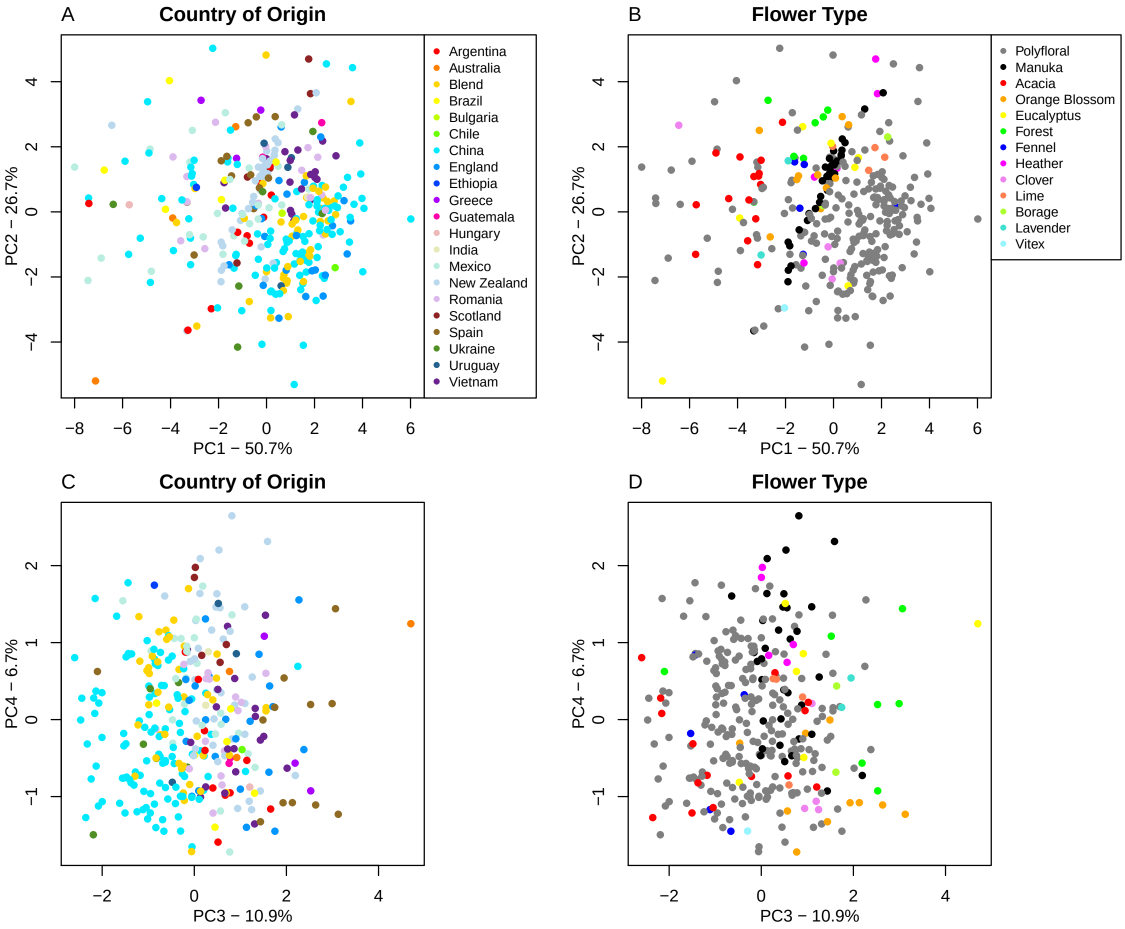


Figure 4.3: PCA scores for the first four components obtained from unscaled data. (A) shows the first two components, coloured by geographical origin, and in (B) the same data is coloured by the flower type. (C) and (D) show the third and forth components, coloured by geographical and floral origin respectively.

training data set consisting of 75% of the observations is formed, taking an equal split of each of the origins where possible. PLS-DA is then used using the origin as a response, and Figure 4.5 shows the resulting X-scores and Y-loadings. Clustering of honey from three geographical origins (China, New Zealand, and Vietnam) can be seen within these first two latent variables (LV).

The variables which account for the separation within these first two components are inspected. Along LV1 the greatest loading, shown in Appendix A.2 Figure A.1 is potentially part of a doublet around 4.65 ppm and for LV2 there are two features, one at 0.97 ppm which has a positive loading and a second at 2.87 ppm with a negative loading. The two features for LV2 both have very small loadings, whereas the doublet is one of the greater intensity sets of peaks.

Based on the root mean squared error of prediction (RMSEP), eight PLS components are needed to suitably model the data, and using these components LDA

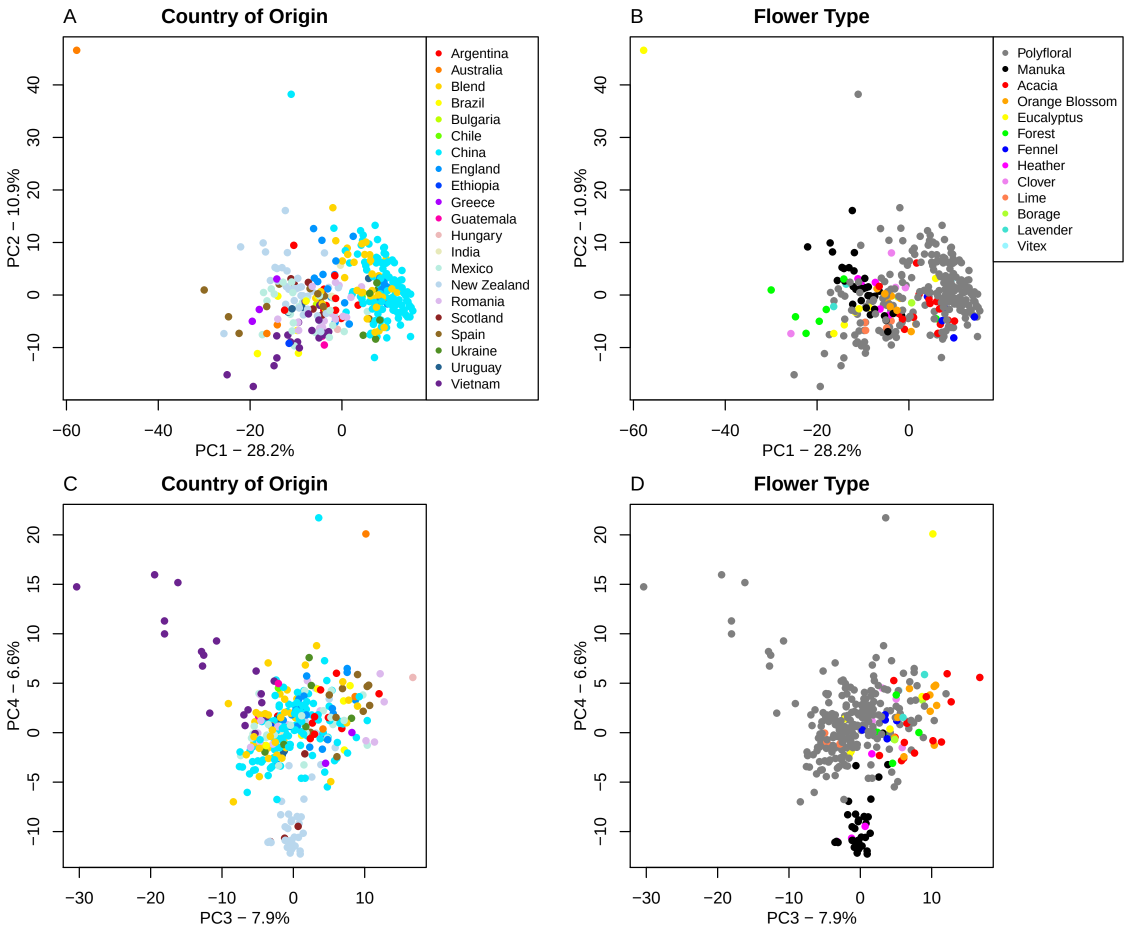


Figure 4.4: PCA scores for the first four components from scaled data. (A) shows the first two components, coloured by geographical origin, and in (B) the same data is coloured by the flower type. (C) and (D) show the third and forth components coloured by geographical and floral origin respectively.

was performed. Whereas this large number of components could lead to over-fitting the training set due to the complexity of the model it forms, an accuracy of 82% was achieved for the training set, only dropping to 79% for the test set. This small drop in accuracy is not enough to suggest it is being over-fit, and instead a large number of components is needed to form the number of classes. The confusion matrix for the test set is shown in Table 4.1. Unsurprisingly, blended honeys have one of the lowest classification rates, with 50% being misclassifed (4 as Chinese, 1 as Spanish). Of the 26 Chinese observations only 2 are incorrectly classified, and these are both predicted as blended honey. Many of the origins with a low number of observations also have a very low accuracy, such as Brazil and Romania (0%) due to the low number of training observations (Brazil has 5, Romania 12).

Due to the low number of observations from some countries the PLS-DA analysis was repeated with all non-Chinese honeys combined to determine how the more traditional mature honeys differ from Chinese immature honey. Blended honeys

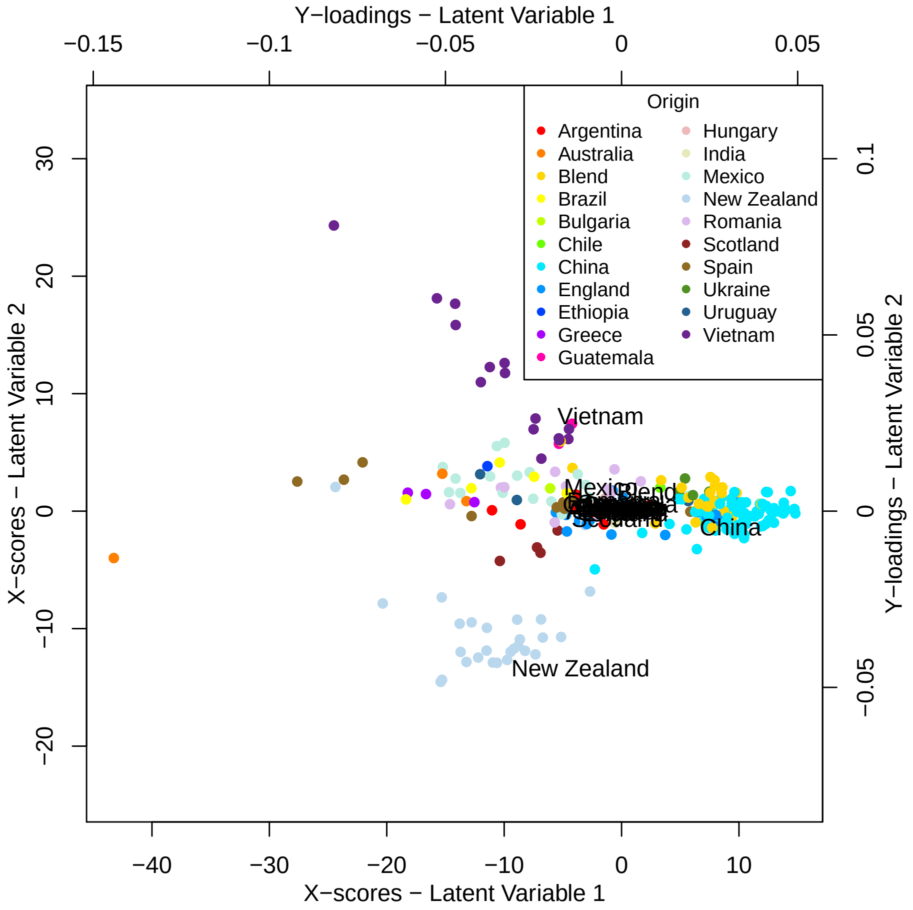


Figure 4.5: X-scores and Y-loadings for the first two PLS-DA components obtained using scaled data with geographical origin as the Y variable.

were omitted from this analysis which achieved 99.5% accuracy for the training set (n = 211), and 97.2% for the test set (n = 72). There are two misclassified observations, one Guatemala honey classified as Chinese and one Chinese honey classified as non-Chinese. VIP scores were calculated for this PLS-DA model and are shown in Figure 4.6A. The four variables with the greatest VIP scores are all related to variables around 4.18 ppm (Figure 4.6B). Although binned data was used in the analysis, the full spectrum is shown for ease of interpretation.

Analysis is now performed to see how polyfloral honeys differ between countries. Due to the nature of polyfloral honeys, being a mixture of multiple unknown floral origins, any floral indicators will not provide consistent differences, and instead theoretically any consistent change found will relate to geographical origins. With the 187 polyfloral observations 91 (48%) are from China, with 21, 20 and 19 observations from Mexico, England, and Vietnam respectively. Bulgaria, Chile, Ethiopia, Guatemala, and India all have only one or two observations each, and so these are combined as a 'miscellaneous' group only represented in the test set. Overall classification rates given exclude these miscellaneous honeys as they cannot be correctly classified, but

		Real Origin												
		Argentina	Blend	Brazil	China	England	Mexico	New Zealand	Romania	Scotland	Spain	Ukraine	Uruguay	Vietnam
Predicted Origin	Argentina	2	-	-	-	1	1	-	-	-	-	-	-	-
	Australia	-	-	-	-	-	-	-	-	-	-	-	-	-
	Blend	-	5	-	2	-	-	-	-	-	-	-	-	-
	Brazil	-	-	1	-	-	-	-	-	-	-	-	-	-
	Bulgaria	-	-	-	-	-	-	-	-	-	-	-	-	-
	Chile	-	-	-	-	-	-	-	-	-	-	-	-	-
	China	-	4	-	24	-	-	-	-	-	-	-	-	-
	England	-	-	-	-	4	1	-	1	-	-	-	-	-
	Ethiopia	-	-	-	-	-	-	-	-	-	-	-	-	-
	Greece	-	-	-	-	-	-	-	-	-	-	-	-	-
	Guatemala	-	-	-	-	-	-	-	-	-	-	-	-	1
	Hungary	-	-	-	-	-	-	-	-	-	-	-	-	-
	India	-	-	-	-	-	-	-	-	-	-	-	-	-
	Mexico	-	-	-	-	-	4	-	-	-	-	-	-	-
	New Zealand	-	-	-	-	-	-	9	-	-	-	-	-	-
	Romania	-	-	-	-	-	-	-	-	-	-	-	-	-
	Scotland	-	-	-	-	-	-	-	-	1	-	-	-	-
	Spain	-	1	-	-	-	-	-	-	-	3	-	-	-
	Ukraine	-	-	-	-	-	-	-	2	-	-	1	-	-
	Uruguay	-	-	-	-	-	-	-	-	-	-	-	1	1
	Vietnam	-	-	-	-	-	-	-	-	-	-	-	-	2

Table 4.1: Confusion Matrix for the test set, showing the results obtained using the eight-component PLS-DA model. Only origins found within the test set are shown in the columns, and similarly only origins which are responses in the model are included in the rows. For readability 0s have been replaced by '-'.

the class assigned is presented in order so trends which appear with their classification can be analysed. An eight component PLS-DA model was created using scaled data from a training set consisting of 70% of the observations. The X-scores for the first two PLS-DA components in Figure 4.7 clearly show clustering based on geographical origin. The Chinese honeys (red) cluster together along both LV1 and LV2 with more spread along LV1 within other groups. LV2 shows separation of Vietnamese honeys (light pink) and, to some extent, Mexican honeys (yellow) although there is overlap of Mexican honeys and Brazilian honeys (black). This is interesting as both are American. Similarly, honeys from Romania (blue) and Ukraine (magenta), both Eastern European countries, overlap.

The PLS-DA model gave an accuracy of 96.0% on the training data (n = 124),

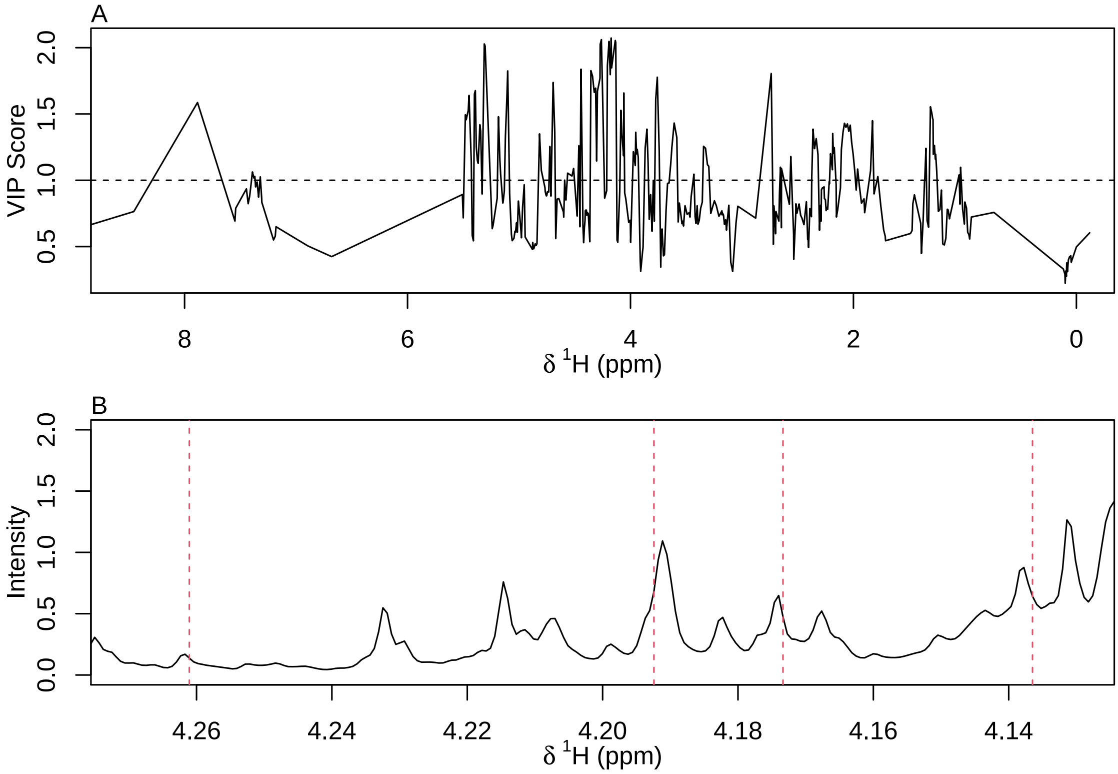


Figure 4.6: (A) VIP scores for the PLS-DA model to discriminate between Chinese and non-Chinese honeys, with a line at VIP = 1. (B) The features with the greatest VIP scores highlighted by dashed lines.

dropping to 87.5% for the test set (n = 56) (excluding the miscellaneous honeys). Whereas eight components is usually considered excessive and can lead to overfitting the accuracy achieved by the test set suggests that this is not the case. These results are shown in Tables 4.2 and 4.3. A total of five honey observations were incorrectly classified in the training set. Two English honeys were classified as Argentinean, two Mexican honeys as Argentinean and Brazilian, and a Romanian honey was classified as Ukrainian. With the exception of the English honeys, there is a geographical link between the real and predicted classes. Within the test data there is again confusion between English and Argentinean honey, with one from each group classified as the other. Another English honey is classified as Brazilian. Again a Romanian honey is predicted as being Ukrainian, as is one Mexican honey.

Of the miscellaneous observations in the test set, Bulgarian honey is predicted as Romanian, continuing the trend of similarity between Eastern European honeys. The Chilean is classified as English, likely based on similarities with honey from Argentina. There is no noticeable trend for the remaining three honeys with an Ethiopian honey classified as Brazilian, a Guatemalan honey as Vietnamese and the two Indian honeys as English.

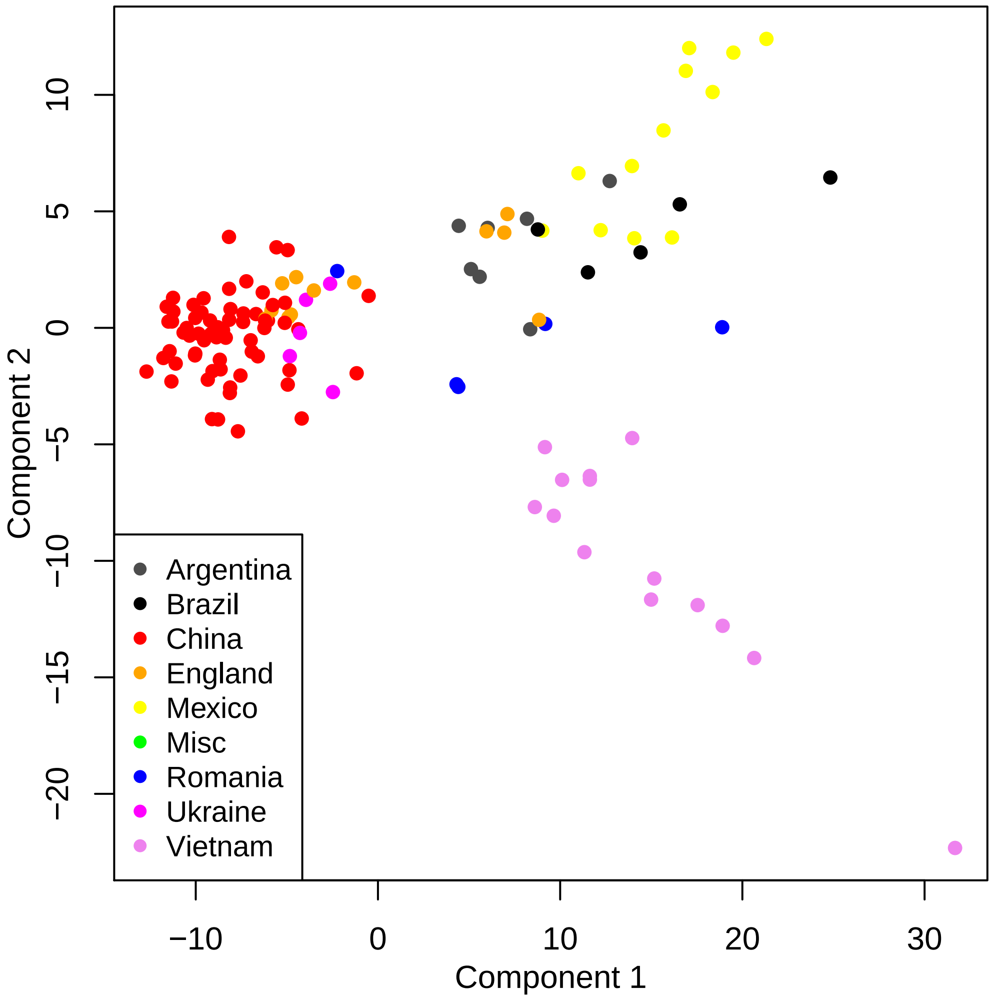


Figure 4.7: X-scores for the first two PLS-DA components for the polyfloral honey analysis, coloured by origin. Clear clustering with geographical origin can be seen.

		Real Origin								
		Argentina	Brazil	China	England	Mexico	Misc	Romania	Ukraine	Vietnam
Predicted Origin	Argentina	7	-	-	2	1	-	-	-	-
	Brazil	-	5	-	-	1	-	-	-	-
	China	-	-	64	-	-	-	-	-	-
	England	-	-	-	10	-	-	-	-	-
	Mexico	-	-	-	-	10	-	-	-	-
	Romania	-	-	-	-	-	-	4	-	-
	Ukraine	-	-	-	-	-	-	1	5	-
	Vietnam	-	-	-	-	-	-	-	-	14

Table 4.2: Confusion Matrix for the polyfloral data, showing the origin classification results for the training using the eight-component PLS-DA model. For readability 0s have been replaced by '-'.

		Real Origin								
		Argentina	Brazil	China	England	Mexico	Misc	Romania	Ukraine	Vietnam
Predicted Origin	Argentina	2	-	-	2	-	2	-	-	-
	Brazil	-	1	-	1	-	1	-	-	-
	China	-	-	27	1	-	1	-	-	-
	England	1	-	-	4	-	1	-	-	-
	Mexico	-	-	-	-	8	-	-	-	-
	Romania	-	-	-	-	-	1	1	-	-
	Ukraine	-	-	-	-	1	-	1	1	-
	Vietnam	-	-	-	-	-	2	-	-	5

Table 4.3: Confusion Matrix for the polyfloral data, showing the origin results for the test set, using the eight-component PLS-DA model. For readability 0s have been replaced by '-'.

Clearly there are differences between geographical origins, however these are predominately low intensity peaks.

VIP scores are calculated from the eight component model, and for many of the features with the greatest scores these are discriminators for Vietnamese honeys, as to be expected from the PLS score plot (Figure 4.7). The feature which provides the greatest split for this it at 2.42 ppm, which is also the feature with the largest score. Chinese and English honeys are also the drivers for separation within some of the top features. The peaks located at 4.182 and 4.192 ppm, which are also discriminators in the Chinese honey vs rest of the world honey modelling have high VIP scores. These can potentially be used as markers for origin, and compound identification is attempted in section 4.1.3.

4.1.2 Differences between floral origins

Bees can travel up to 5 miles collecting pollen, and the floral origin will depend on the surrounding flowers. Honey from a single floral origin, or monofloral honey, is often more expensive and the ability to discriminate between honeys from different flower types could allow for the detection of fraud and adulteration. The identification of markers of various honey types has been a goal in many studies [8], [11]. Many monofloral honeys claim to have benefits over polyflorals, for instance the natural anti-bacterial properties of manuka honey [116], [117].

Here, decision tree methods are used to classify the floral origin of the honey observations.

Figure 4.8 shows a decision tree formed from a training set consisting of 75% of the observations from each flower type. Despite there being 13 different flower types in total, the tree only has six different types as its leaf nodes. This tree achieves 83.7% accuracy on the training data (n = 227) with 73.2% of the monofloral honeys correctly classified (n = 55). Flower types such as 'Orange Blossom' which has five of the eight test observations classified as Acacia cannot be classified correctly by this tree due to the lack of leaf nodes. This issue can be addressed by specifying a lower threshold for the minimum number of observations required in a leaf node. However, although this increased the accuracy on the training set, the test set accuracy decreased as a result of overfitting, see Figure 4.9.

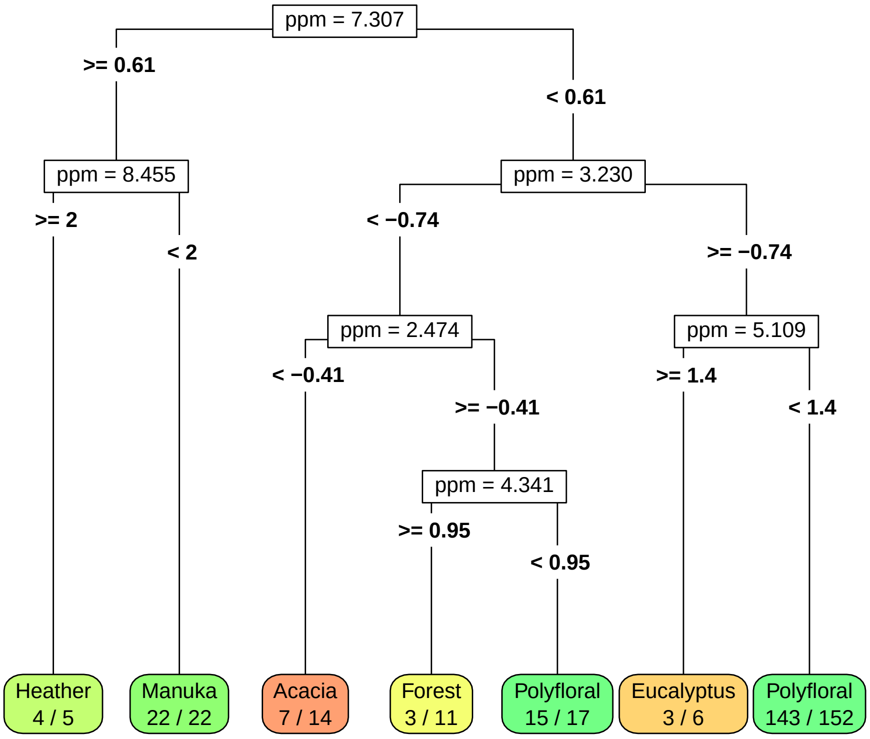


Figure 4.8: Decision tree for the classification of honey floral types, with results shown for the training data.

With the original threshold (5) the accuracy on the test set is 81.6% (n = 98). The confusion matrix is shown in Table 4.4. After just two branches, the manuka honey is classified with 100% accuracy (n = 11), however there are three honeys (one Heather, two polyfloral) which are wrongly classified as manuka.

Random forests were used to determine the variables that were used in multiple decision trees, with the importance of each variable measured in terms of the mean decrease in accuracy (MDA). 5000 trees are used, each only using 20 variables out of the 371. The MDAs are shown in Figure 4.10. The variable with the highest MDA has ppm 7.307, which is also the variable at the root node in the single tree shown in Figure 4.8. For each of the floral origins with a significant number of observations, the MDAs were inspected to see which features appear as markers. For manuka while

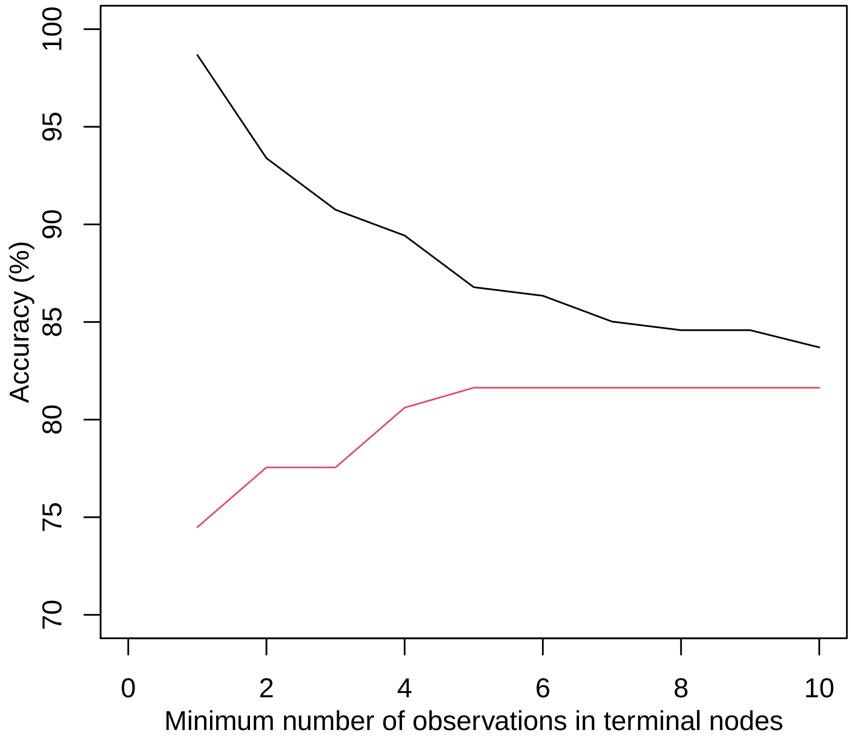


Figure 4.9: Accuracy of the training set (black) and test set (red) for the decision tree when varying the number of observations allowed in terminal nodes. As the number reduces, over-fitting gets worse.

		Real Floral Origin									
		Acacia	Borage	Clover	Eucalyptus	Forest	Heather	Lime	Manuka	Orange Blossom	Polyfloral
Predicted Floral Origin	Acacia	2	-	-	-	-	-	-	-	1	-
	Eucalyptus	-	-	1	-	-	-	-	-	-	1
	Forest	-	-	-	1	3	-	-	-	-	3
	Heather	-	-	-	-	-	1	-	-	-	-
	Manuka	-	-	-	-	-	1	-	11	-	2
	Polyfloral	3	1	2	-	-	-	2	-	-	63

Table 4.4: Confusion Matrix showing the classification results for the test set using the single decision tree model. For readability 0s have been replaced by '-'.

leptosperin is already known as a marker [11], the features at 7.307 and 2.305 ppm also provide separation. Many of the greatest MDAs for heather honey also have a high MDA as manuka too. Acacia and Orange Blossom have the same greatest feature in terms of MDA, however with so few observations this is unreliable.

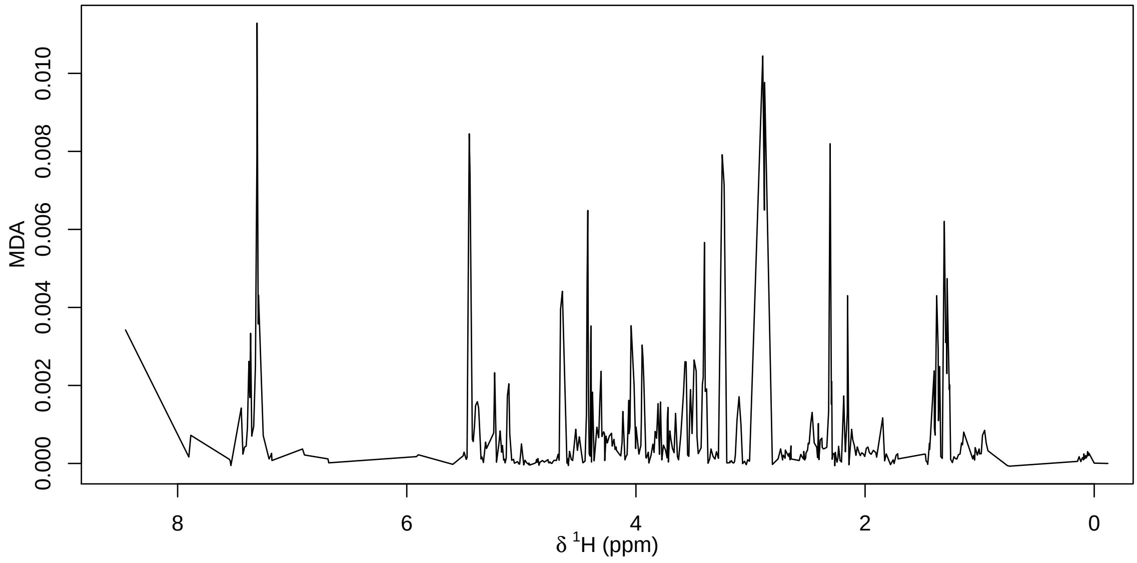


Figure 4.10: Mean decrease in accuracy of origin for each variable calculated from the random forest.

4.1.3 Tentative Compound Identification

Spectral Database for Organic Compounds [118] was used for identifying the corresponding compounds for the ppms of the variables identified in this section to investigate the key compounds. Of the three compounds identified in the PLS-DA analysis to investigate geographical origin (Figure 4.5), the first, located at 4.65 ppm could potentially be Quercitrin, a flavonoid previously found in honey [119], [120]. The second, found at 0.965 ppm is believed to be 2-Methylbutanal [121], and the third feature at 2.87 ppm could not be identified.

Three of the four compounds with the highest VIP score when using PLS-DA to find differences between Chinese and non-Chinese honey have also been tentatively identified. The compound at 4.26 ppm is found to match D-glucopyranose [122], the feature at 4.192 ppm methyl glycolate [123], and finally the feature at 4.135 ppm is possibly sodium gluconate [124]. The fourth feature with ppm 4.17 was not able to be identified with a match found in honey.

The model testing the origin of polyfloral honeys which used eight PLS components provided clear separation for the Vietnamese honeys, clearest at the peak around 2.42 ppm which matches 2-chloroethanol, seen in honeys previously [125]. One of the features that was able to discriminate Chinese and English honeys could potentially be Divinyl Ester however this has previously been found in only a select few honeys [126].

Finally, the key feature used for floral identification at 7.307 ppm could be Thiophene, a volatile compound similar to benzene and known to be in honey [127]. For the remaining features in the tree some of these have been tentatively identified. For the feature at 3.23 ppm this is found to potentially be 2-isobutoxyethyl acetate [128], 2.47 ppm is a match with ethyl pyruvate [129], and finally for the feature which separates Eucalyptus and polyfloral honeys at 5.11 ppm 3-ethyl-2-pentene is tentatively considered a match [130]. The feature found in the random forest for manuka, paired with the feature at 7.307 ppm, is found to possibly be cyclopentene [121].

4.1.4 Discussion

The data here has slight shifts over the spectra, and due to the low magnitude of these shifts adaptive binning, using a third level non-decimated wavelet transform on a reference spectrum, was able to calculate suitable bins to be able to account for the shifts. This provided a suitable example of when this method can work successfully on real data, and allowed for analysis to be performed focusing on differences relating to groups, instead of the variance caused by shifts.

Despite the belief that Chinese honeys are ‘fake’ and immature, analysis here has shown that the differences between Chinese and honey from the rest of the world are minimal. The exploratory analysis on unscaled data showed no separation within the first two components, and there was only slight clustering occurring within the third. Floral origin instead provides the majority of the variance within these components. The scaled exploratory analysis did show improved clustering, suggesting that the differences between Chinese honey and others are within low intensity peaks. These differences have been attributed to several compounds, one of which is D-glucopyranose, a sugar. Within polyfloral honeys Vietnamese honey had clearer separation than Chinese, and this was reflected in the VIP scores made from a PLS-DA model. Several markers were also found here, such as 2-chloroethanol for these Vietnamese honeys.

Due to the number of groups with low observations classification of several origins and floral types was not possible. Many of these added extra layers of complexity to the models, and the resulting accuracies were therefore affected. This was especially clear in the floral analysis, with a decision tree unable to separate these without over-fitting the training data. This highlights the importance of having a sufficient number of observations when performing classification, as well as the need for model validation. Where there is a low number of observations models are unable to find a suitable fit for the data, and will either include the noise in the model which can lead to overfitted models, or it can also lead to underfitting. In order to verify that neither of these occur in the model validation is essential.

4.2 Initial Analysis of Coffee Data

Exploratory analysis of the ^{1}H NMR data from different coffee origins was carried out using PCA. Figure 4.11 shows the score plots for the first two principal components obtained from both unscaled (A) and UV-scaled (B) data. Apart from within the Javan and Colombian observations, there is little evidence of clustering in both PC1 and PC2 according to origin in the unscaled data, with a pattern instead appearing between the observations. The scaled analysis shows no clustering within the first two PCs either. The PC1 loadings for the unscaled data show an interesting pattern, with similar magnitudes for positive and negative loadings of adjacent variables. This produces the mirror effect seen in Figure 4.11C, suggesting an issue with the data. A similar situation was observed with the honey data, where adjacent data points had loadings with opposite signs, and the problem with unwanted chemical shifts was identified. Here too the problem is due to shifts between observations, with an example shown in Figure 4.12.

The variable with the greatest absolute loading is plotted in Figure 4.12 and shows serious shifts affecting the peak. A bin width of 0.05 ppm would be required to ensure this peak falls in the same bin for every observation. Bins of this width would be too large for other areas of the spectra where multiple peaks occur so that fixed-width binning cannot be used. Furthermore, the separation of two observations (shown in blue) from the rest means that adaptive binning would result in two separate bins without a level of smoothing that would also be unsuitable for other areas of the spectra. However, as this peak gives rise to the loading with greatest magnitude, it is possible that this is the worst case and that adaptive binning could still improve the situation overall.

4.2.1 Adaptive Binning

The median spectrum over all observations was used to provide the reference. After smoothing via a 3-level wavelet transform, deemed appropriate for the resolution of the ^{1}H NMR data, a total of 700 non-noise bins were obtained. Figure 4.14 shows the same region as Figure 4.12, however the bin regions have been highlighted in alternating colours. The majority (22 out of 24) of peaks are located within a large

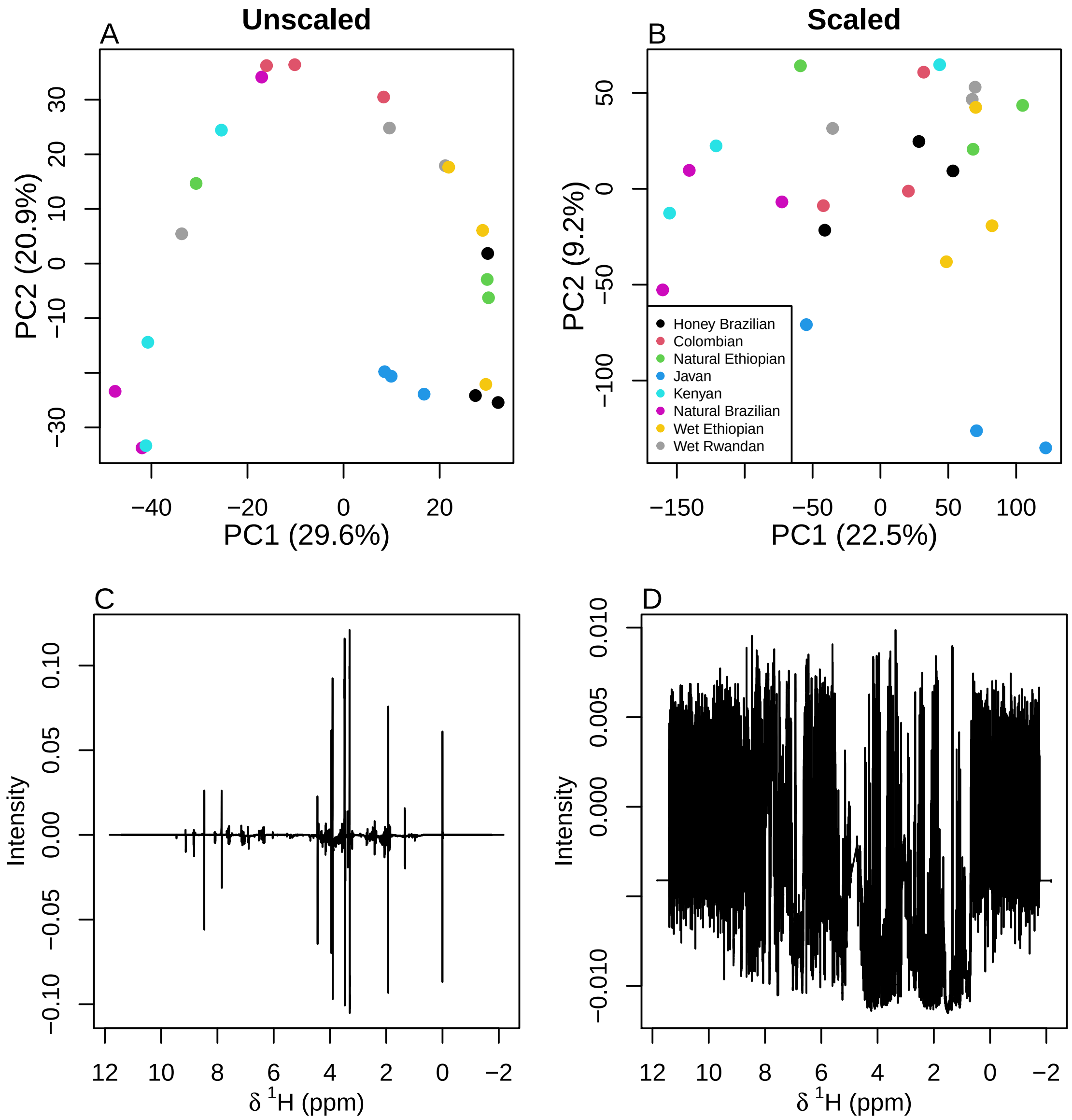


Figure 4.11: PCA scores (A and B) and loadings (C and D) plots for scaled and unscaled data. PC1 and PC2 scores are shown for both, with observations coloured by origin. Only PC1 loadings are shown.

bin with a width of 0.033ppm, but as expected the two Javan observations at the extreme left are assigned to a separate bin. The right hand tails of some of the Natural Brazilian and Kenyan observations have also been cut off.

In order to see whether third level adaptive binning was able to resolve most negative effects caused by the shifts, PCA was again applied to the unscaled binned data. The scores for the first two components are shown in Figure 4.13(A) with the loadings for the first component shown in (C). While some clustering related to origin can be seen, at least for Brazilian and Wet Ethiopian coffees, the loadings show that the issue is still present. The 'U'-shaped distribution of observations follows the shifting shown in Figure 4.12, which suggests that this issue has not been resolved.

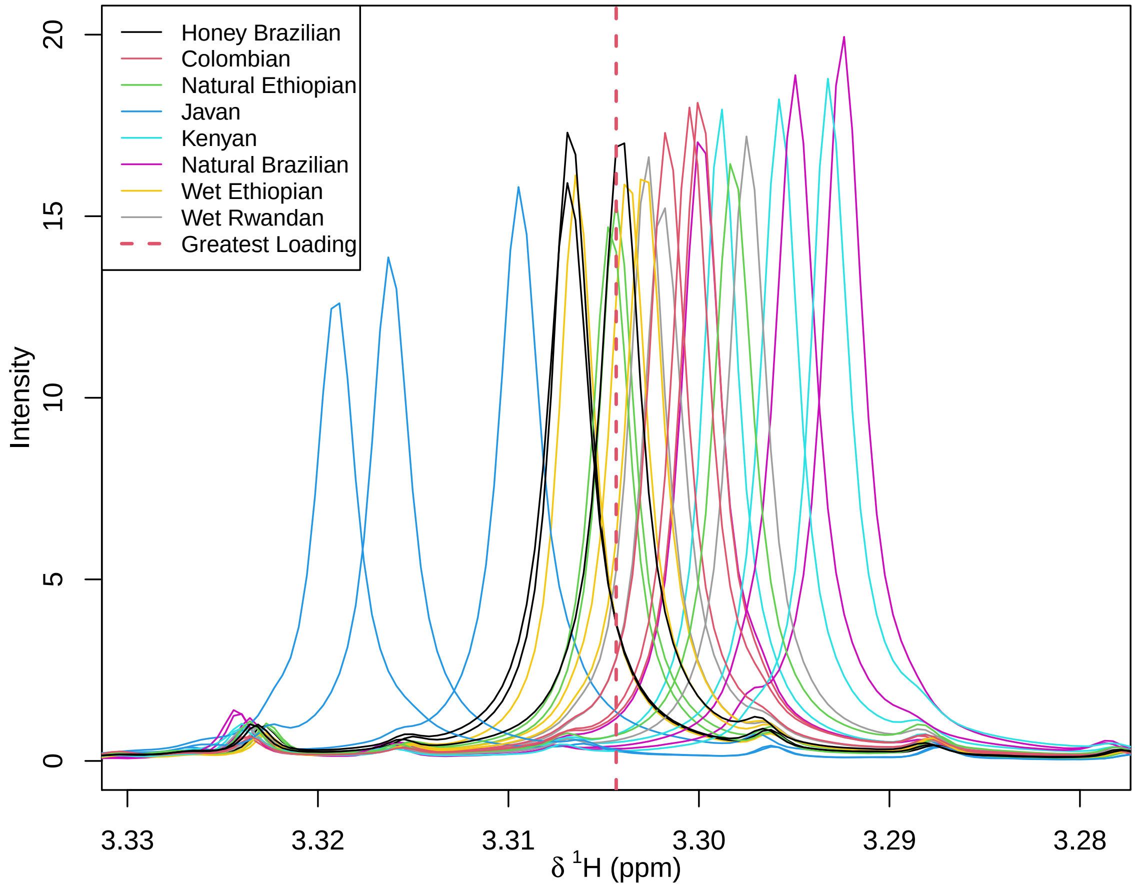


Figure 4.12: Region of the ^{1}H NMR spectra indicated by the loading with greatest magnitude for PC1 in unscaled data. This shows that a large shift between observations for a single peak is responsible for much of the variance in PC1. The dashed line represents the data point with the greatest loading in PC1.

Other wavelet levels were tested and it was found that the lowest level transform which accommodates the shift shown in Figure 4.12 is the sixth, which over-smooths many other areas of the spectra. The PCA scores for the first two components after adaptive binning with six-level smoothing are shown in (B), and the PC1 loadings in (D). While the issue is dampened, this is still an inappropriate approach for resolving the issue due to the over-smoothing of some regions, an example of this shown in Appendix A.2 Figure A.2. The shifting pattern is also apparent when the scores in (B) are projected onto the line $PC2 = -PC1$.

As neither fixed binning nor adaptive binning can successfully deal with the large shifts between observations, alignment methods were also investigated.

4.2.2 Correlation Optimised Warping

The first alignment approach used is correlation optimised warping (COW), described in section 2. The spectra were split into segments of length 82, and the warping

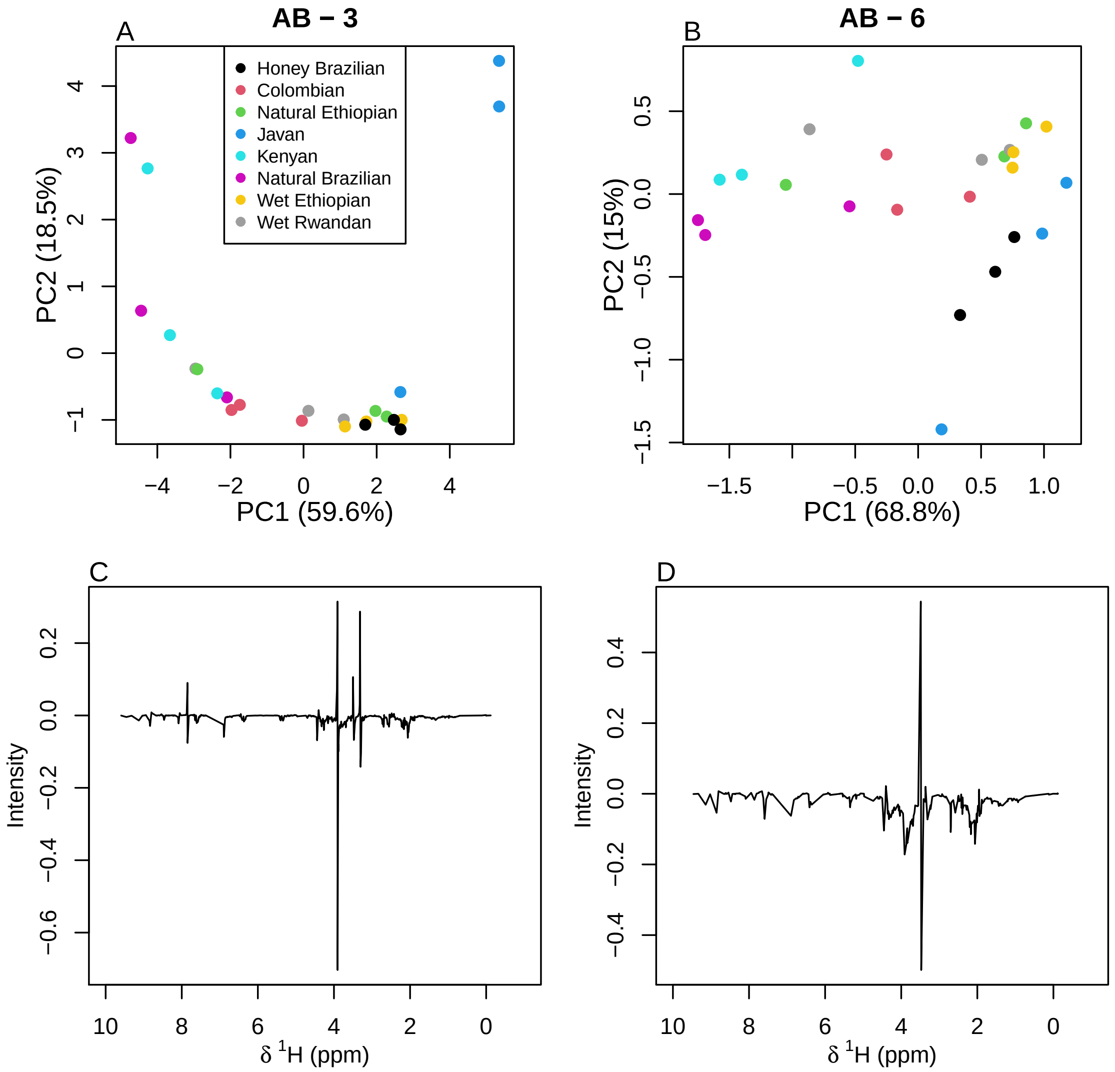


Figure 4.13: PCA Scores and PC1 loadings for the ^{1}H NMR spectra after using adaptive binning at third and sixth level wavelet transforms. (A) shows the scores for third level AB data, and (B) shows the scores at sixth level. (C) and (D) show the PC1 loadings associated with the PCA models for the third and sixth level wavelet approximation data.

slack parameter was set to 82. These parameters were selected via the optimisation algorithm provided by Skov et al. [131] The section of the spectra shown in Figure 4.12 can be seen after alignment in Figure 4.15. Although the shift of the major peak has been accounted for, shifting has occurred for the neighbouring small peaks. The alignment was tested using PLS-DA to classify the observations based on geographical origin. Using the full spectra this only achieved 4% accuracy via LOO-CV. After alignment, adaptive binning (AB) was applied with third level wavelet smoothing, and this increased the accuracy to 33% when using 5 components. Although major shifts appear visually to have been accounted for, further methods need to be applied to extract useful information from this data set.

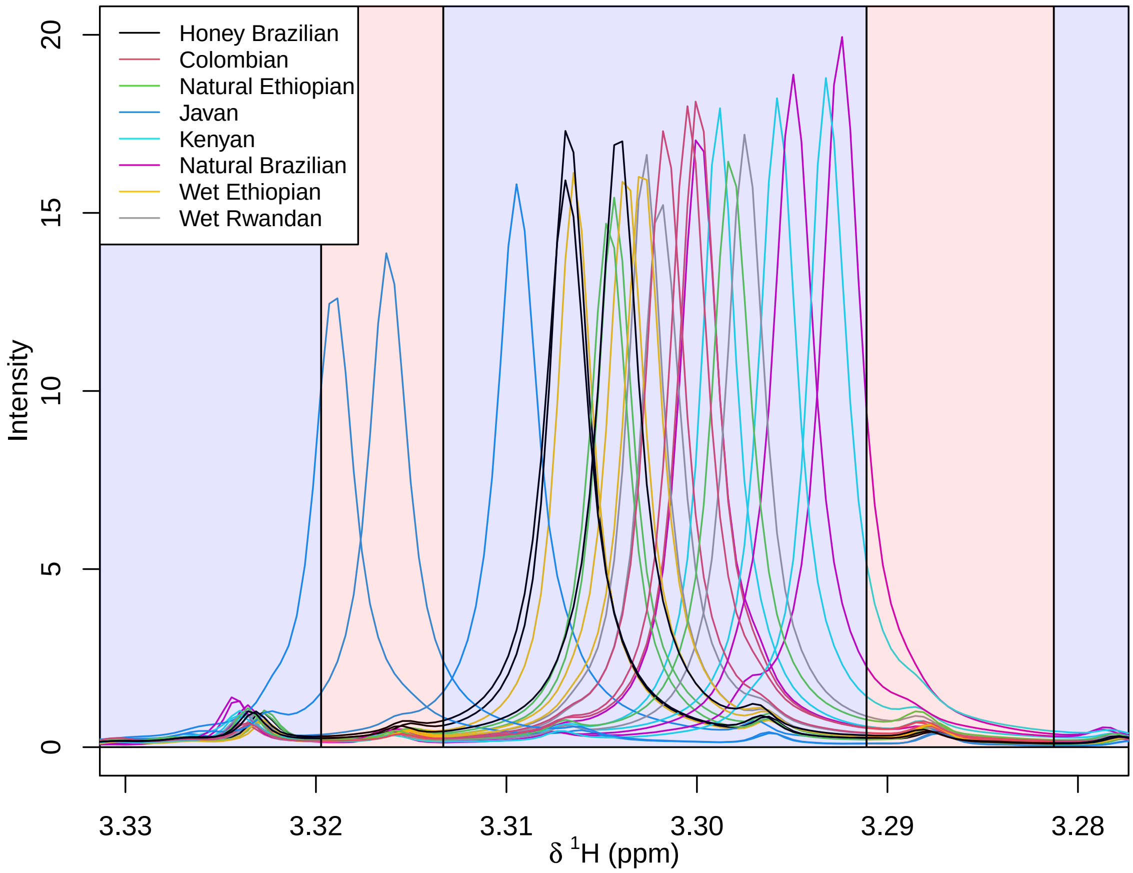


Figure 4.14: The region shown in Figure 4.12, but with the bins obtained by adaptive binning with third level wavelet smoothing overlaid. Alternating colours represent different bins, with black vertical lines at the boundaries.

4.2.3 Speaq 2.0

The second alignment method applied here is Speaq 2.0, as described in Section 2.2.2.5. The baseline threshold was set using the Donoho-Johnstone thresholding technique with the window width set to the recommended 100 data points (∼0.064ppm). One consideration is whether to include nearby peaks in the algorithm, which determines whether small peaks, such as those seen in Figure 4.12 around 3.323 ppm, should be considered in the alignment. As this is recommended, the algorithm was first run with small peaks included.

Using these default parameters, Speaq 2.0 finds 501 peaks, but inspection of the region shown in Figure 4.16 shows that as a result of the significant shifting of the peak around 3.30 ppm the major peak for two Javan observations once again split from the rest of the observations. Running the algorithm again with small nearby peaks ignored (but other parameters the same), the issue is resolved for this region. However, new problems are introduced in other regions. For example, well defined peaks that were previously aligned are sometimes discarded, as shown in Figure 4.17 by the two peaks around 5.40 ppm now only being represented by a single feature.

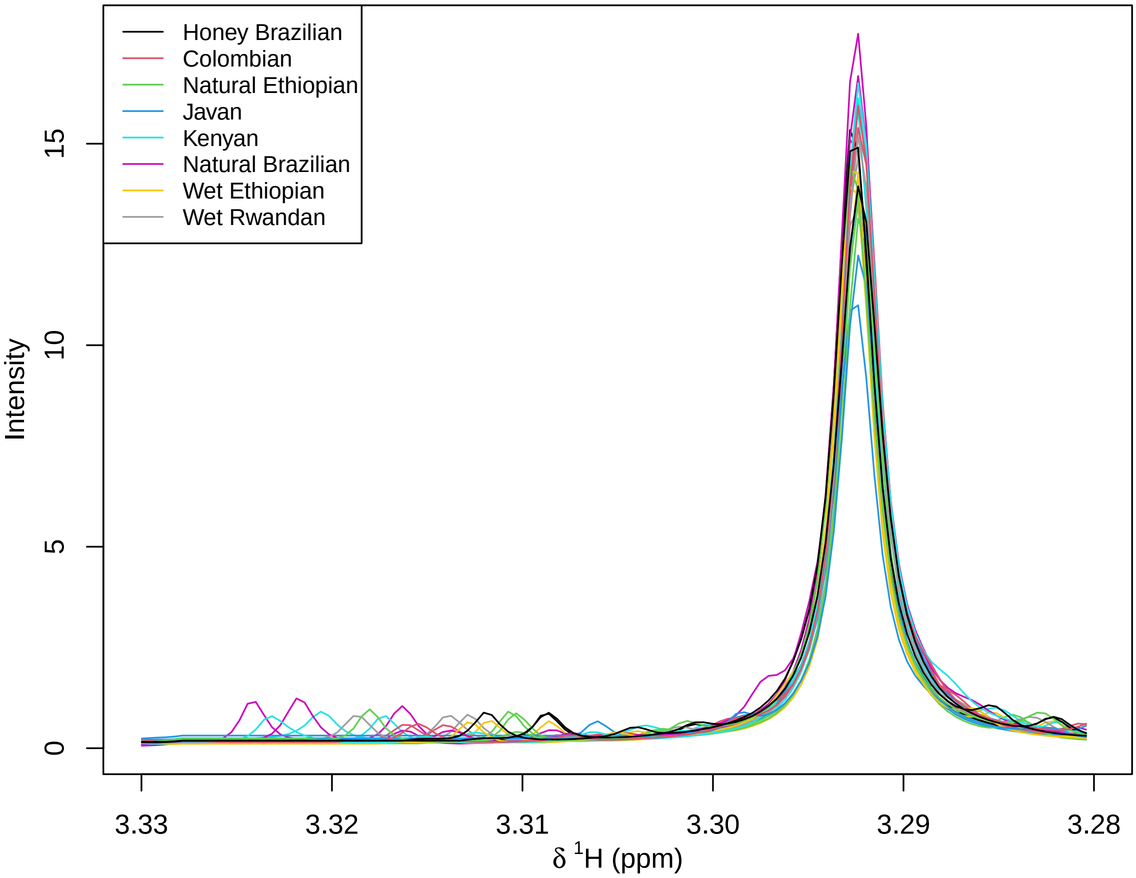


Figure 4.15: Section of the spectra after alignment via COW. This is the same range as shown in Figures 4.12 and 4.14.

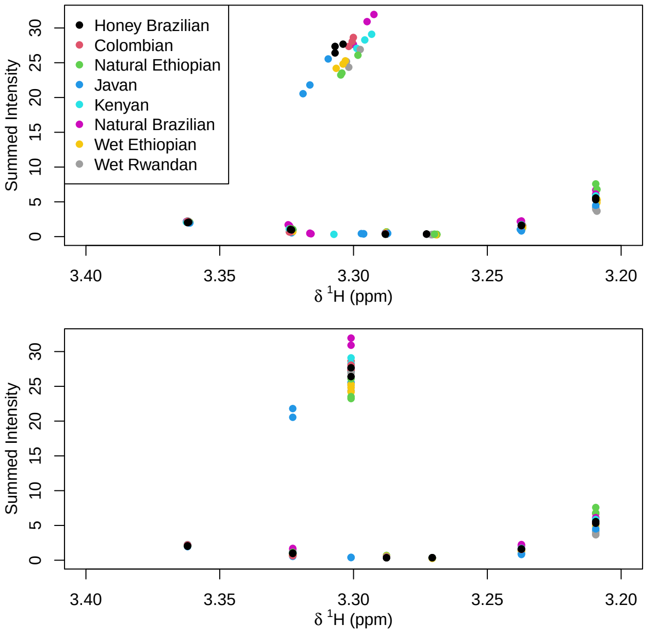


Figure 4.16: The alignment of region shown in Figure 4.12 using Speaq 2.0, taking small neighbouring peaks into account. The upper plot shows the detected peak locations and the lower plot shows how these peaks are clustered. It can be seen that the major peak for two Javan observations cluster separately so that these peaks are not aligned with the rest of the observations.

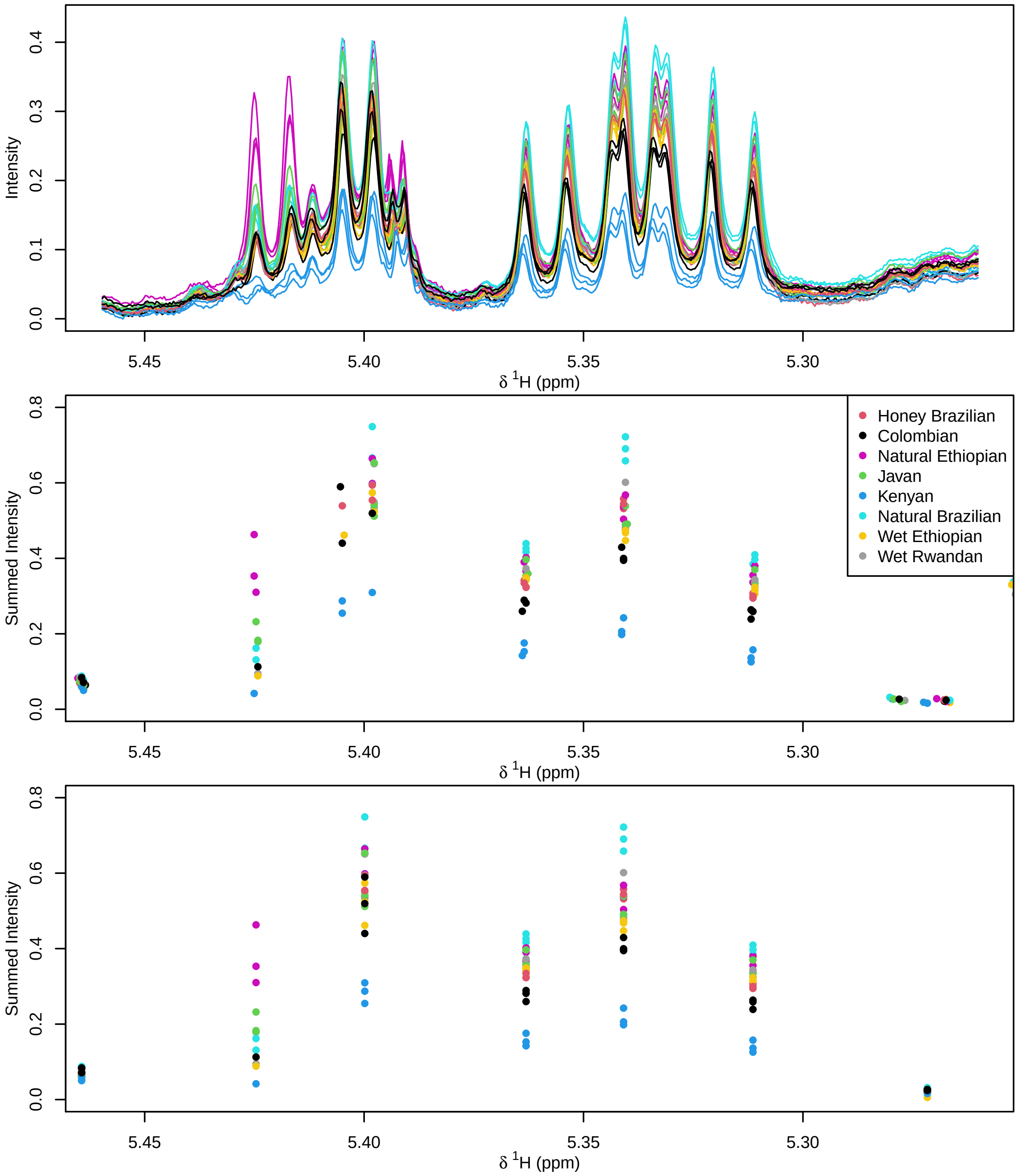


Figure 4.17: A second region of the coffee spectra aligned via Speaq 2.0 with small nearby peaks ignored. The first plot is the unaltered region, the second plot shows them clustered including nearby peaks, and the third plot shows the results when small peaks are ignored. It can be seen that this causes several prominent peaks to be ignored completely.

4.3 Multistage Feature Extraction

Peaks in spectra can have varying unwanted shifts which need resolving by different intensities of correction. An example showing the effect such shifts can have on the Adaptive Binning method (AB), and therefore the resulting features, is shown in Figure 4.18 with the bins obtained in each case represented by the alternating coloured regions. Spectra are coloured according to coffee origin as in all examples shown in this chapter. Two regions of the coffee NMR data are shown with AB performed using a reference spectrum derived from the median at each point. The first has a series of peaks, some of which appear to be doublets, which are well aligned. When AB is performed using a third level Haar wavelet transformation, the doublets are smoothed to give single peaks and then each bin corresponds to a single peak as required. The second region shows one major peak with a significant shift across the observations. The peaks are split between bins when the same third level wavelet transform is applied, making this level inappropriate for the region. The lowest level of WT that does not have peak splitting is the sixth, which Figure 4.18(B) shows over-smooths the first region, combining multiple peaks into single bins.

As there can be very different shifts within the same spectra, a single stage single method to extract meaningful features is not always appropriate. To resolve this issue, a method called Multi-stage Feature Extraction (MSFE) was developed and is described in this chapter. By extracting variables at different levels MSFE minimises the risk of over-correcting the data while attempting to maximise the quality of the extracted data.

The overarching aim of MSFE is to identify the optimal alignment for individual regions by considering shifts at decreasing wavelet scales. This allows regions with either large or small shifts to be corrected at a suitable level. In addition the algorithm makes use of correlations between peaks to further improve feature extraction. It is well known that single compounds can result in multiple peaks within the spectrum, e.g. each hydrogen environment in the compound causes a peak in ^{1}H NMR. Shifts occurring within such peaks may therefore have similar shifts. MSFE makes use of this by identifying the corrections for well-separated peaks and using them to correct the less reliable alignment of overlapping peaks with highly correlated shift patterns.

By considering the quality of peak alignment in different regions of the spectra as well and the amount of correction necessary, regions are categorised as follows:
Class 0: Noise Regions.
Class 1: Clearly defined with little or no unwanted shift occurring.

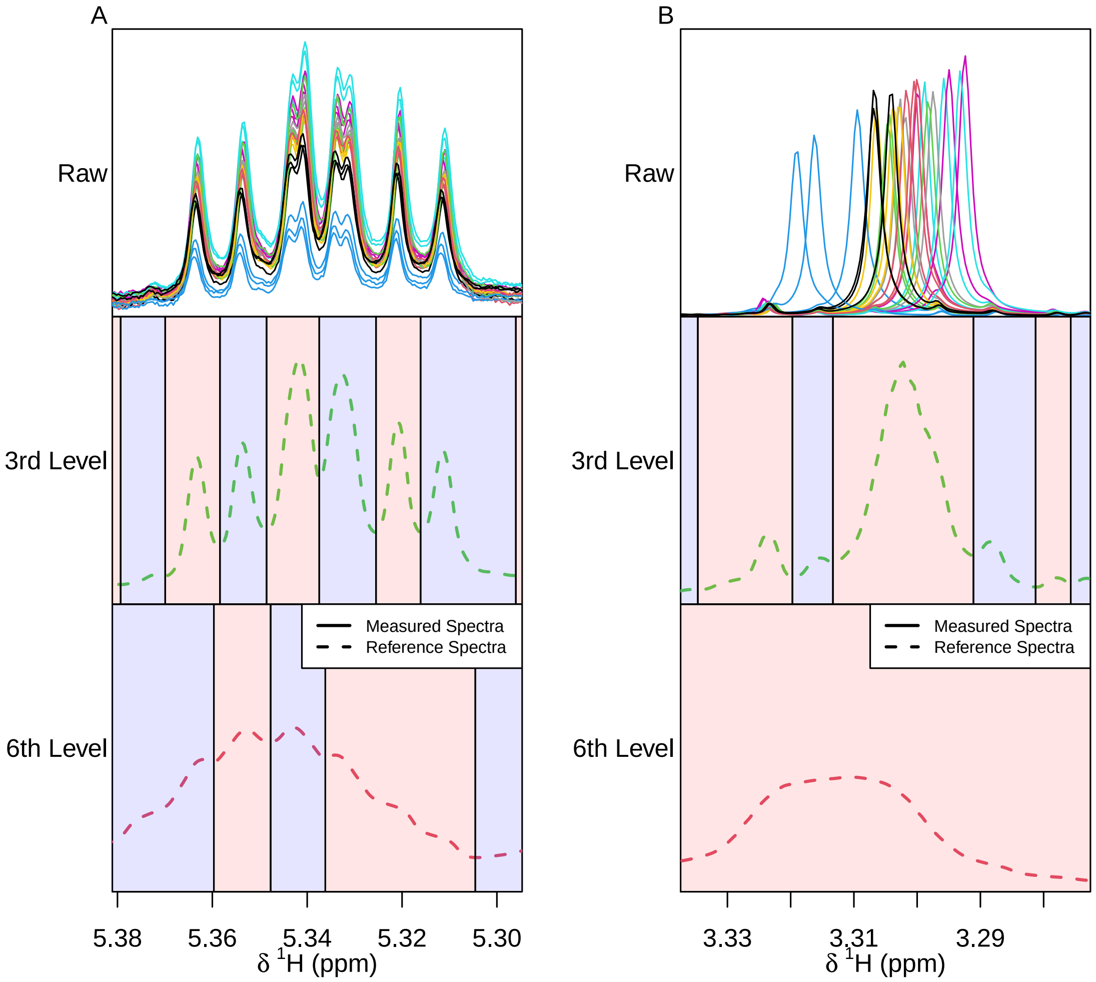


Figure 4.18: NMR spectra of two sets of peaks. (A) shows a series of peaks where the third level approximation allows accurate binning, whereas the sixth level transform over-smooths. (B) shows peaks for which the third level approximation is too low and the sixth level is needed. The magnitudes for the reference spectra (dotted lines) have been enhanced for illustration.

Class 2: Shift that alignment can account for.

Class 3: Many peaks with shifts that cannot be corrected.

In fact, the classes that peaks are assigned can also offer insight for compound identification. For instance knowing that shifting has occurred could imply that it is an acidic compound, known to often cause shift.

The method to classify regions is summarised in the flowchart shown in Figure 4.19 before each step is described in detail in the following sections, and the pseudo-code is shown in Appendix A.2 Algorithm 1 and Algorithm 2.

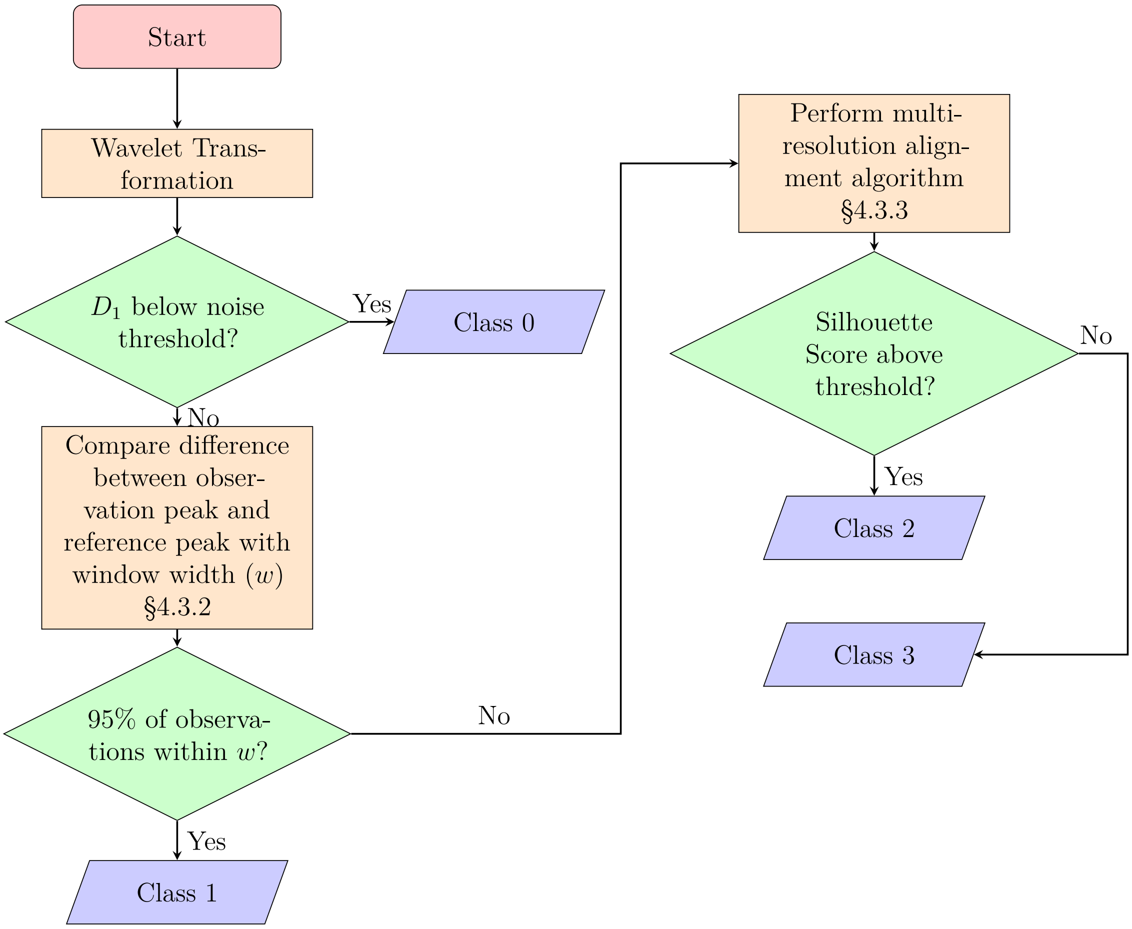

Start
Wavelet Transformation
D_1 below noise threshold?
Yes
Class 0
No
Compare difference between observation peak and reference peak with window width (w) §4.3.2
95% of observations within w?
Yes
Class 1
No
Perform multi-resolution alignment algorithm §4.3.3
Silhouette Score above threshold?
Yes
Class 2
No
Class 3

Figure 4.19: Flowchart for multi-stage feature extraction algorithm.

4.3.1 Denoising

The initial step in the MSFE algorithm is to locate regions containing only noise, and this is achieved within the Adaptive Binning routine. Here, a reference is obtained as the median value at each data point after wavelet smoothing, where the level of the wavelet transform is chosen according to the resolution of the data [27]. Bin ends are then identified from the minima in a non-decimated approximation to this reference spectrum. Using the Donohoe and Johnstone noise threshold on the smallest detail wavelet coefficients, some bins can be identified as containing only noise (Class 0). These noise regions are excluded from further analysis.

4.3.2 Class 1 feature extraction

After denoising, MSFE makes further use of AB as a quick and effective method of extracting features from peaks with little or no shift. Bins identified as comprising a single significant peak for the majority of observations (95%) with their maximum matching within a few data points can be considered as Class 1 regions for which no correction is necessary and the binned intensities taken as features for the region.

For the peaks for each observation X_i in the data matrix $\mathbf{X}$, any maxima greater than 30% of the maximum intensity within a bin are considered as genuine peaks and are denoted by $p_{ij}, j \in \{1, \ldots, N_i\}$, where N_i is the number of peaks in observation X_i. Similarly, the peaks in the reference spectrum are denoted by $\rho_s, s \in \{1, \ldots, N_\rho\}$.

A window width w is chosen and the peak p_{ij} associated with the reference peak ρ_s if $|p_{ij} - \rho_s| \leq w$. The reference peak ρ_s will be paired with at most one peak (the closest) from each observation. The number of peaks paired with each reference peak, i.e. the number of observations with such a peak, is determined for a given w. This window width depends on data resolution but can be varied to determine the optimum value by considering the proportion of observations associated with a reference peak. This can also be calculated from the typical width of a peak at half-height, usually less than 1.5 Hz, which can be converted into data points if the spectral width is known.

This method is applied to every observation, with the window increasing in size from 1 data point up to a predetermined width W, here set to be a quarter of a typical peak width in the data set. If the maximum allowed window width W is too low then there will be almost no peaks selected as Class 1, providing a false sense of lower quality data. If it is too high then the risk of mistakes being caused by incorrect grouping is introduced, a key issue that this approach is trying to avoid. Furthermore, allowing peaks with quite small shifts to be considered as Class 2 variables rather than Class 1 is useful for the correction of other Class 2 peaks as will be explained later.

4.3.2.1 Example peak width calculations

As the choice of W is important in distinguishing between Class 1 and Class 2 peaks, an example of the process with simulated data is given here. In this example there are three observations, each with three peaks, as shown in Figure 4.20. The first peak is perfectly aligned, the second has greater shifts between the three spectra and the third is only slightly shifted between spectra. For simplicity the reference spectra shown is taken as the maximum over the observations at each point. While the first and third sets of peaks each result in a single maximum in the reference spectra, the shifts in the second set of peaks are severe enough to form two maxima. As we do not consider wavelet smoothing in this example (as would be done when using AB), this gives four peaks in the reference spectra.

The resolution of the data here is 0.001ppm and there are four reference peaks to be matched with peaks in the data: $\rho_1 = 0.030, \rho_2 = 0.053, \rho_3 = 0.060$ and $\rho_4 = 0.100$. Table 4.5 shows the number of observations with peaks matching each reference peak

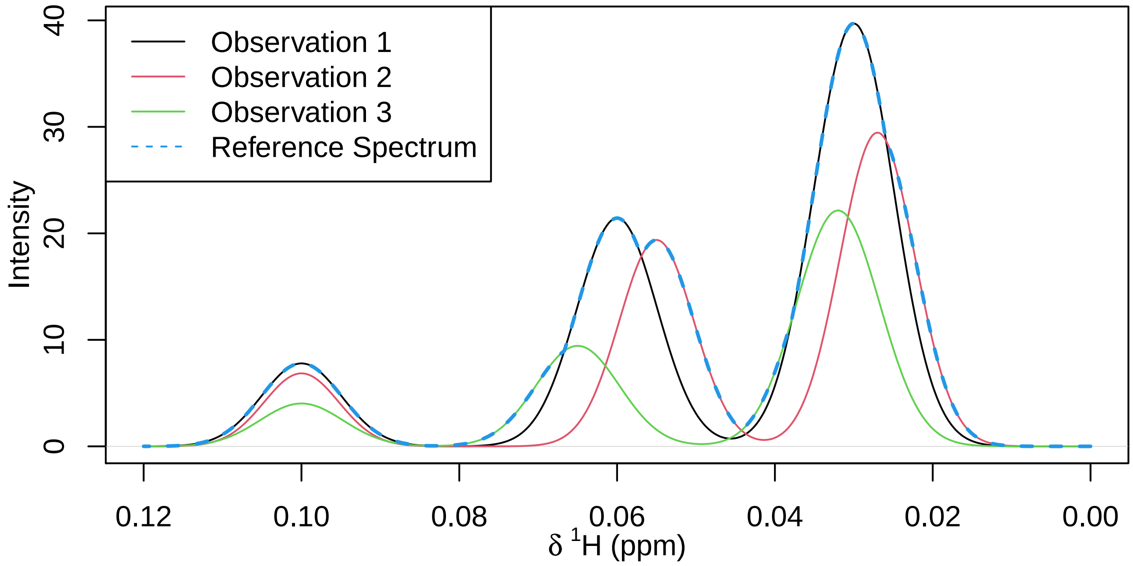


Figure 4.20: Example showing three spectra, each with three peaks, along with a reference spectrum formed from the maximum intensity at each data point. The first peak at 0.10 ppm is well aligned and produces one peak in the reference spectrum. The second peak around 0.06 ppm has greater shifts, leading to two peaks (0.053 and 0.06 ppm) in the reference spectrum. The third peak, located around 0.03 ppm has a slight shift between spectra, but not enough to cause a second local maximum in the reference spectrum.

for four different window widths, up to $w = 5$.

	$w = 1$	$w = 2$	$w = 3$	$w = 4$	$w = 5$
ρ_1	1	2	3	3	3
ρ_2	1	1	1	1	1
ρ_3	1	1	1	1	2
ρ_4	3	3	3	3	3

Table 4.5: The number of peaks assigned to each reference peak for different window widths.

4.3.2.2 Example Class 1 classification

Using the information in Table 4.5, an allowable shift and tolerance on the total percentage can be set. A high percentage of observations matched within a low window width suggests minimal shifts on a well-defined peak, plausibly resulting from the same compound. It is recommended that 95% of observations are present, which in this case means a peak from every spectrum needs to be matched to the reference peak so, with a window width $w = 1$, only ρ_4 is designated as Class 1. Although ρ_1 is matched to a peak in all three spectra with $w = 3$, this window width would be considered too great for a resolution of 0.001 ppm and so this peak would not be classified as Class 1, requiring further correction. This not only avoids the risk of comparing peaks that are not from the same compound, but allows for peaks with slight shifts to potentially be used to correct correlated shift patterns in related Class 2 peaks where the shifts are greater, as discussed later.

4.3.3 Multi-resolution Alignment

With Class 1 peaks extracted as variables for further analysis, the remaining regions require some form of alignment to ensure the same compounds are compared in different spectra. This is shown in Figure 4.21, and each stage discussed in this section. Rather than considering the bins identified by AB, which have potential to be wrong for greater shifts, regions are re-defined as contiguous areas containing peaks that are not considered Class 1 or noise. Then, the number of peaks within each region are counted for each observation, according to the criteria described for determining Class 1 peaks. Ideally, there would be the same number of peaks per observation in each region, but this is rarely the case and region based clustering is necessary to match peaks between spectra.

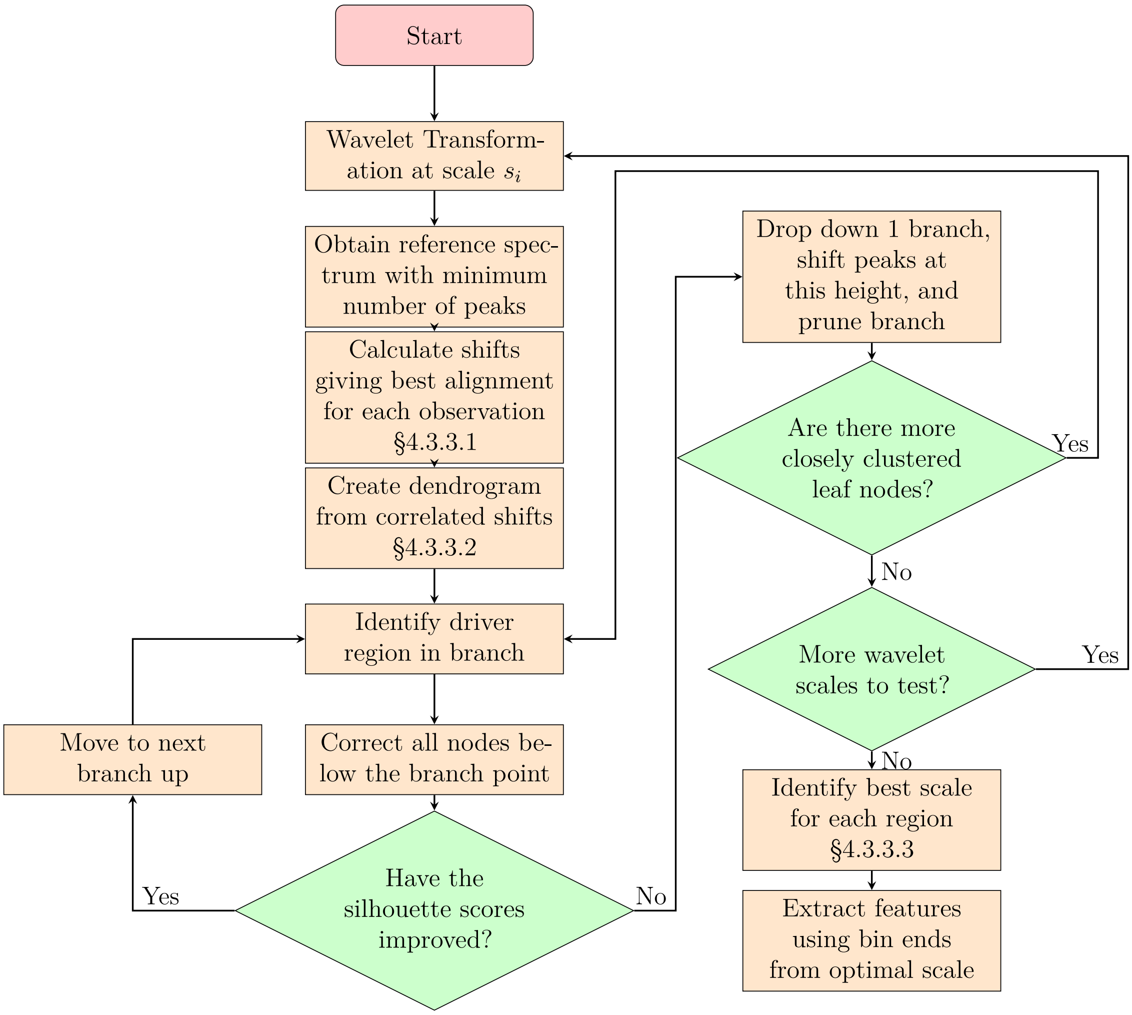

Start
Wavelet Transformation at scale s_i
Obtain reference spectrum with minimum number of peaks
Calculate shifts giving best alignment for each observation §4.3.3.1
Create dendrogram from correlated shifts §4.3.3.2
Identify driver region in branch
Correct all nodes below the branch point
Have the silhouette scores improved?
Yes
No
Move to next branch up
Drop down 1 branch, shift peaks at this height, and prune branch
Are there more closely clustered leaf nodes?
Yes
No
More wavelet scales to test?
Yes
No
Identify best scale for each region §4.3.3.3
Extract features using bin ends from optimal scale

Figure 4.21: Flowchart showing the multi-resolution alignment routine.

4.3.3.1 Region Based Clustering

A dynamic method to group peaks is necessary when the number of peaks present in a region varies between observations. Several simple methods exist, however not without some problems. One such method would be to perform k-means clustering with k as the lowest number of peaks found in any observation. While this might work well for some regions, Figure 4.22 shows an example where the shifts result in peaks being clustered inappropriately.

Two problem clusters have been circled in Figure 4.22, where multiple peaks have been clustered erroneously. It is clear that a more advanced method that can account for such shifts is required, potentially using the distances between peaks within a cluster to generate more reliable clusters. Furthermore, the ordering of the peaks can provide valuable insight for choosing optimal clusters and the k-means algorithm does not take this into account.

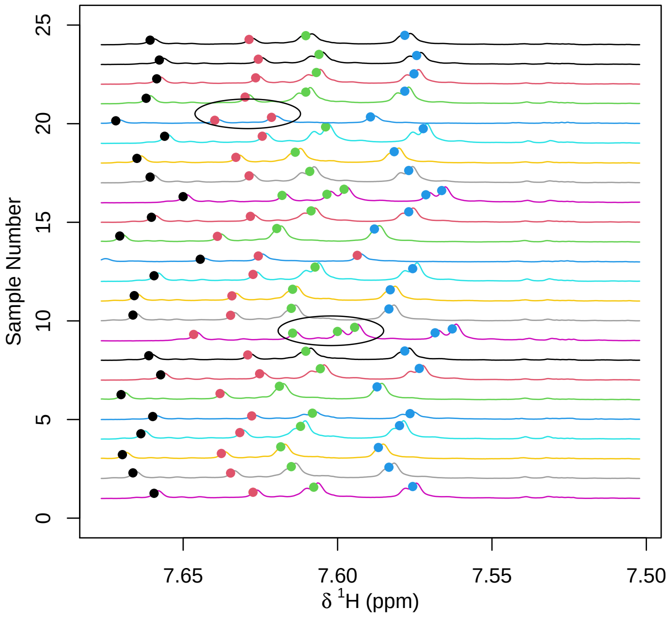


Figure 4.22: Example region with the number of peaks varying between observations. Peaks are highlighted by dots coloured according to the cluster identified by k-means clustering, with $k = 4$. While this works for many observations, visual inspection shows a few observations with erroneous peak clustering (circled).

This problem with k-means clustering can be mitigated by shifting the regions first, but determining the shift for each observation can be difficult. Figure 4.23 shows three example regions that together cover most problems that can arise. Figure 4.23(A) shows the region from Figure 4.22 with 4 peaks in most observations. However, the last two peaks are doublets in two observations, resulting in 6 peaks for these cases. The region shown in Figure 4.23(B) has 6 peaks in the majority of cases but 3 observations have an extra peak located between the second and third peak (compared to other observations). Figure 4.23(C) shows a region with minimal shifting, but one observation has an extra peak on the left. A number of methods to align regions prior to k-means clustering were investigated using these example regions with peaks identified in a third level wavelet approximation.

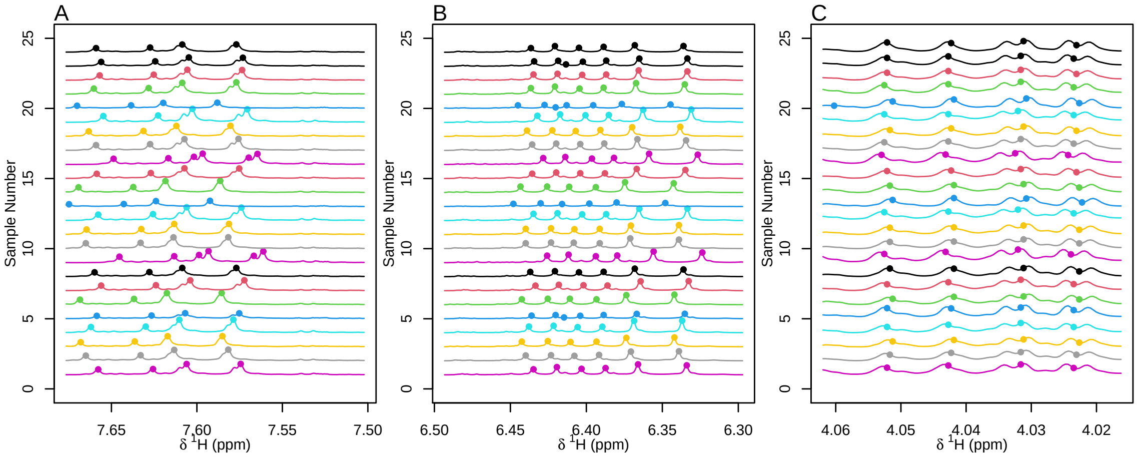


Figure 4.23: Three example regions used to benchmark region alignment. Together these examples cover the majority of issues that can arise in the alignment process. Here peaks identified for k-means clustering are highlighted by dots with observations coloured by origin with $k = 4$ in (A), $k = 6$ in (B) and $k = 4$ in (C).

Clustering to the global maxima within a region. A simple method to temporarily align a region is to shift the observations so that the peak with the maximum intensity in each is aligned. Although quick and easy to perform, multiplets and overlapping peaks can seriously affect the results. Figure 4.24 shows the regions from Figure 4.23(A) and (B) clustered after pre-alignment to the maxima within the region. This method does not work well on region (A), with the clusters obtained by k-means after pre-alignment being no better than those with no alignment. However, the method does work well for the region shown in (B), where distinct peaks are well clustered after pre-alignment, ready for further processing. In region (A) there are some observations where two overlapping peaks have been identified and others where just one of the two peaks passes the criteria. The amount of overlap affects the peak intensities and can therefore affect which peak has the maximum intensity. This is the case in (A), where the highest intensity peak is not always the peak that should be aligned.

Clustering to the geometric mean of the peaks within a region. Rather than pre-alignment based on a single peak in each spectrum, multiple peaks can be used to avoid the issues with overlapping peaks. Alignment to the geometric mean of the identified peak positions within each observation is less sensitive to subtle differences in intensity, but was found to be too sensitive to outlying peaks such

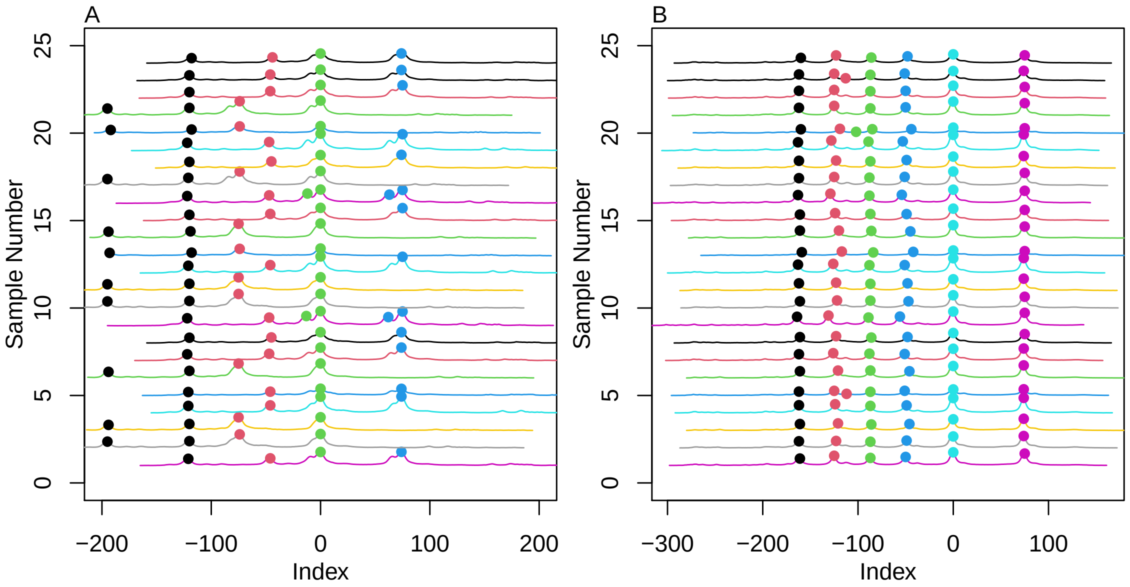


Figure 4.24: Clustering of the two regions shown in Figure 4.23(A) and (B) after pre-alignment to the peak with the greatest intensity. Peaks are identified by dots coloured according to cluster with the spectra shifted to show the alignment. The maximum of each spectrum is represented by 0 on the x-axis.

as that in observation 20 shown in Figure 4.23(C). The clustering obtained after pre-alignment to the geometric mean is shown in Figure 4.25 for the regions (A) and (C) of Figure 4.23. The clustering in (A) shows two observations with a doublet split between 2 clusters and the outlier in region (C) has affected the calculation of the geometric mean and hence caused serious problems using pre-alignment.

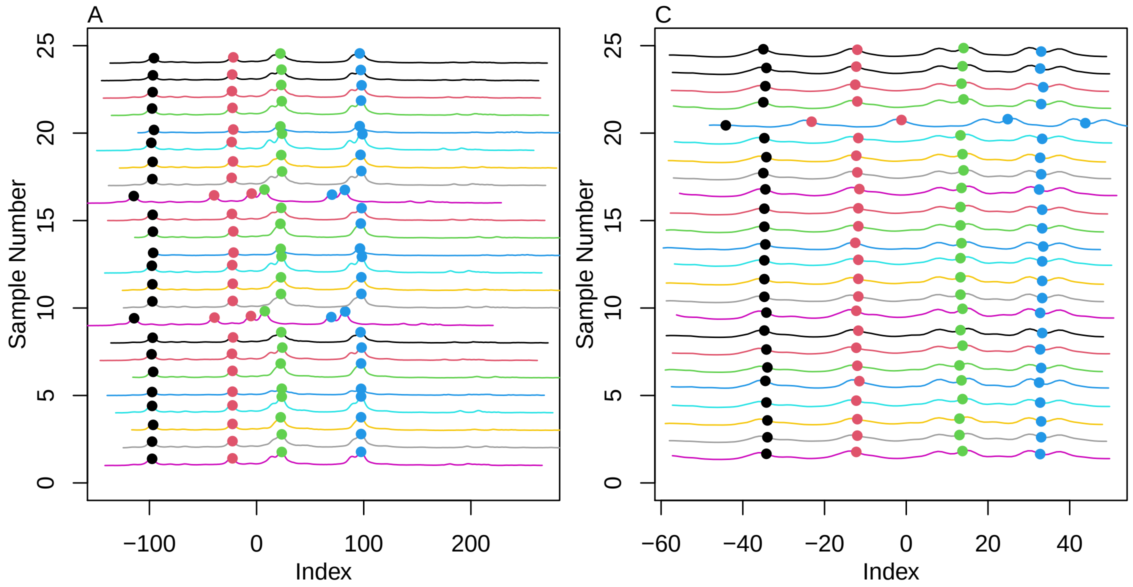


Figure 4.25: Clustering of the two regions shown in Figure 4.23(A) and (C) after pre-alignment to the geometric mean of the identified peak positions within each observation. Peaks are identified by dots coloured according to cluster with the spectra shifted to show the alignment. The geometric mean of the peaks in each spectrum is represented by 0 on the x axis.

As small outlying peaks seriously affected the calculation of the geometric mean, an alternative method using the geometric mean of only the top k peaks in each observation was investigated. The clustering obtained after pre-alignment with this approach is shown in Figure 4.26 for the same two regions shown in Figure 4.23(A) and (C). It can be seen that, although this works well for region (C), region (A) no longer has split peaks but is still poorly clustered. It was found that this method could be very sensitive to multiplets.

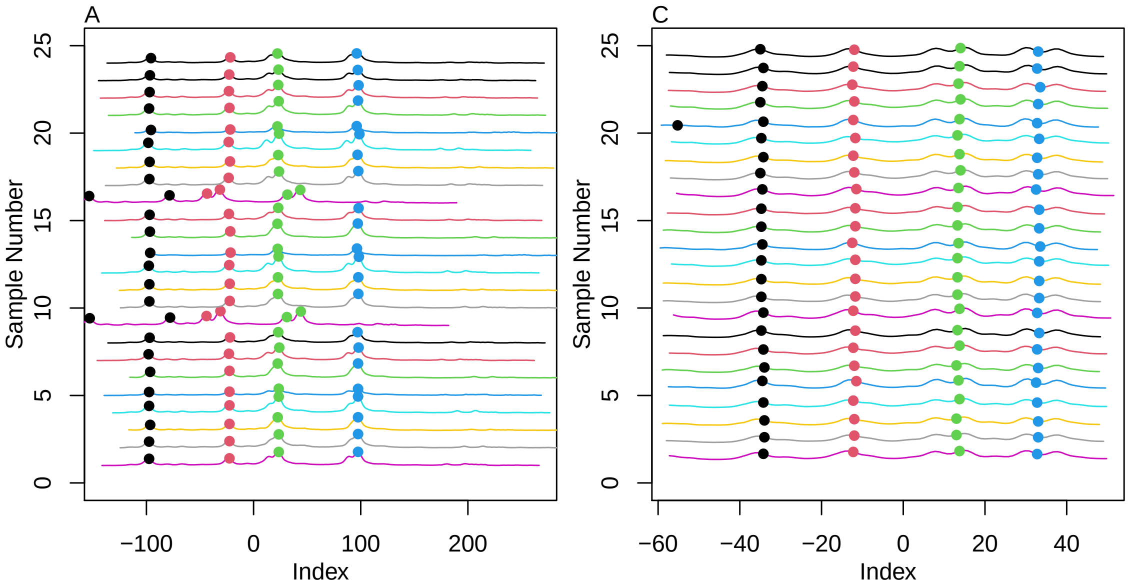


Figure 4.26: Clustering of the two regions shown in Figure 4.23(A) and (C) after pre-alignment to the geometric mean of the k highest peaks, where k is minimum number of peaks in the region for any observation. Peaks are identified by dots coloured according to cluster with the spectra shifted to show the alignment. The geometric mean of the k greatest peaks in each spectrum is at 0 on the x axis.

Correlation-based clustering Instead of analysing the results of a single pre-alignment, correlation-based clustering tests the alignment of every peak in an observation with the maximum peak in a reference spectrum. This reference spectrum is obtained from the subset of observations with the minimum number of peaks (k) for the region. For each of these k peaks, the geometric mean is calculated over this subset and then, for each observation in the subset, the average shift to the k geometric means is determined. After aligning this subset of observations by applying the appropriate shift for each spectrum, the reference spectrum is formed by taking the median of these shifted spectra.

Figure 4.27 illustrates this process for the three observations with the minimal 4 peaks (observations 4, 15 and 20) in the region from 4.23(A). The geometric means for each of the four peaks are shown by the dashed lines. For observation 4 the first two peaks are shifted to the right by 0.00156 ppm, and the second two by 0.00213 ppm. This gives an average shift of 0.00188 ppm from the geometric means, equivalent to 5 data points. Similarly, observation 15 is shifted 12 data points to the right and observation 20 is shifted 16 data points to the left. The lower plot in Figure 4.27 shows the three spectra after these shifts have alignment to the geometric means using the appropriate shift for each observation. The reference spectrum for this region would then be calculated from these three aligned spectra.

The best alignment of every observation to this reference spectrum is determined

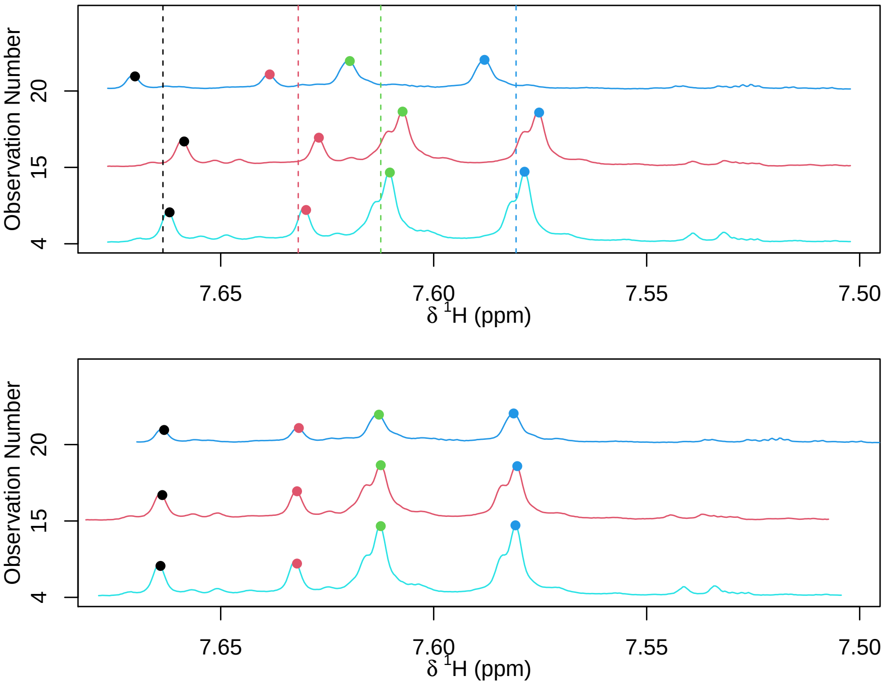


Figure 4.27: Example alignment of observations with the minimal number of peaks. The upper plot shows three spectra from region (A) with the geometric means of the $k = 4$ peaks indicated by dotted lines. The average deviation between the peaks in an observation and the geometric means provides the shift for that observation. The lower plot shows the same spectra after they have been shifted to provide an aligned subset from which the reference spectrum for correlation-based clustering is created.

by comparing the Pearson correlation (ρ) obtained when each peak identified in the observation is aligned to the maximum peak in the reference. The correlation between the unshifted observation and the reference is also considered. The alignment with the greatest correlation is considered to be the most suitable for clustering, ready for further analysis.

Zero padding is used to avoid peaks being introduced from neighbouring regions when the spectra are shifted and affecting the correlation coefficient. This method is illustrated in Figure 4.28, using region (B) from Figure 4.23. The unshifted spectrum for one example observation is shown below the reference spectra formed from the subset of observations containing only 6 peaks. The observation appears to be shifted to the right in comparison with the reference spectrum. Each of the possible alignments of a peak in the observation with the maximum in the reference is shown, with the aligned peak circled as well as the maximum peak in the reference spectrum. The correlation coefficients show the alignment of the sixth peak in the observation (pink) to the maximum in the reference spectrum to be optimal, with $\rho = 0.887$, whereas the next highest correlation is just $\rho = 0.400$. Figure 4.29 shows that this correlation-based method also works well for regions (A) and (C) from Figure 4.23.

For (A), alignment to the fourth peak (blue) gives the best correlation, and for (C), the correlation is the same for both the raw spectrum and alignment to the second peak (red), so either may be used in further analysis.

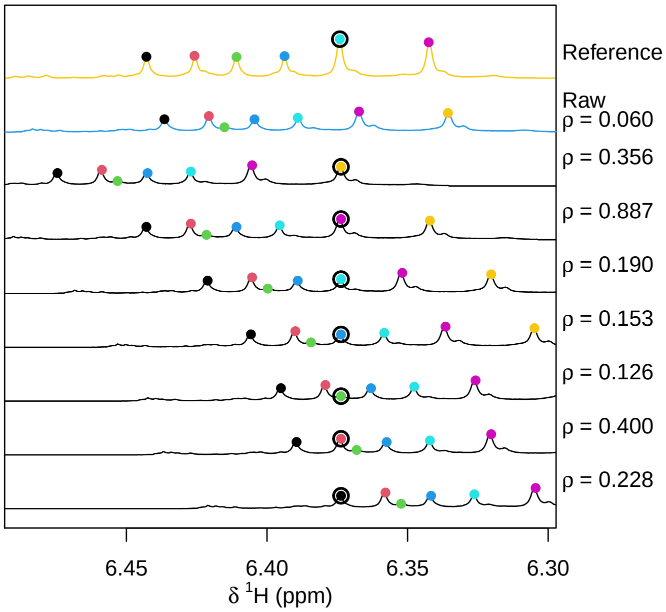


Figure 4.28: The correlated alignment of an observation, using the region shown in Figure 4.23(B) with $k = 7$ peaks. The unshifted spectrum (blue) is show below the reference spectrum (yellow) with each of the possible alignments below (black). The correlation coefficient (ρ) for each alignment is shown. The peak in the observation that is being aligned to the maximum peak in the reference is circled.

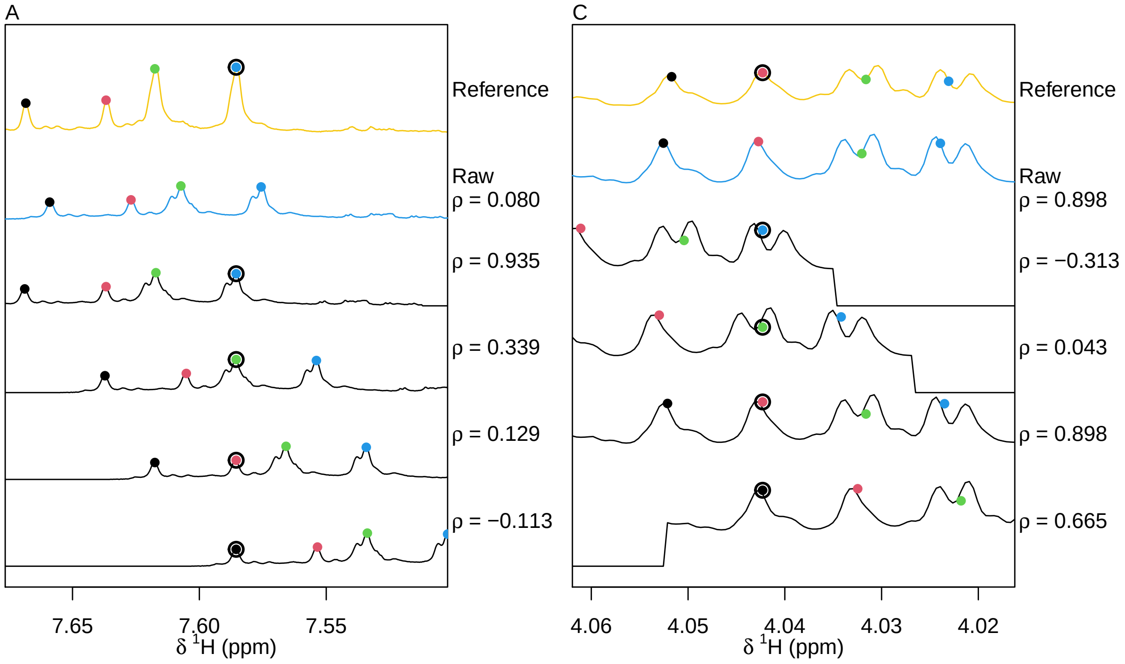


Figure 4.29: The correlated alignment of an observation, using the regions shown in Figure 4.23(A) and (C). For (A) the alignment with the blue peak is selected, and for (C) no shifting and alignment to the second peak (red) are equivalent.

With all observations aligned (shown for the example regions in Figure 4.30), k-means clustering is performed with k again being the minimum number of peaks in any observation within a region. This allows the regions to be binned using the minima between peaks as bin ends for each cluster, providing an equal number of variables for each observation ready for further analysis.

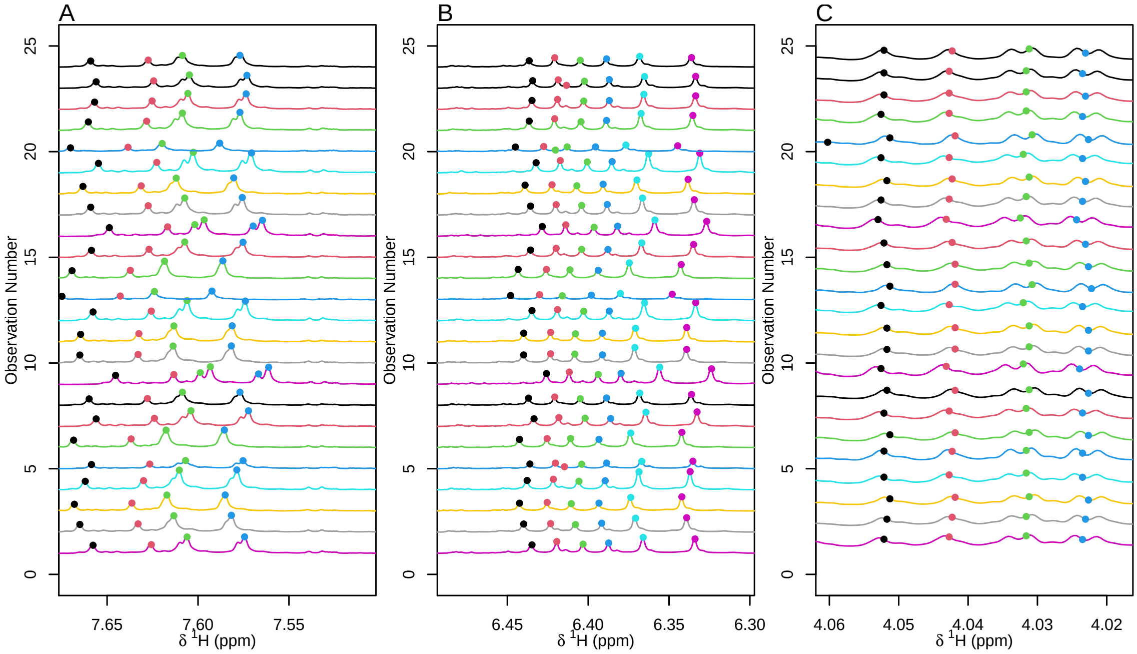


Figure 4.30: Results of the k-means clustering after the correlation-based correction process has been applied. Peaks are highlighted by points coloured by their peak index. For each of the peaks it is clear that the corresponding peak in each observation has been identified.

4.3.3.2 Supervised Alignment

The alignment described in the previous section may be improved by exploiting the fact that multiple peaks can be related to the same compound and therefore be shifted in a similar way. These related peaks can be identified from the correlation of ppm values for the observations maximum values. Spearman's rank correlation is used due to its insensitivity to translations between data points. As shifts can occur in either direction, the absolute values of the correlation are used to create a distance matrix and a dendrogram formed using average linkage. Figure 4.31 shows an example section of the dendrogram formed from the third level wavelet approximation to the coffee NMR data. Clearly four nodes (coloured blue, green and red) cluster tightly with a significantly lower distance than any others and these nodes will be used to provide an example of the supervised alignment.

Starting at the lowest branch point of the dendrogram in Figure 4.31 (coloured blue), the silhouette scores are calculated for clusters represented by the nodes

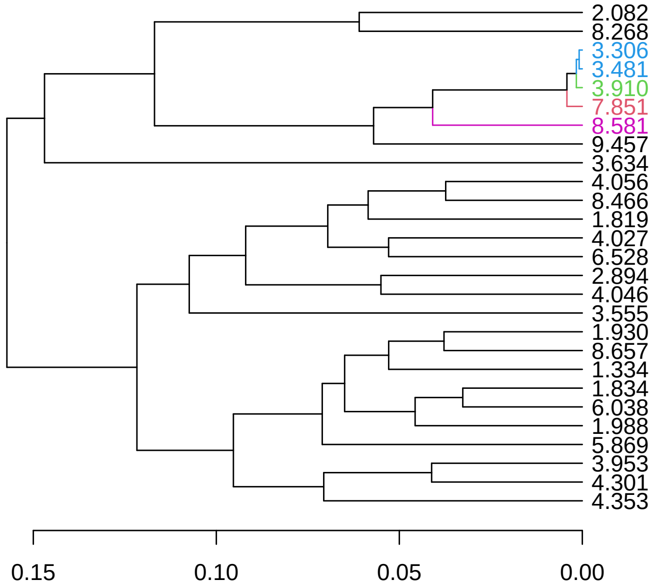


Figure 4.31: Section of the coffee dendrogram, using data from the third level wavelet approximation. Nodes are labelled by their region's maximum ppm.

(regions) in this branch. This provides a measure for how 'complicated' the peaks involved are. The four (unaligned) regions with the smallest distances in the dendrogram are shown in Figure 4.32 together with their silhouette scores. With the spectra coloured according to their origin, the similarity in the pattern of shifts can be clearly seen, although the magnitude of the shifts differ. The best alignment is more obvious for some regions than for others and the idea is to use the most obvious alignment as the 'driver' region to improve the alignment of other correlated regions. The region with the greatest absolute silhouette score is taken as the driver, as it is has the clearest clustering. For the example in Figure 4.32, the first region is selected. The correction used for this driver region is proposed for the other region in this branch of the dendrogram, after scaling by the average shift for the non-driver region. The scaling allows negative correlations, which were replaced by absolute values, to be accounted for.

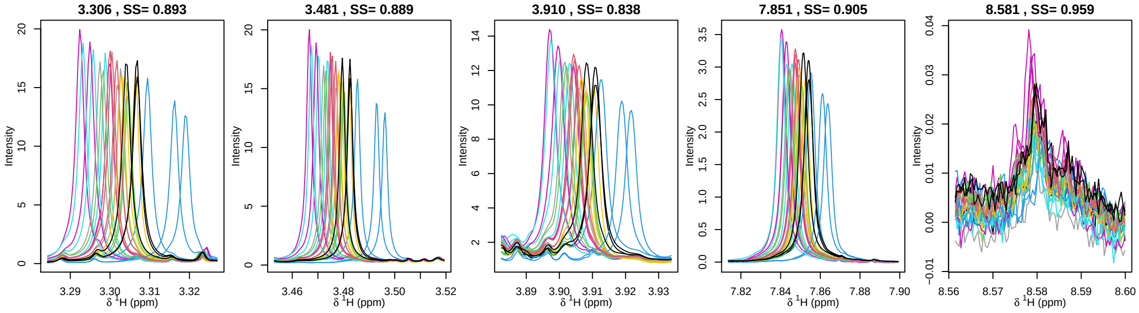


Figure 4.32: Regions with the smallest distances in the dendrogram of Figure 4.31 labelled with the ppm of the region's maximum and the silhouette score for the region.

Before committing to the correction based on the driver peak, silhouette scores are recalculated based on the proposed correction. If the majority of silhouette scores for regions that would be corrected at this point have improved then the process is repeated at the next branch point in the dendrogram with a potentially new driver peak chosen according to the original silhouette scores. If no significant improvement in the scores is found then the previous branch point is chosen and the corrections calculated up to this point are applied. The dendrogram is then pruned of these regions, and the process continued from the next lowest branch point until either all branch points have been inspected, no nodes remain, or a fixed height in the dendrogram is reached. This fixed height is set as the 50^{th} percentile of all branch heights as to increase efficiency while minimising the effect it has on valid correction. This dynamic pruning approach allows a broader range of corrections to be tested, whilst stopping at the most appropriate branch.

After the first iteration in the example, the silhouette scores for the first two regions are recalculated as 1 (as all peaks are aligned perfectly) and 0.996 respectively. Due to the improvement, the algorithm goes to the next level, which now includes the region labelled 3.910 ppm. In this case, the driver is again chosen as the first region and again an improvement in silhouette scores is obtained. At the third branch point, now spanning four regions, a new driver, that labelled as 7.851 ppm, is selected and again increases the silhouette scores. However, with the region labelled 8.581 ppm chosen as the driver at the fourth branch point, there is an overall decrease in the silhouette scores. Therefore, the previous branch is selected and the four nodes within this branch are removed from the dendrogram, and their corrections (using the fourth region as the driver) are recorded. This method also provides information about which peaks are likely to be from the same compound although for each set this would need to be confirmed manually.

4.3.3.3 Optimal Scale Selection

The alignment described relies on the peaks identified within each observation and this depends on the resolution of the spectra. Smoothing the spectra can resolve issues due to additional small peaks or multiplets in some spectra, but over-smoothing can result in overlapping peaks being combined. As no single resolution is likely to suit all regions, a method to determine the best possible alignment has been developed, using wavelet approximations to provide different levels of smoothing.

The region-based clustering and supervised alignment algorithms are applied to the data at multiple levels of the wavelet transformation and the results compared using the silhouette scores for corresponding regions. The resolution with the greatest silhouette score is considered to be the most suitable and is chosen as the optimal scale. In the case that identical silhouette scores are obtained for different wavelet levels, the lowest resolution is chosen as this one would be least affected by noise. The chosen scale determines the bin boundaries which are then applied to the raw data for consistency with the Class 1 features. As the scale can differ between regions, bins obtained at lower resolution can encompass neighbouring regions. If this occurs, the lower resolution is selected so that the two regions are combined.

Figure 4.33 illustrates this multi-resolution approach using region C from section 4.3.3.1 as an example. At the highest resolution, six bins are formed as indicated by the different coloured regions. The resolution is reduced with approximations at different levels of the wavelet transform, until at the fourth level, the two pairs of doublets on the right are smoothed down to single peaks.

Comparing silhouette scores for the region around 4.05 ppm shows the fourth level transform to be optimal with a score of 0.983. According to the silhouette scores, the fourth level also provides the optimal scale for the second region. The third level approximation gives the best score for the next region which merges the third and fourth regions of the highest resolution. Although the fourth level transform is found to be optimal for the region that is fifth at the highest resolution, the first level is better for the sixth region. These last two regions illustrate the issue of overlapping regions, where here the fourth level is chosen and the features combined.

The final step in the algorithm applies the method used for Class 1 peaks (Chater 4.3.2), although the allowed window width w around the ppm corresponding to a region's maximum intensity is doubled to allow greater shifts. Corrected regions are considered as Class 2 if they now meet the criteria, otherwise they are considered Class 3.

4.3.4 Comparison with existing methods

The performance of the MSFE method was compared with other feature extraction methods using the coffee NMR data (n = 24), unscaled and normalised to the TSP peak. The raw data was used as a baseline with variables obtained using uniform binning, adaptive binning and Speaq 2.0 in PLS-DA with the country of origin as response. LOO-CV was used to compare the accuracy of each method. In addition to using three subsets of the MSFE features, Class 1 only, Class 1 and Class 2 together, and Classes 1, 2 and 3 together, two versions of adaptive binning and Speaq were also tested. Adaptive binning was used with a reference spectrum at two different resolutions (AB-3 and AB-6) and two versions of Speaq were tested, one including

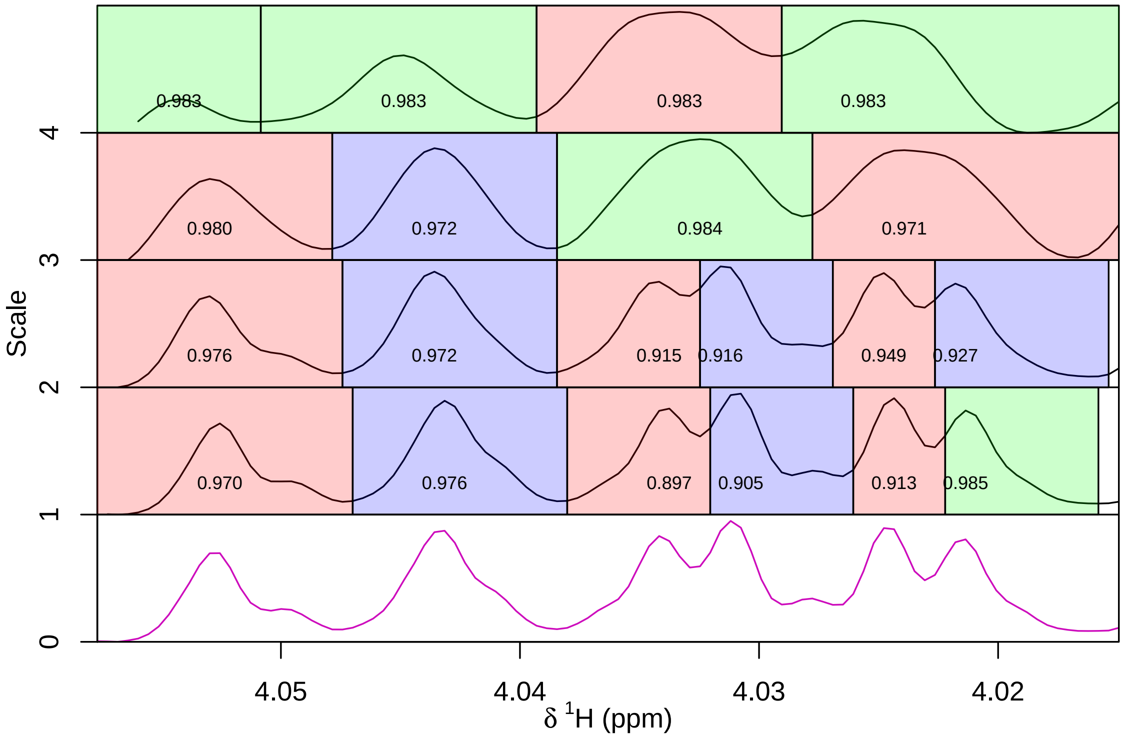


Figure 4.33: Optimal scale selection, using region C from Section 4.3.3.1 as an example. The green regions have the greatest silhouette scores corresponding to peaks found at the first level. Red and blue regions are coloured only to be able to distinguish between neighbouring regions.

neighbouring peaks (denoted NP), and one that does not make use of them (NNP). As the optimal number of components could differ between pre-processing methods, various numbers of components were considered and the accuracies obtained are shown in Figure 4.34.

As expected the raw data achieves a low accuracy, with the greatest being 16.7% for 3 components not much higher than random selection (12.5%). Uniform binning, using a fixed bin width of 0.04ppm and adaptive binning with a 3rd level reference spectrum (considered most appropriate for this resolution) perform similarly, achieving 25% and 30% respectively with 5 PLS components. Using a higher wavelet level (6th) improves the results, suggesting that peaks with larger shifts are useful for classification, although the variables extracted are too broad to be useful for determining the compounds responsible. As discussed in section 4.2 COW was used twice. After applying the COW algorithm PLS-DA was performed on the full spectra, as well as after AB was applied using a third level wavelet transformation. These were also very poor at classifying the coffee based on origin. Both Speaq 2.0 methods have surprisingly low accuracy, not much better than the raw data.

For the MSFE method, a window width limit of 2 was applied as this approximates the half-height peak width at the resolution of the data. Using only Class 1 data increases the accuracy significantly, achieving 62.5% for 5 components. Including Class 2 data increases the accuracy further, with 91.7% for 5 components, although it plateaus after 3 components. However, including the Class 3 data causes a slight decline in accuracy showing that these few regions can have a detrimental effect on the quality of the data, suggesting they should be excluded from the analysis in this set of data. In the model formed including the Class 3 features, one of the Class 3 peaks has the second greatest VIP score (shown in Appendix A.2 Figure A.3. Due to the noise present in this feature, this explains why the accuracy is reduced.

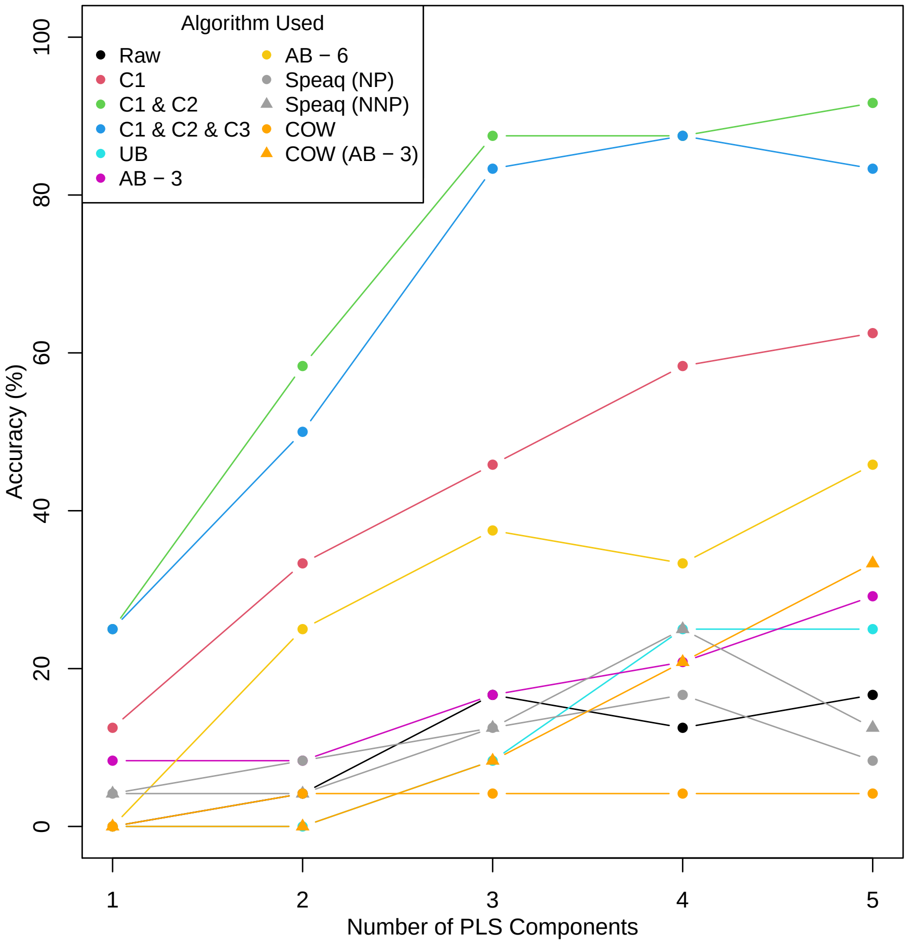


Figure 4.34: PLS-DA classification accuracy obtained for various feature extraction methods. Only adaptive binning and MSFE achieve an accuracy significantly higher than random classification.

It is clear that, by considering multiple scales to extract as many useful variables as possible and using different techniques appropriate for the quality of the data, it is possible to achieve good results from NMR data exhibiting large shifts between observations.

4.3.5 Summary

By classifying regions of the spectra according to the perceived quality of the data, the maximum number of reliable variables are extracted for further analysis. This reduces the risk of false conclusions drawn from poor quality data, for example as shown in Figure 4.12.

This has been achieved making use of a wide range of techniques, each found to be the most appropriate for handling the data at each stage. The Adaptive Binning algorithm allows for noise regions to be identified and discarded. Examination of the bins formed in this algorithm also allows for extraction of features which do not have a significant shift first means that these features will not have any new issues introduced, and these features can be used in analysis safely. For features which do contain shift these were corrected by first finding an appropriate correction within a group, and then correlating this with all other features in order to both validate the correction in shift and to identify features which might be from the same compound. A metric was used to test if the resulting correction improves the quality of the data, and for the features which did display an improvement these form the Class 2 features, and where there was no improvement these are Class 3.

The features extracted from this method was compared with features from existing popular methods, and if was found to result in notably higher accuracies with complex data.

4.4 Analysis of coffee data after MSFE

As the Class 3 features obtained by MSFE were shown to decrease classification accuracy, they were not included in this analysis. PCA was performed using only Class 1 and Class 2 features and the first 4 components are shown in Figure 4.35. Only the 10 greatest loadings for each pairing are shown, with four ppm values (3.80, 7.58, 1.33, and 3.91) appearing in both plots. Good clustering with origin can now be seen although PC1, which accounts for 64.4% of the total variance, shows the least separation between origins with PC3 having the greatest.

4.4.1 Classification of origin

The accuracies obtained using PLS-DA with up to 5 components can be seen in the comparison in Figure 4.34. Here the PLS model built using just three components is used, as the accuracy can be seen to plateau at this point. The classification results achieved through LOO-CV are shown in Table 4.6 where the overall accuracy is

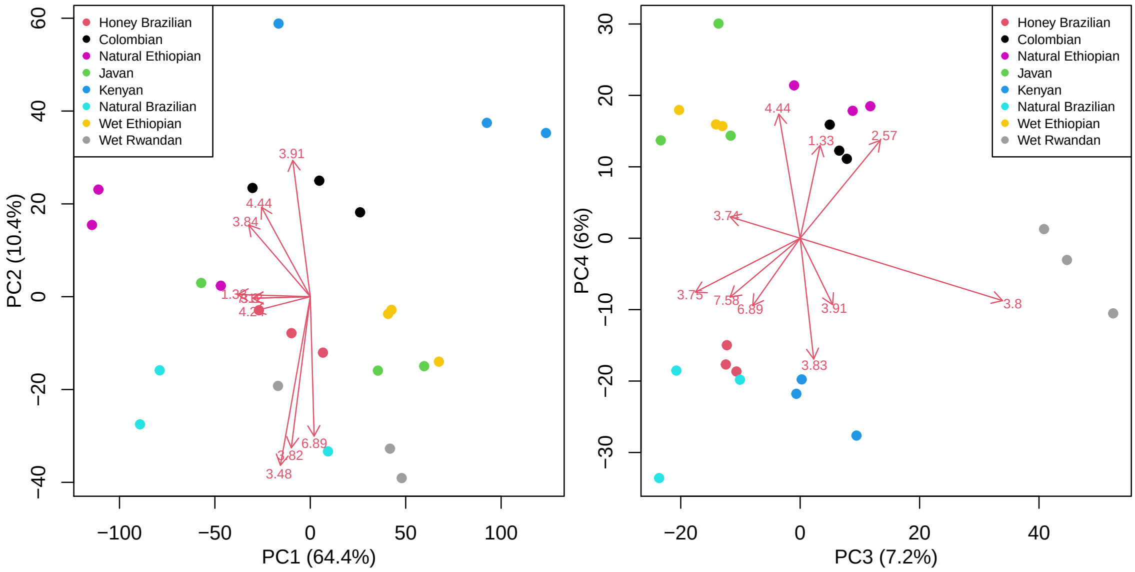


Figure 4.35: Biplots for the first 4 components obtained by PCA of the Class 1 and Class 2 features. There is a clear separation between origins. For clarity, only the greatest 10 loadings are shown for each pair of components.

87.5%. Just three observations are misclassified. Of these, one natural Ethiopian observation is misclassified as natural Brazilian and one natural Brazilian is classified as natural Ethiopian. The reason for these two observations being confused is clear from the PCA scores plot in Figure 4.35 where one observation from each group lies between the two groups, closer to each other than to the other replicates from their group. The final misclassified observation is a honey Brazilian observation classified as wet Ethiopian.

		Predicted							
		Honey Brazilian	Columbia	Natural Ethiopian	Java	Kenya	Natural Brazilian	Wet Ethiopian	Wet Rwandan
Reference	Honey Brazilian	2	0	0	0	0	0	1	0
	Columbia	0	3	0	0	0	0	0	0
	Natural Ethiopian	0	0	2	0	0	1	0	0
	Java	0	0	0	3	0	0	0	0
	Kenya	0	0	0	0	3	0	0	0
	Natural Brazilian	0	0	1	0	0	2	0	0
	Wet Ethiopian	0	0	0	0	0	0	3	0
	Wet Rwandan	0	0	0	0	0	0	0	3

Table 4.6: Confusion matrix showing LOO-CV classification results for the NMR data, obtained using PLS-DA with origin as the response.

VIP scores were used to determine the variables' importance in this model. Figure 4.36 shows these ordered by MSFE class rather than ppm, with Class 1 coloured black and Class 2 in red. The largest VIP scores are from Class 2 with no Class 1 variables achieving a score greater than 1, an accepted threshold for VIP scores. In comparison, 41 of the Class 2 variables are above this threshold. The feature with the greatest VIP score (~ 8) is at 3.801 ppm, followed by the features at 3.907 ppm and 3.743 ppm.

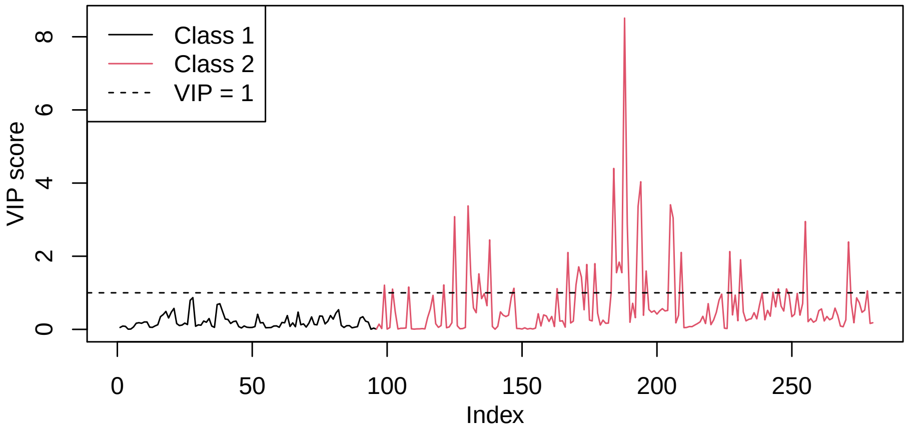


Figure 4.36: VIP scores for the Class 1 and Class 2 features, ordered and coloured by MSFE class. A dashed line is added at $VIP = 1$, which is the usual threshold to determine whether a variable is important or not. This shows that all important features belong to Class 2.

Each of the features with high VIP scores were inspected to see how the intensities change with origin. For the feature at 3.801 ppm the wet Rwandan observations have a higher intensity to many of the other origins, which are all similar. This feature therefore could act as a marker for this origin.

4.4.2 Analysis of taste intensity

For this analysis three of the origins (Natural Brazilian, Wet Ethiopian and Wet Rwandan) were omitted to avoid any potential differences in taste that may arise with the different processing steps. This reduced the number of observations to just 15, which were used in PLS-R using the Class 1 and Class 2 features and the expert taste scores as response variables. A triplot of the scores and Y-loadings is shown in Figure 4.37, which highlights the similarities between the X and Y scores and suggests clear differences with regard to taste. As expected from the taste scores, the Brazilian observations cluster with the nut and chocolate Y-loadings, and the Ethiopian observations cluster well with berry, citrus and floral tastes.

VIP scores were calculated from 3 PLS components for the response variables combined. The 12 most important features were selected (limited due to only having 15 observations), all of which are Class 2. Linear regression models were created for each taste individually using these 12 variables. Step-wise regression to minimise the AIC was performed and the p-values for two-sided t-tests were obtained from

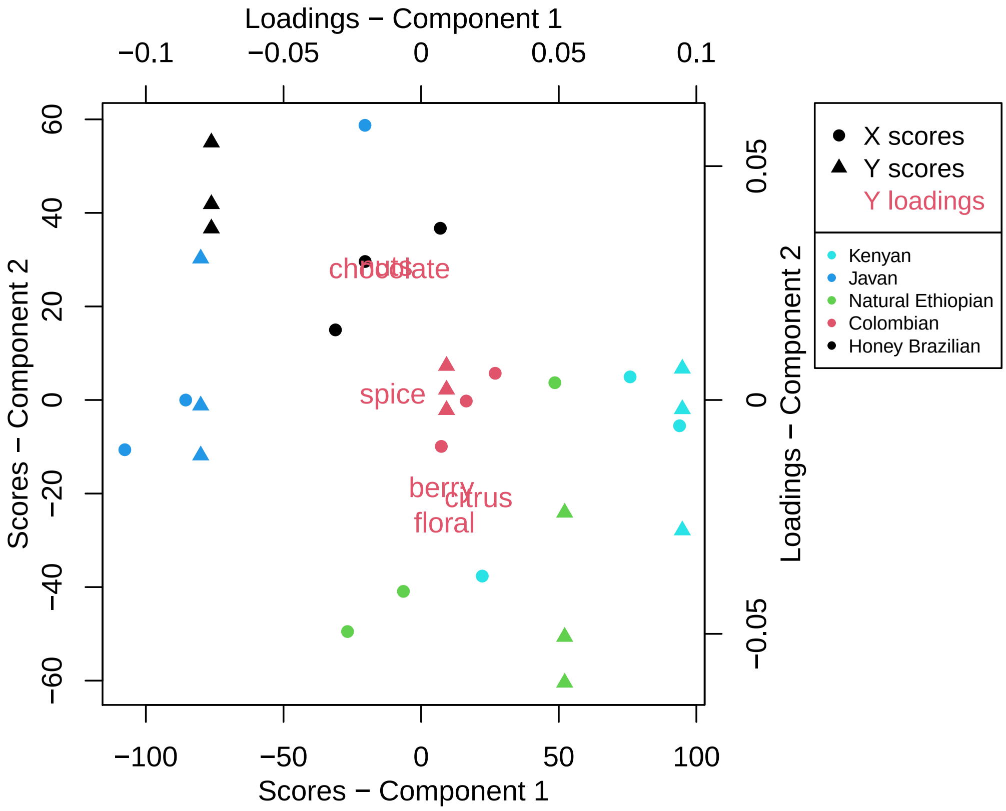


Figure 4.37: Triplot showing the first two PLS components' scores for both X and Y together with the loadings for Y, obtained using MSFE features extracted from the NMR data. Clustering according to origin can be seen.

the final models. The null hypothesis for the t-test is that the coefficient for each parameter is 0, and the alternative hypothesis is that it is not 0. Boot-strapping applied using the algorithm described by Finos and Salmaso [132] in order to produce accurate p-values. The fitted values calculated for each taste are plotted in Figure 4.38. Although there are quite large residuals, there is still a noticeable relationship between the fitted values and the expert taste scores.

Figure 4.39 shows the significance levels for the 12 features in the regression model for each taste response. Out of the twelve components, three of them have p-values less than 0.05, and two of these contribute to a second taste with p-values less than 0.1. The feature at 3.478 ppm contributes to chocolate at the 95% confidence level, and at 10% confidence it also contributes to nuts. Based on the PLS-R latent variable plot (Figure 4.37) we would expect to see similar compounds for these two tastes. The second feature 4.235 ppm is linked to to the spice taste, and at a lower confidence level also citrus. Finally, the feature at 4.443 ppm only has a significant contribution to the feature at 4.443. Despite being selected based on their high VIP scores, one of the features found at 3.821 ppm is not included in any of the step-wise models.

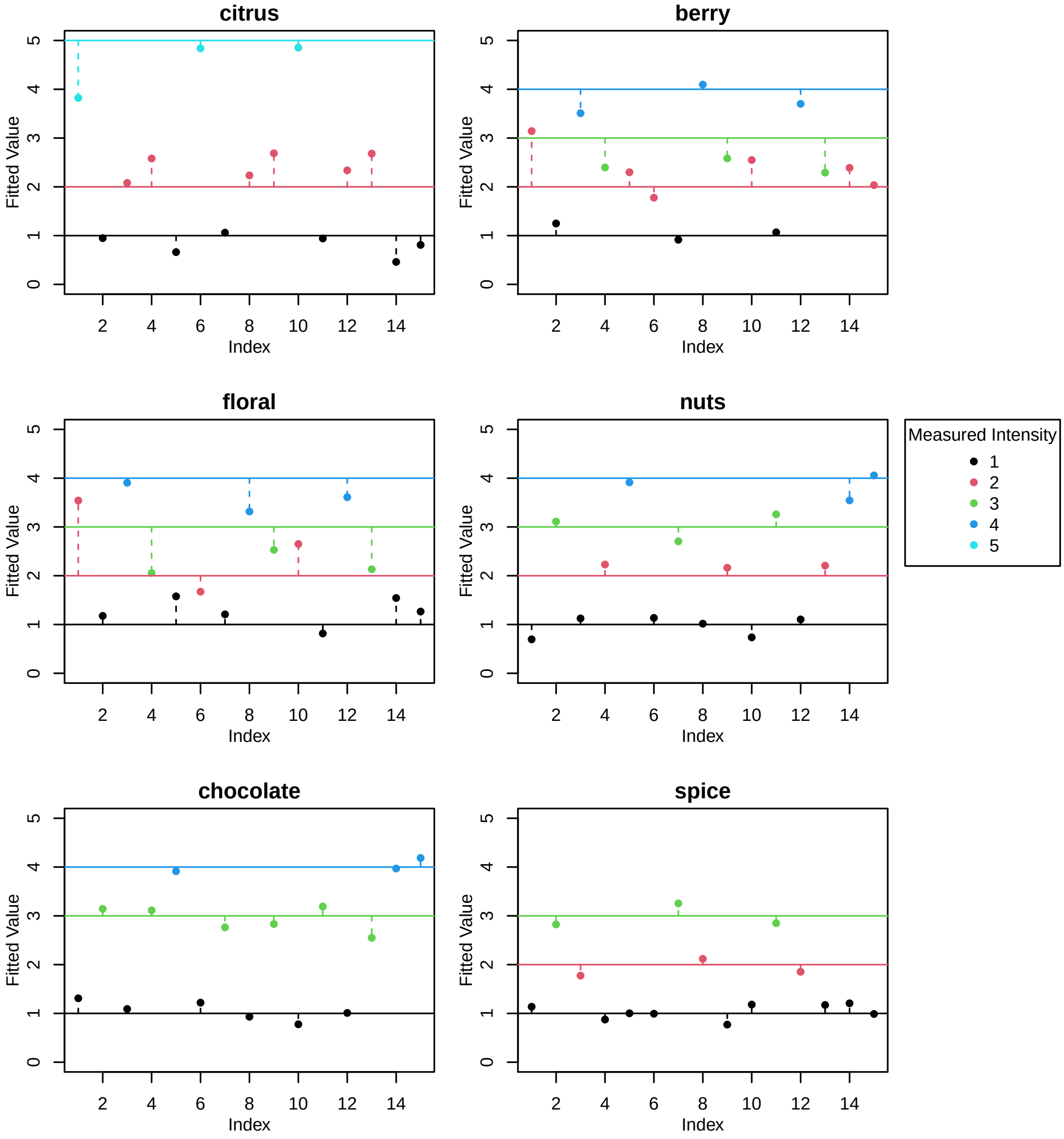


Figure 4.38: Fitted values for the taste levels obtained by linear regression using just 12 variables with the greatest VIP scores in the PLS model, coloured by the expert taste score. Vertical dashed lines are added so that the residuals are easier to visualise.

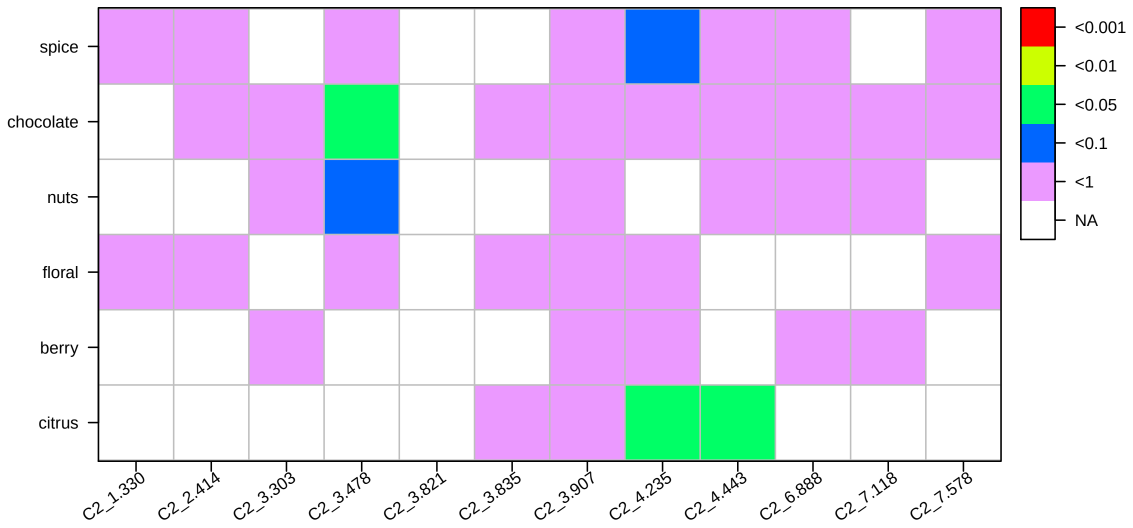


Figure 4.39: Significance of the NMR features with greatest VIP scores in linear models for each taste. p-values are colour-coded with white indicating variables that were not used in the step-wise models.

4.4.3 STOCSY applied to key compounds

STOCSY was used in order to try to find features which are strongly correlated to three of the features identified as being important for classification of either taste or origin.

The first driver peak chosen was the variable with the greatest VIP score (seen in Figure 4.36) location at 3.801 ppm. STOCSY only highlights small peaks located around the driver peak as being highly correlated with this driver (Figure 4.40). By making use of this information when attempting identification of this driver it should help reduce the number of potential compounds.

The second driver chosen is the feature located at 3.478 ppm, that contributed to multiple tastes. Here a significantly larger region of the spectra is highlighted as being correlated compared to the previous driver (3.801 ppm). Two regions containing the features with the greatest correlation are also shown. Due to the large number of correlated regions this has found, only the top few are suitable for aiding with compound identification.

Figure 4.42 shows the STOCSY analysis for one of the correlated regions shown in Figure 4.31 in MSFE. There are three features, at 3.91 ppm and 7.85 ppm and 3.48 ppm which have a high correlation coefficient (0.95), all of which were also identified as correlated in MSFE. These three peaks, when combined with the driver feature, should make potential identification simpler.

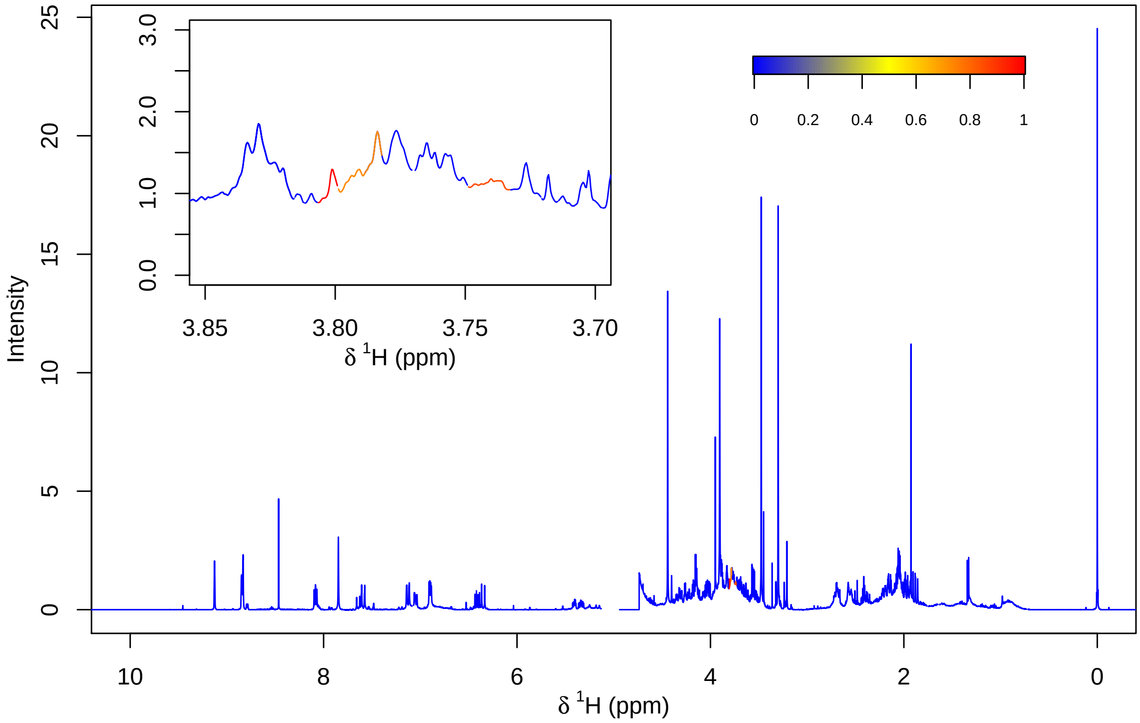


Figure 4.40: STOCSY analysis for the driver peak at 3.801 ppm. The spectrum is coloured based on the correlation between each MSFE region and the driver region. A close up of the only correlated features is also shown.

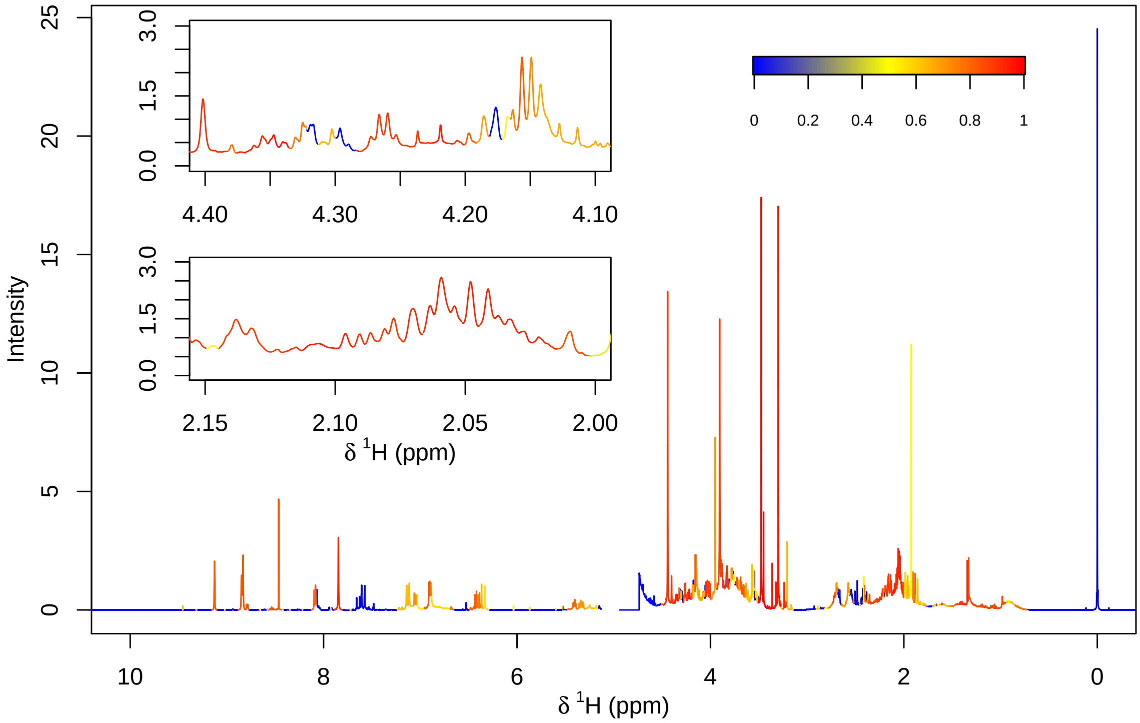


Figure 4.41: STOCSY analysis for the driver peak at 3.478 ppm, coloured based on correlation. Close ups of two correlated regions are also shown.

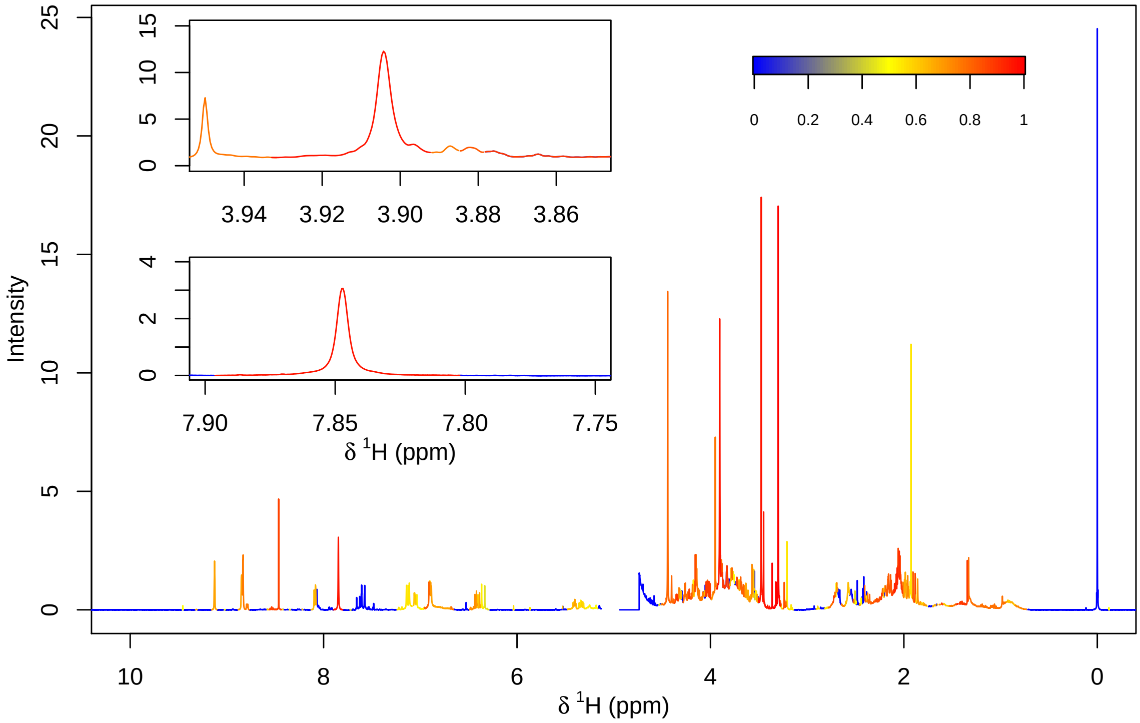


Figure 4.42: STOCSY analysis for the driver peak at 3.303 ppm, coloured based on correlation. Close ups of two correlated regions are also shown.

4.4.4 Tentative compound identification

In order to try to identify the compounds which were found to have interesting properties, the Spectral Database for Organic Compounds [118] was searched using their ppm values. Not all features were able to be identified, however those with potential matches are discussed here.

First, the peak around 3.478 ppm, found in many of the tastes, could potentially relate to caffeine and eosin Y [133]. Using the correlated peaks from both MSFE and STOCSY the feature at 3.30 ppm is also found to be one of the caffeine peaks. Considering that caffeine is known to be linked to roasting and grinding level [134], [135] it is possible that this is not the feature directly responsible for the tastes, and instead suggests that the features which are are also linked to the coffee processing. This also can explain why a significant portion of the spectra was found to be correlated with it via STOCSY (Figure 4.41).

The second possible match, for the compound responsible for the peak at 2.414 ppm, could be 5-Methyl furfural, which has a caramel flavour [136].

The final tentatively identified compound is for the feature with the greatest VIP score from the PLS-DA model with origin as the response (3.801 ppm). This could be either 2-(aminomethyl)pyridine or dimethylsilane, both known to occur in coffees

[137], [138].

4.4.5 Discussion

The ^{1}H NMR data for coffee was prone to errors relating to extreme shifts in peaks between observations. Multiple methods were applied to the data, such as Adaptive Binning, which was shown in Section 4.1 to work well with small shifts, however these were unable to account for the shifting without introducing new adverse affects to the data, such as over-smoothing region with high level wavelet transforms. A novel approach was devised which builds up from Adaptive binning which is able to account for extreme shifts, while minimising any negative effects introduced. The classification of regions allows for control over which features to use in analysis. For the coffee data, likely due to the number of regions with shifts, it was shown to significantly improve the classification rate in a PLS-DA model for the origin when including the Class 2 regions, and then the few Class 3 features made a negligible negative difference.

Using the Class 1 and 2 features, models were created for the origins and taste intensities. Using PLS-DA a model was formed which is able to accurately predict the origin of these samples, validated using LOO-CV. Primarily Class 2 features have the greatest importance within this model, and this was found to potentialy be one of two compounds found in coffees before. When modelling the intensity of the six tastes a good fit of the data was achieved. A subset of the features were used in this model, and these were selected based on their VIP score in a PLS-R model for the data. Only three features were found to be significant in contributing to the taste, and one of these was identified as potentially being caffeine. STOCSY was able to identify a number of other features which were correlated with the driver peaks, and this provided extra features to be used when trying to identify compounds. A number of other features were able to be identified, however it was not possible for each feature of interest. By combining the MSFE features with other data sources via data fusion these can hopefully be identified.

Chapter 5

Mass Spectrometry data analysis

5.1 Analysis of Coffee LC-MS Data

As detailed in Section 3.2.6 LC-MS has both a positive and a negative mode, which results in two separate sets of measurements. While it is common to combine these two via data fusion, here each set is analysed separately initially. The aims of the analysis undertaken here is to explore the data to identify any trends between the observations relating to both origin and taste intensity, and to form models for these parameters. From these models the variables which have the greatest contributions can be found, and the compounds corresponding to these variables are able to be identified.

5.1.1 Positive Mode

Exploratory analysis of the positive mode data using Principal Component Analysis (PCA) showed batch differences to be a major source of variance. Figure 5.1 shows the scores calculated from unscaled data coloured according to the origin of the observation, biological replicates, analytical replicates and the order of data collection by batch. There is clear separation of the observations into two groups, one tightly clustered and a second with more variance. From Figure 5.1A and 5.1B it can be seen that the split has no relationship with either origin or biological replicate. In fact, Figure 5.1D shows that the separation is related to batch with the split occurring towards the end of batch 5. This is also evident in Figure 5.1C which shows the first set of replicates in the tightly clustered group, whilst the second set is split and the third set of replicates all appear in the second group. Further investigation revealed that the solvent was topped up at this point in the experiment, resulting in the variance in PC1, which accounts for 98.2% of the total variance. Only in PC6 does separation begin to appear based on the origin of the coffee observations, previously visible in the first component for the NMR data. Score plots for the first eight components are shown in Appendix A.3 Figure A.4. Traditional batch correction methods, using either QCs or background correction, were unable to correct the problem.

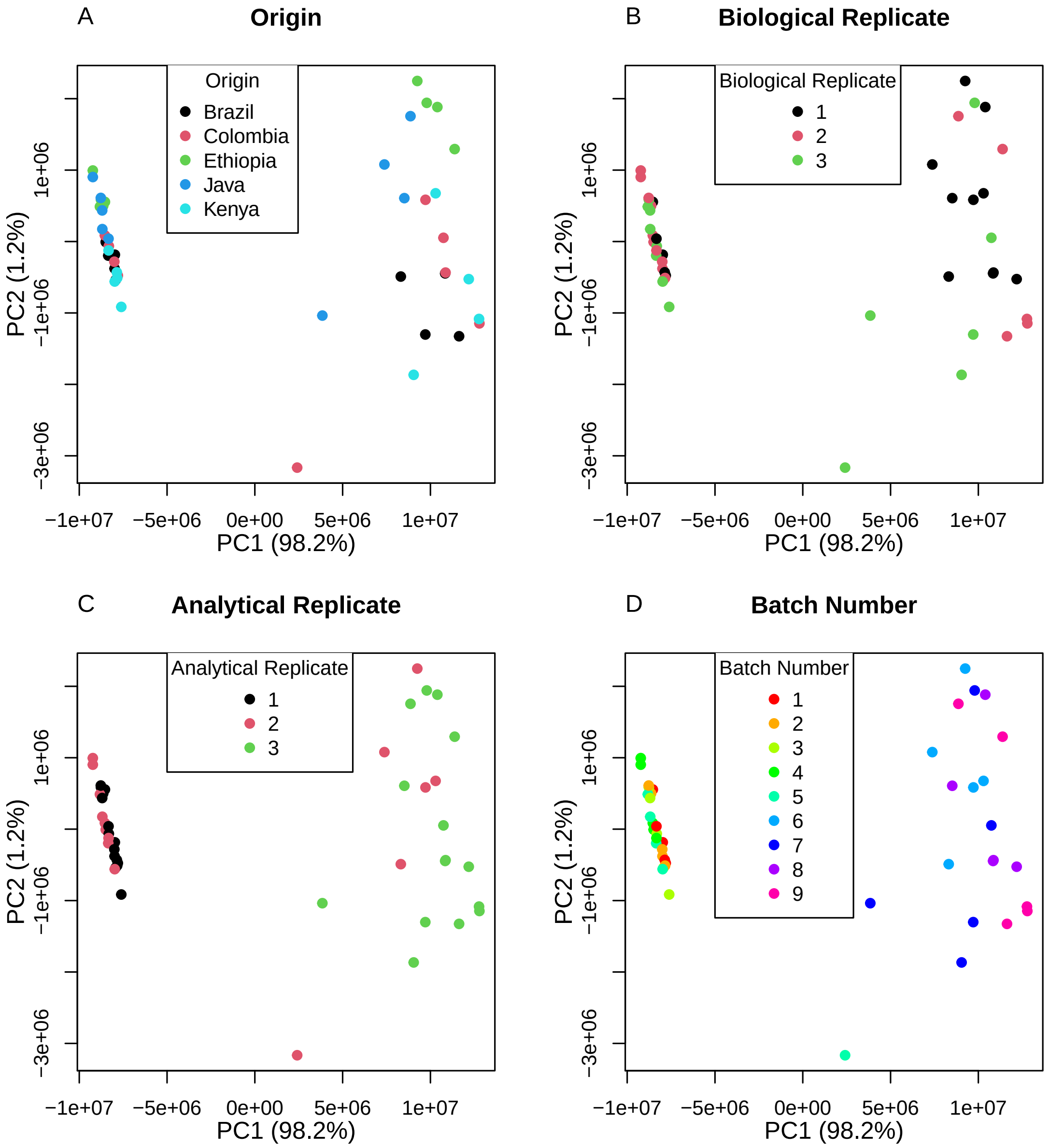


Figure 5.1: Scores plots of the first two principal components for positive mode LC-MS data, coloured by different factors. There is clear separation into two groups, unrelated to either geographical origin (A) or biological replicate (B). The split occurs within the second set of analytical replicates (C) towards the end of batch 5 (D).

Comparison of the total intensities for observations shows clear differences between the batches as can be seen in Figure 5.2. This data was normalised in Progenesis prior to feature extraction so further normalisation is not appropriate.

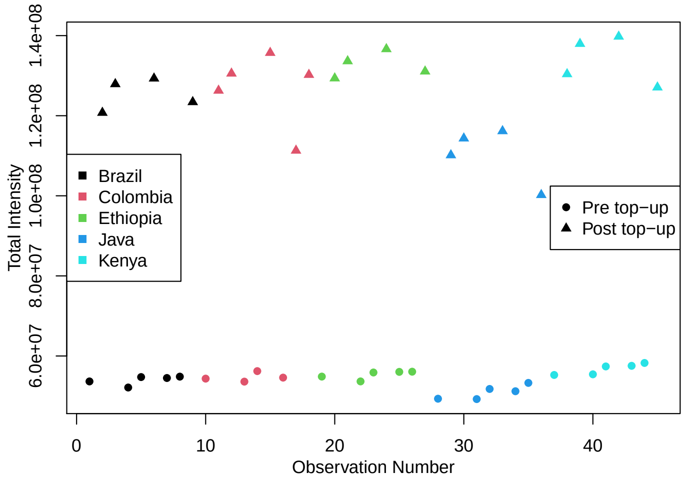


Figure 5.2: Total intensity for each observation in the LC-MS positive mode data. Here observations are coloured by coffee origin, and the symbol shows whether they were run before or after the solvent top-up.

Group standardisation transforms variables via UV-scaling within groups rather than across all groups and is usually applied to pre-defined groups, such as batch or origin. Here, however, the two groups considered were based on whether the LC-MS analysis was run before or after the solvent top-up. PCA was performed on the standardised data and Figure 5.3 shows the scores obtained for the first two components. The split of the two groups is no longer apparent and PC1 now shows clear separation according to origin with the most noticeable separation for the Javan observations.

LDA was applied to the first principal component to classify the data by origin and achieved 80% (n=45) using LOO-CV, with the confusion matrix shown in Table 5.1. The greatest confusion is between the Colombian and Ethiopian coffees with three Colombian observations predicted as Ethiopian, and following the trend seen in the PC1 scores, all Javan observations are correctly classified. LDA was repeated using two PC components to see how this affected the accuracy, despite no clear separation along this PC. In fact, the accuracy found via LOO-CV increased to 97.8%, with only one Ethiopian observation classified as Colombian. However, due to the scaling, many smaller intensity variables contribute to the variance and it is not possible to highlight particular variables that contribute significantly.

PLS-DA was used in order to generate VIP scores to identify variables that discriminate between origins. The X-scores show clear separation between origins (Figure 5.4), although again the only real separation is for the Javan coffees. The Y-scores are also found close to the corresponding X-scores suggesting a good fit for the two component model. VIP scores were calculated using the first latent variable (LV), and the top percentile of these were then explored for identification. Within

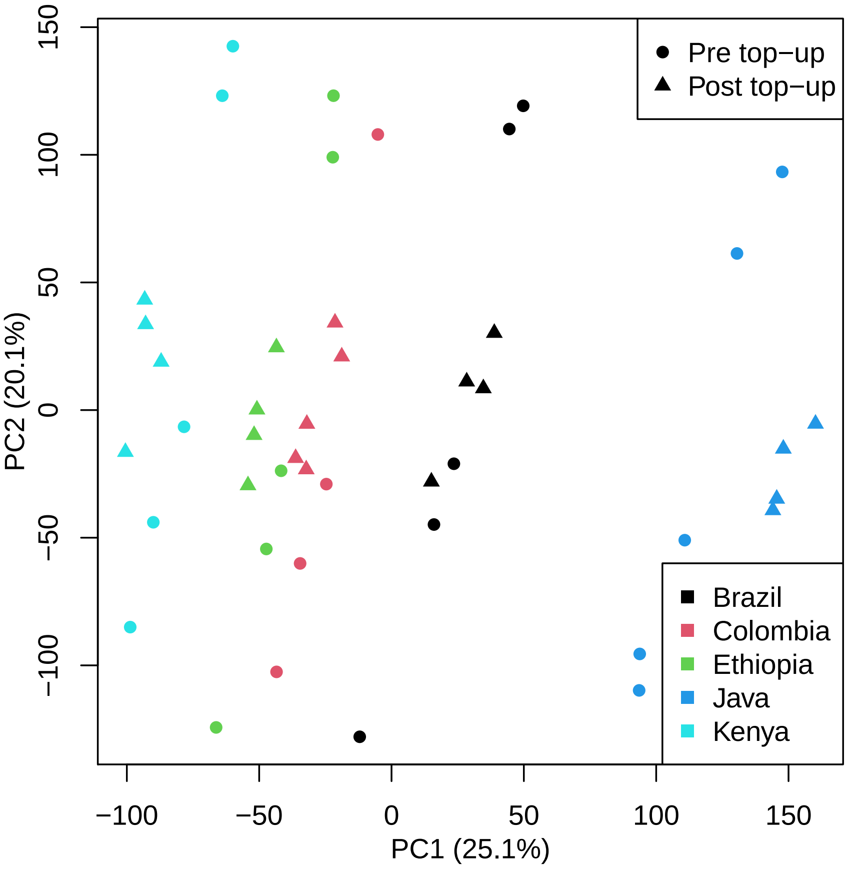


Figure 5.3: Scores plots for the first two principal components for the positive mode LC-MS data after group standardisation. Observations are coloured by origin with symbols indicating whether they were run before or after the solvent top-up.

		Reference				
		Brazil	Colombia	Ethiopia	Java	Kenya
Prediction	Brazil	8	0	0	0	0
	Colombia	1	6	2	0	0
	Ethiopia	0	3	6	0	2
	Java	0	0	0	9	0
	Kenya	0	0	1	0	7

Table 5.1: Confusion Matrix of the PCA-LDA model for origin using the positive mode LC-MS data, and only the first PC.

the top VIP scores many of the features have retention time of either 14.38 minutes or 16.76 minutes, and so potentially two compounds, and their fragments, are the key compounds for the separation seen in the first LV.

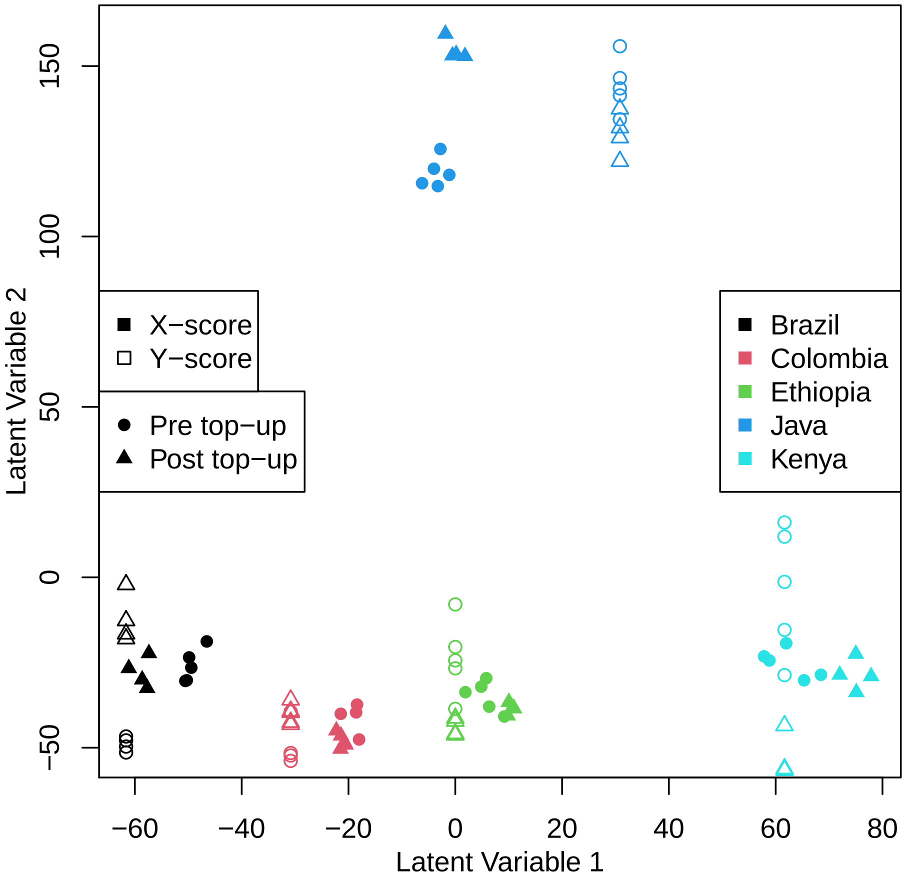


Figure 5.4: PLS-DA scores plots for the first two latent variables obtained for the positive mode LC-MS data with origin as the response, after group standardisation. Observations are coloured by origin with symbols indicating whether they were run before or after the solvent top-up and whether they are X or Y scores.

5.1.1.1 Analysis of taste intensity

PLS-R was used with the group standardised positive mode LC-MS data (X) and the taste scores as responses (Y). The X and Y scores along with the Y loadings for the first two LVs are shown in Figure 5.5. While the Y scores somewhat resemble the corresponding X scores the fit is still quite poor suggesting that more LVs are needed. This is also evident in the Y loadings, which show the correlation with origin is not as high as the taste scores in Figure 3.2 suggest. For example, Brazilian coffee and the spice Y loading are clustered, however it was measured to have a score of just 1 out of 5. In order to find how many components are needed to accurately predict the taste intensities the root-mean-square of error prediction (RMSEP) is calculated for each of the 6 models using all observations. This suggests that, while chocolate and spice only need three components, floral and nuts would require five to approximate taste scores.

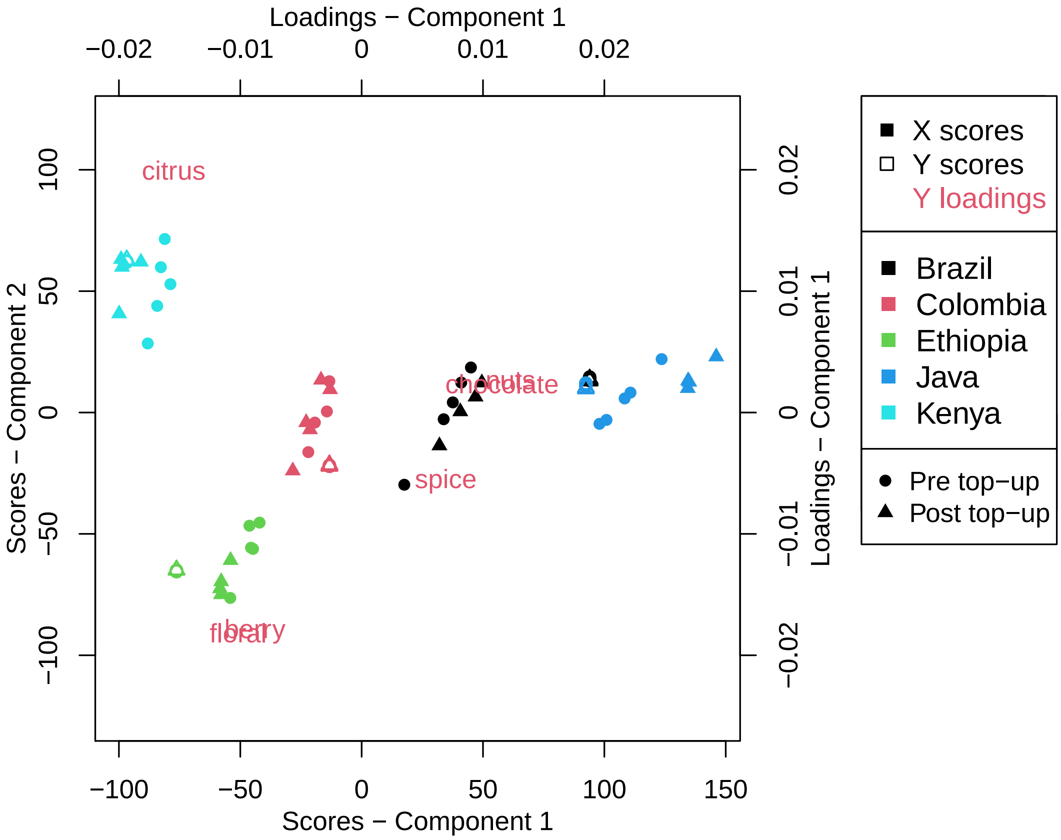


Figure 5.5: Triplot showing both X and Y scores as well as Y loadings for the first two PLS components obtained from the group standardised positive-mode data. The X and Y scores are distinguished by symbols, as are the pre- and post-solvent top-up groups, whilst colour indicates origin.

To test the prediction of individual tastes, a linear regression model was created for each of the six tastes using only the top 40 variables selected by VIP scores formed from five components. The same method as was used for the NMR analysis was used here, with step-wise regression to minimise the AIC being performed with boot-strapping applied to produce accurate p-values. Both the residuals for the models, as well as the significance levels for each variable can provide valuable information regarding how appropriate the model is. If there are large residuals and poor fitted values it suggests that there were too few LVs in the model as there is not enough information available to model the tastes, and the lowest p-values will be related to the compounds most likely to contribute to taste. k-fold validation is used with $k = 13$ in order to create training and test sets, and the predicted intensities for the test sets from each fold are shown in Figure 5.6. The models produced here produce notably higher residuals than observed with the NMR data (Figure 4.38), suggesting that while these components do have high VIP scores for modelling the taste they are not able to fully fit models. While including more components might improve the fit for these tastes, it would increase the risk of over-fitting the intensities, and so the five component model will still be used.

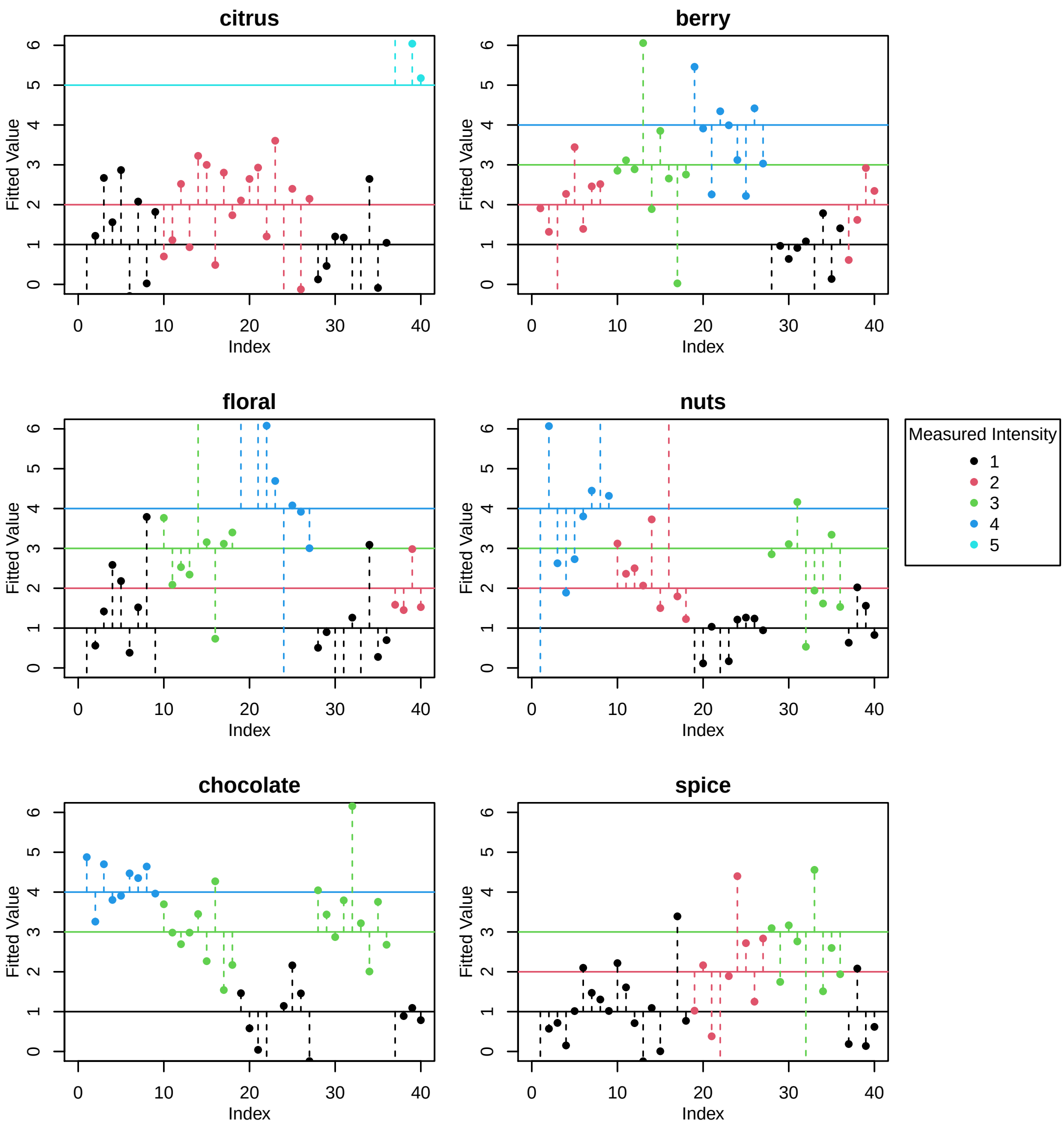


Figure 5.6: The fitted values for models created using step-wise regression with the 40 positive-mode LC-MS variables with the highest PLS-R VIP scores, coloured by taste scores. Vertical dashed lines are added so that the residuals are easier to visualise.

When plotting the p-values for these models (Figure 5.7) and colouring based on their significance level it is clear to identify how each of the 40 features affects the intensity of the tastes. Sixteen of the features are found to contribute to these tastes at the 95% confidence level, with one at the 99.9% confidence level. This feature, with m/z value 517.226 and retention time of 13.58 minutes, only contributes towards the chocolate taste. Within these 40 features, there are no p-values that a sufficiently low for the berry or citrus tastes. Identification of the features responsible, especially those that appear in multiple models, should provide important information about which compounds result in each of these tastes.

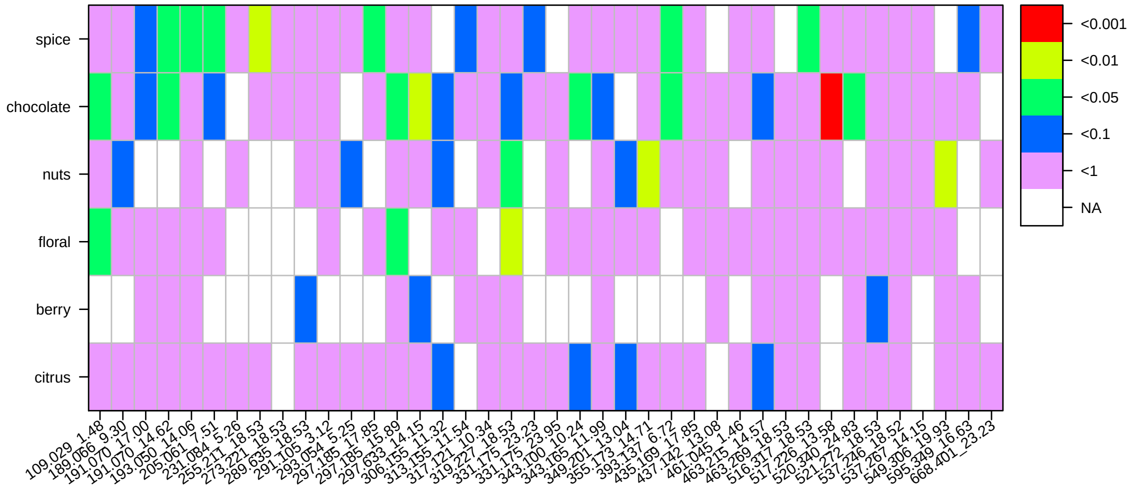


Figure 5.7: p-values for each feature from the linear regression models built using the 40 positive-mode LC-MS variables with highest VIP scores. Significance levels are represented by colour, with red being most significant (p < 0.001) and pink the least (p < 1). White shows variables that are not included in the final step-wise regression model.

5.1.1.2 Tentative Compound Identification

For the positive mode LC-MS several features were identified as being useful for analysis. For the origin the PLS model showed clear separation within the first component (Figure 5.4), and the features with the greatest VIP score within this eluted at two times. For the features with a retention time of 14.38 minutes these possibly are all from Quercetin-3-O-glucoside, a flavonoid [139]. The other important features have a retention time of 16.76 minutes, however no compound was found that fits the parameters. For the taste intensity models, the feature with m/z 349.201 and retention time 13.04 minutes, which appeared in each model with low p-values is potentially Fumaric Acid [140]. The feature with m/z 293.054 and retention time 5.25 minutes was found to possibly be Sebacic acid [140].

5.1.2 Negative Mode

PCA scores plots for the negative mode data (Figure 5.8) show the observations cluster very well by geographical origin. It can also be seen that the South American coffees, from Brazil and Colombia, cluster together, and although there is a difference along PC2 (8.4% of the total variance), African coffees, from Kenya and Ethiopia, are similar in their PC1 scores. There is no pattern with either biological or analytical replicate and one observation from each batch occurs in each cluster, suggesting no issue with data collection time.

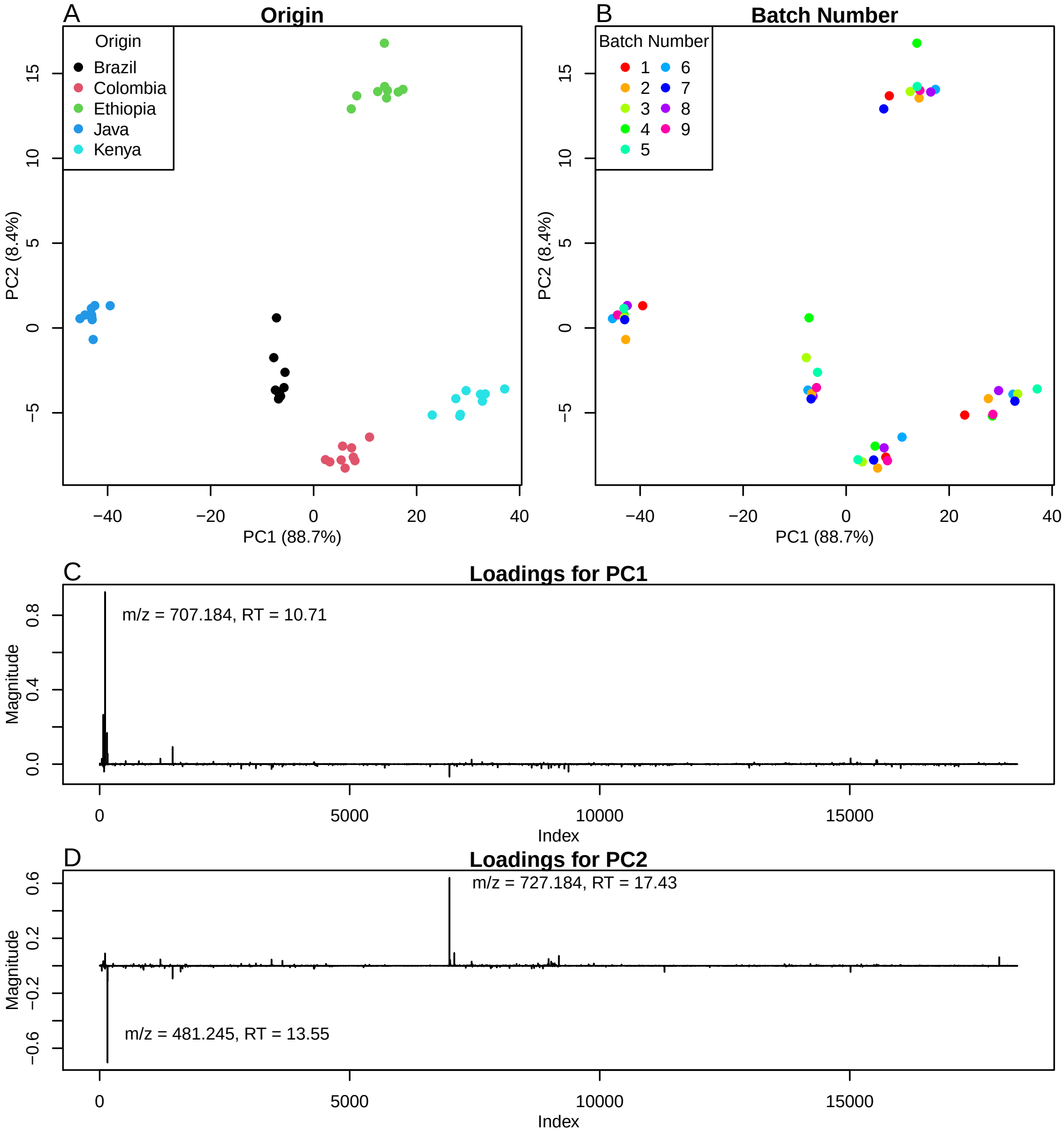


Figure 5.8: Scores plots of the first two principal components for negative mode LC-MS data, coloured by origin (A) and batch number (B). This shows clearly that geographical origin is the major source of variance. Line plots of the loadings for PC1 (C) and PC2 (D) show three variables with three significant loadings, labelled by their m/z value and retention time.

Figure 5.8 also shows the loadings for the first two principal components. One variable, m/z value 707.184 and retention time 10.71 minutes, has the greatest absolute magnitude in PC1 by a significant margin. Just two compounds, one with m/z value 727.356 and retention time 17.43 and another with m/z value 481.245 and retention time 13.55, stand out as responsible for the variance in PC2 where the separation of Ethiopian coffee can be seen. Figure 5.9 shows the relative intensities for these three peaks. It can be seen that, for the first feature plotted, the Javan observations have much lower values in comparison to other coffee origins. For the first of the two peaks related to differences along PC2, Ethiopian coffee stands out with higher values as expected from the scores plot. However, both Ethiopian and Javan observations have low intensities for the second of these two peaks, with Brazilian observations showing greater variance. The m/z values and retention times for these all initially do not show any clear connection with each other. Some decrease in intensity can be seen for several variables with the first Brazilian observation in particular (the first observation run) often having noticeably higher values.

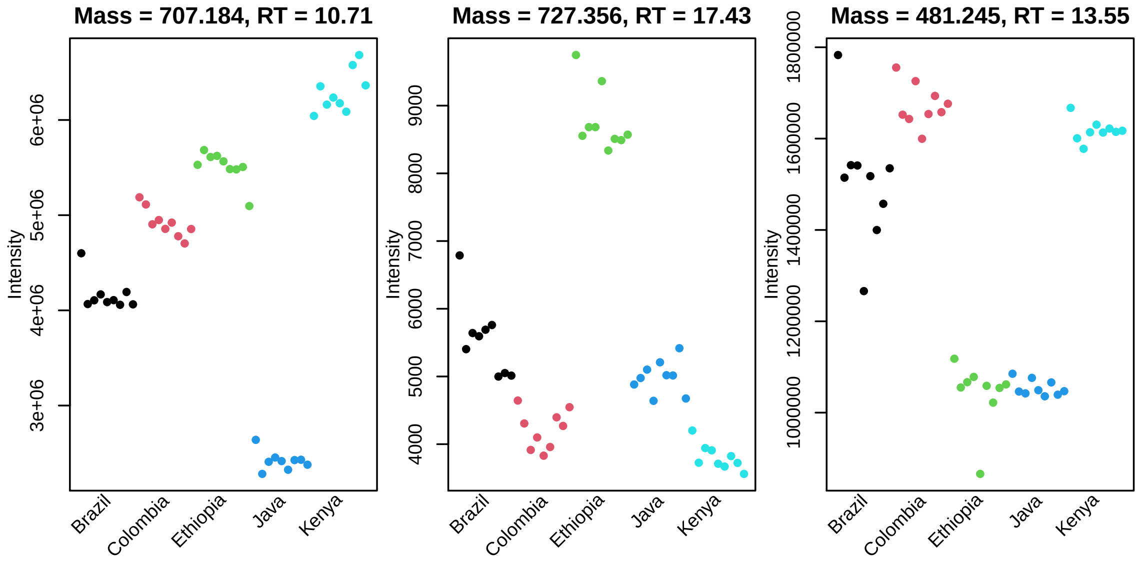


Figure 5.9: Relative intensities for the three variables corresponding to the loadings for PC1 and PC2 with greatest absolute values. The first plot shows a clear separation with the Javan observations with less, but still noticeable, separation of the remaining four origins. The second plot shows a large separation of Ethiopian coffee as seen in the PCA scores plot, and the final plot shows separation of Ethiopian and Javan coffees but greater variance within some origins.

The importance of the three variables identified by PCA in the discrimination of coffee origin is demonstrated by the dendrogram (Figure 5.10) obtained by hierarchical clustering using only these three variables (unscaled) and centroid based linkage. If the dendrogram is cut at the height represented by the dotted line, every observation is grouped with all other observations from that origin, with the only exception being observation 1 (Brazilian) which clusters with the Colombian observations. One interesting observation to be made from the dendrogram is that, at a height shown by the dashed line, three clusters are form, corresponding to the continents from where the observations originate, a surprising result considering only three variables were used. This suggests that there is an inherent connection between the observations based on their respective continent.

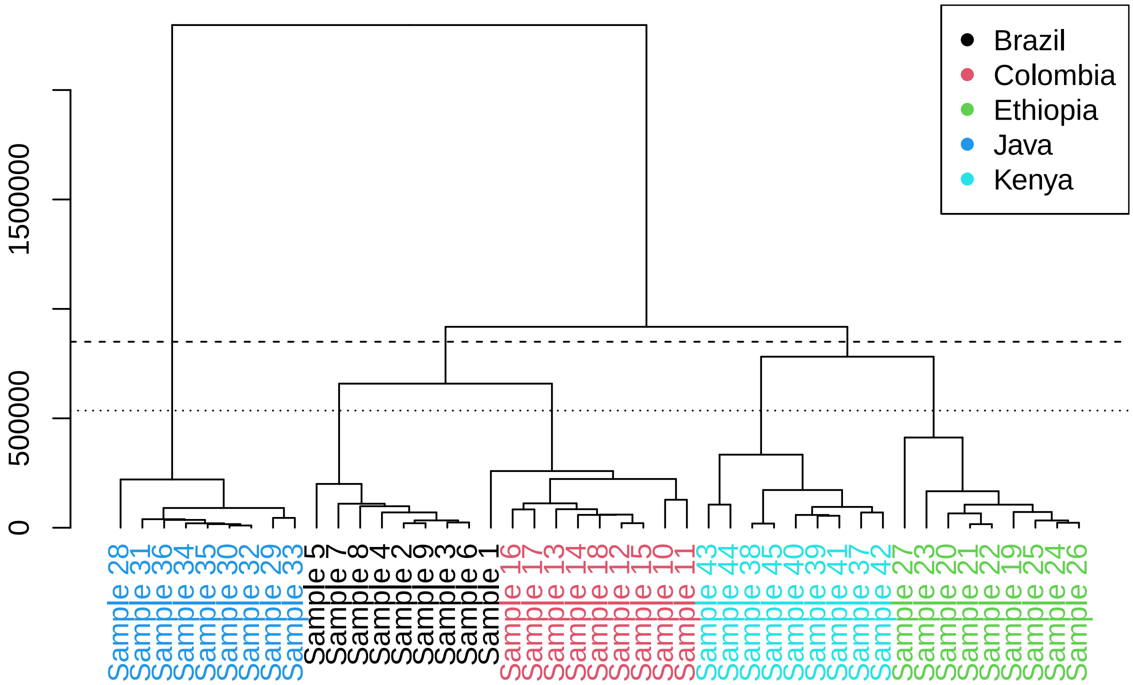


Figure 5.10: Dendrogram showing the results of hierarchical clustering using centroid linkage and the three key variables from LC-MS negative mode data, with the leaf nodes coloured by origin. The dotted line shows the height at which the origins form separate clusters, and the dashed line is the equivalent for continents.

PCA was repeated with UV-scaled data to give low intensity variables greater influence and prevent the greater intensity variables dominating the analysis. The three variables highlighted so far are among the greatest intensity in unscaled data, which is potentially why they have such large loadings. The scores plot for the UV analysis shows a tighter clustering for each of the origins (Figure 5.11) with the first two PCs accounting for 48.3% of the total variance. This suggests that lower intensity features do also discriminate between origins.

The loadings for two of the three variables identified as important for unscaled data (m/z values 707.184 and 727.356) remain high in these first two components from the UV-scaled analysis, but the third variable now has lower impact. The features with the greatest loadings in this scaled analysis are all within a small range of retention times with the top 7 all eluting between 19 and 20 minutes. The three variables with greatest loadings in PC1 and PC2 combined are shown in Figure 5.12.

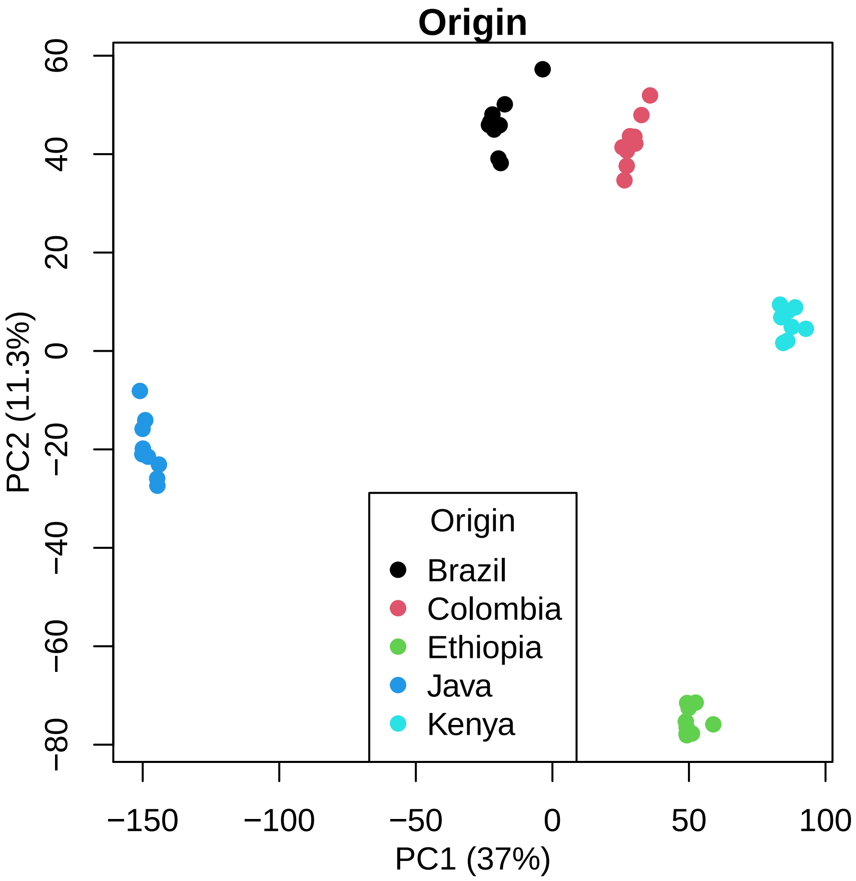


Figure 5.11: PC1 and PC2 scores for the negative mode LC-MS data, after UV-scaling. A similar pattern to that found for unscaled data (Figure 5.8) is seen but with less within-group separation in PC1 here.

All have very similar patterns and clearly aid the classification of Ethiopian, Javan and Kenyan coffees.

5.1.2.1 Supervised analysis of taste and origin

PLS was used with the taste scores as responses (Y variables) to identify any negative mode LC-MS (X) variables associated with different tastes. Figure 5.13 shows a triplot combining the scores plot for the first two latent variables for both X and Y with the loadings for the response variables. The X and Y scores for each origin appear quite close to each other, although the distance is greater for the Brazilian and Javan coffees. The Y-scores along the first component for Brazilian and Javan observations have similar values, showing that multiple components are necessary. The similarity between X and Y scores shows the PLS model provides a good prediction of the data, and so the greatest loadings should provide insight based on the origin and taste. The Y loadings reflect the taste scores given by experts shown in (Figure 3.2). Nuts and chocolate are located near the Brazil observations, which have the highest taste strength, spice appears nearest Javan scores, citrus close

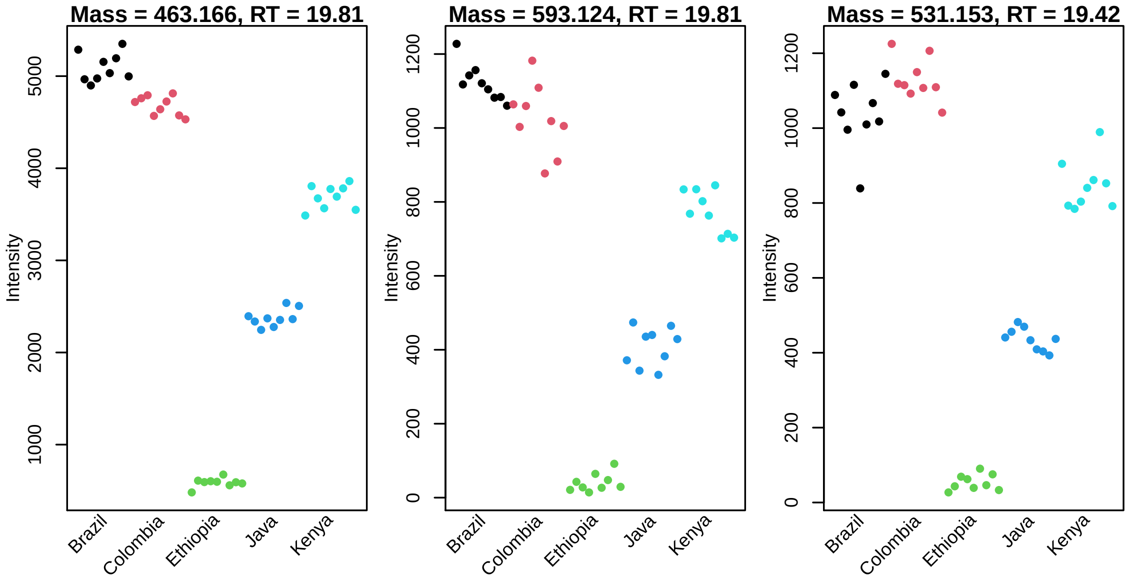


Figure 5.12: Relative intensities for the three variables corresponding to the loadings with greatest absolute values for PC1 and PC2 combined in the scaled analysis. Based on the intensities of the peak it suggests that these might be from the same compound.

to Kenyan and floral and berry flavours near the Ethiopian scores. The centrally located X and Y scores for Colombian coffee support the fact that maximum taste score for this origin is just 2, which it attains for floral, berry and chocolate flavours.

Similar to the taste analysis used with the positive mode data, in order to find the most important variables in the model, VIP scores are calculated, and based on the RMSEP calculated from the PLS-R model, three components are used. A linear regression model was created for each of the six tastes using only the top 40 variables selected by VIP scores. Step-wise regression to minimise the AIC was performed and the p-values for t-tests were obtained from the final models. The fitted values for the models are shown in Figure 5.14 with the significance levels for the p-values presented in Figure 5.15. The magnitude of the residuals from each of the tastes appears to correspond to the RMSEP; chocolate and spice both display relatively low residuals, whereas for floral and nuts they are high. While including more components might improve the fit for these latter two tastes, it would increase the risk of over-fitting the intensities which are already well fit, and so the five component model will still be used. This implies that the components found in the negative-mode contribute more to the taste than those in the positive-mode.

Despite the good fit achieved from the k-fold models for the test, when inspecting the p-values in order to see which features best describe each of the individual tastes only three of them have p-values less than 0.05 (Figure 5.15). For citrus, this is the feature with m/z value 542.227 and retention time of 17.08 minutes and has the

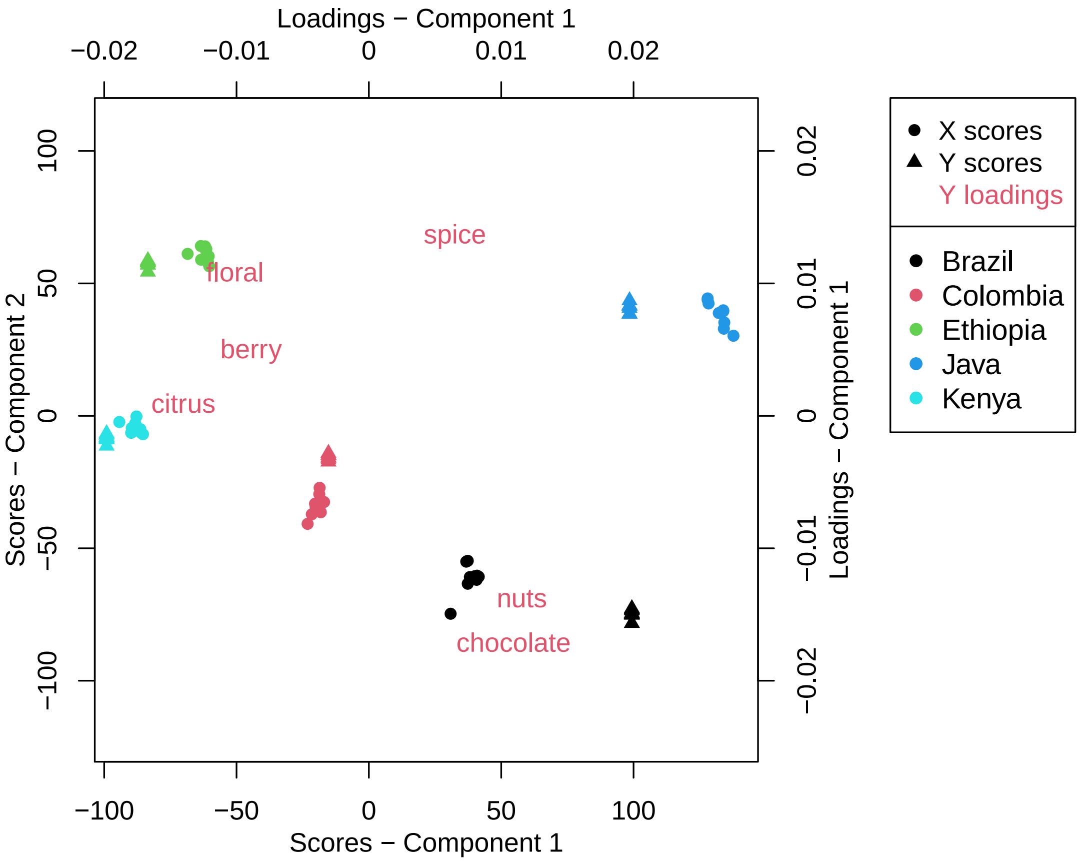


Figure 5.13: Triplot showing both X and Y scores as well as Y loadings for the first two PLS components of the scaled negative-mode PLS data. The X and Y scores are distinguished by symbols whilst colour is based on origin. The relationship between X and Y scores is clear with the Y scores all relatively close to their X counterparts. The Y loadings, shown by the red text, mimic the taste scores provided by experts (Figure 3.2).

lowest p-value. Berry has one feature, with m/z value 533.131 and retention time 14.60 minutes, which is significant at the 99% confidence level, which is significant for the floral taste, however this is at the lower 90% confidence level. Spice has the best fit based on the k-fold models, however there are no features with sufficiently low p-values which are able to be identified.

5.1.2.2 Use of STOCSY to aid identification

Several variables were identified as being discriminatory between the geographical origins of coffee observations and related to taste profiles. Statistical Total Correlation Spectroscopy (STOCSY) was applied to these variables to potentially identify further variables highly correlated with these that could, for example, represent different adducts, and thereby aid compound identification.

The first variable used as a ‘driver peak’ in STOCSY is one of the three key variables identified in PCA, with m/z value of 707.184 and a retention time of

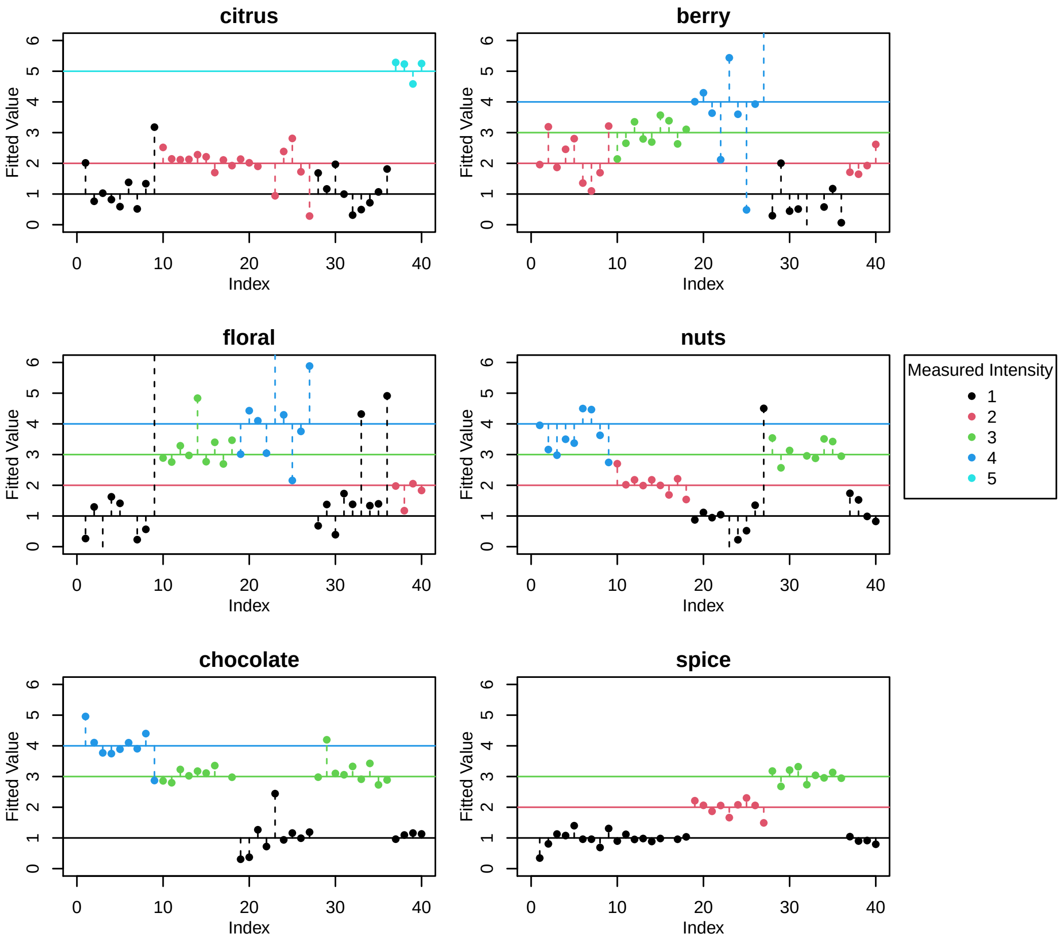


Figure 5.14: The fitted values for models created using stepwise regression with only the 40 variables with highest VIP scores in PLS-R, coloured by the measured taste intensity. Is is is clear that less than 40 variables are sufficient to explain the tastes of each coffee origin.

10.71 minutes. Two of the variables with the largest absolute correlation with this driver peak have the same retention time, and a third elutes just 0.01 minutes later suggesting they may be related compounds (Figure 5.16).

A second driver peak used is the compound with m/z value of 481.285 and a retention time of 13.55 minutes also identified as a key variable for classification of coffee origin. The four variables with the greatest correlation with the driver are shown in Figure 5.17. Again, three of the four variables with the greatest correlation with this driver, have the same retention time but a range of m/z values.

The STOCSY method was also applied using the variable with m/z value 542.227 and retention time 17.08 minutes, identified in stepwise linear regression as having p-value <0.001 for citrus. Figure 5.18(A) shows the covariance this driver has with all other variables, coloured based on the correlation they have with the driver. The relative intensities for the three features with the greatest correlation with the driver

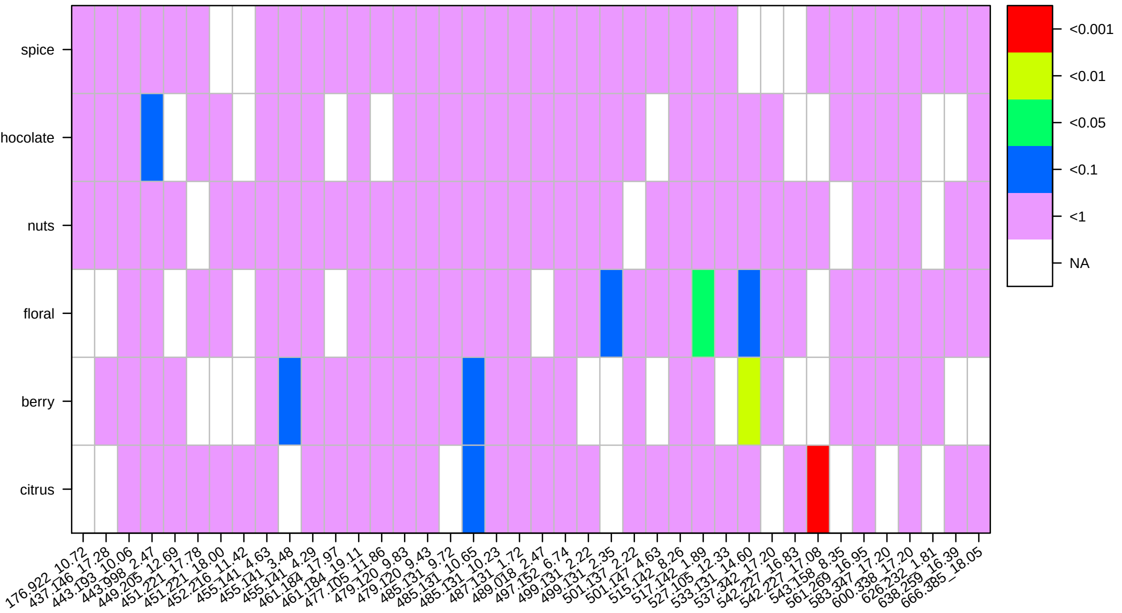


Figure 5.15: Significance levels for variables in linear regression models built using the 40 variables with the highest VIP scores. Here each significance level is represented by a different colour, with red being most significant ($p < 0.001$) and pink the least ($p < 1$). White is used to show variables that are not included in the final step-wise regression model.

are also shown. Based on the high correlation these have with the driver peak, as well as the similar masses suggets that these are adducts of a common compound.

5.1.2.3 Tentative compound identification

Using the STOCSY method together with Massbank, an online database for mass spectrometry reference spectra [141], it is possible to attempt to identify the potential analytes. The difference in m/z values between high correlates of the first driver peak appear to be from oligosaccharides (OS), specifically DP5. The abundance of larger oligosaccharides is known to increase with a darker roasting (i.e. higher temperature and long duration) of the beans [142], and so this is potentially linked. As we do not have roasting information for these beans we can not confirm if this applies here however.

For the driver peak with m/z 481.245 and a retention time of 13.55 minutes, by using the driver properties as well as the masses from the compounds identified as similar via STOCSY and searching via Massbank again this is potentially Apigenin C-glucoside malonylated, a flavonoid previously identified in coffees [143].

Applying the same technique to the remaining unidentified peaks starting with the remaining key feature selected for origin classification with m/z value 727.356 and retention time 17.43 minutes, this was found to potentially be Nicotinamide adenine

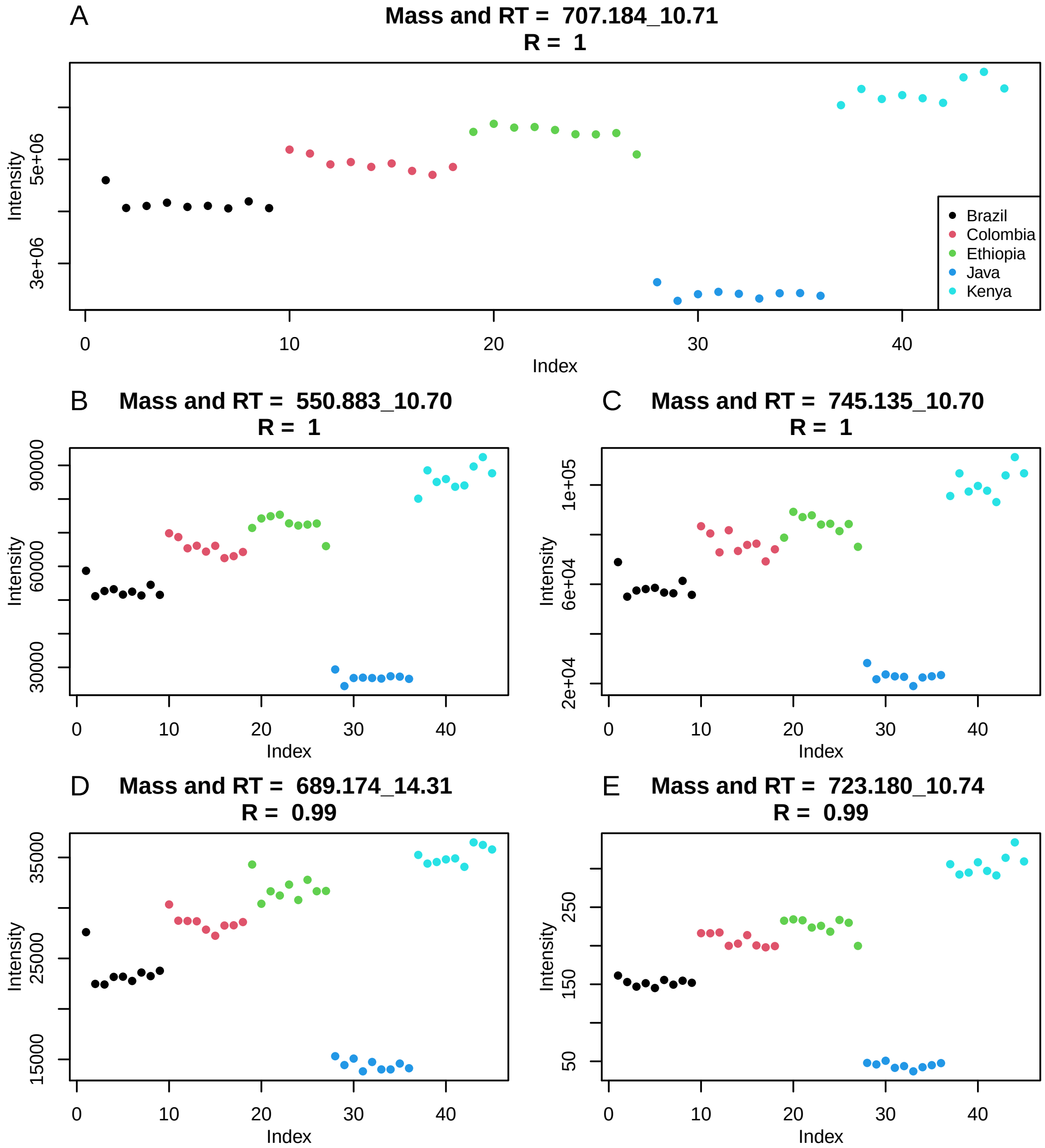


Figure 5.16: Relative intensities for the driver peak (A) and the four variables having the highest correlation with the driver peak. Three variables (B, C and E) all have a very similar retention times to the driver.

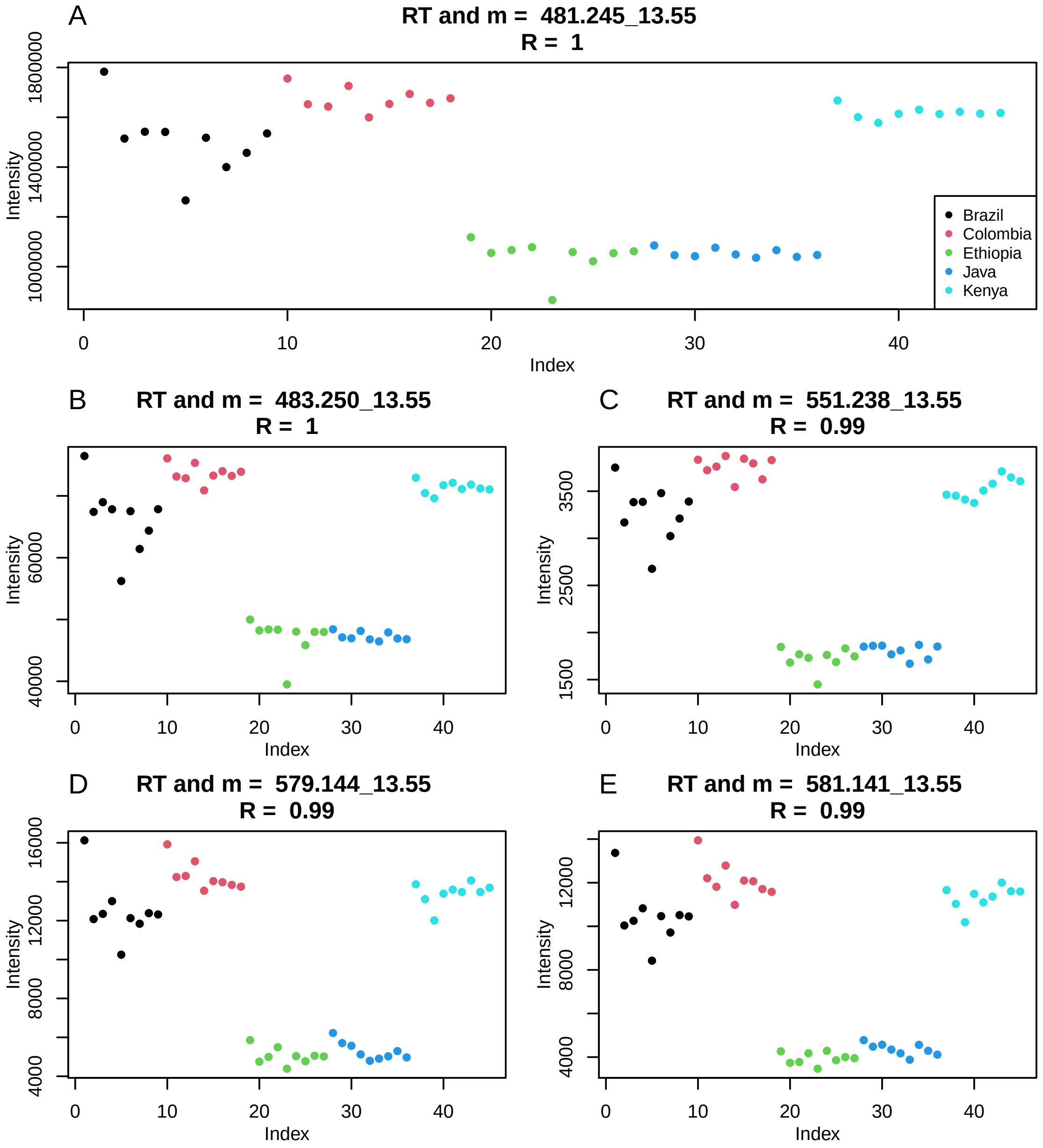


Figure 5.17: Relative intensities for the driver peak (A) and the four variables with the highest correlation with the driver peak. Three of the variables (B, D and E) have the same retention time as the driver.

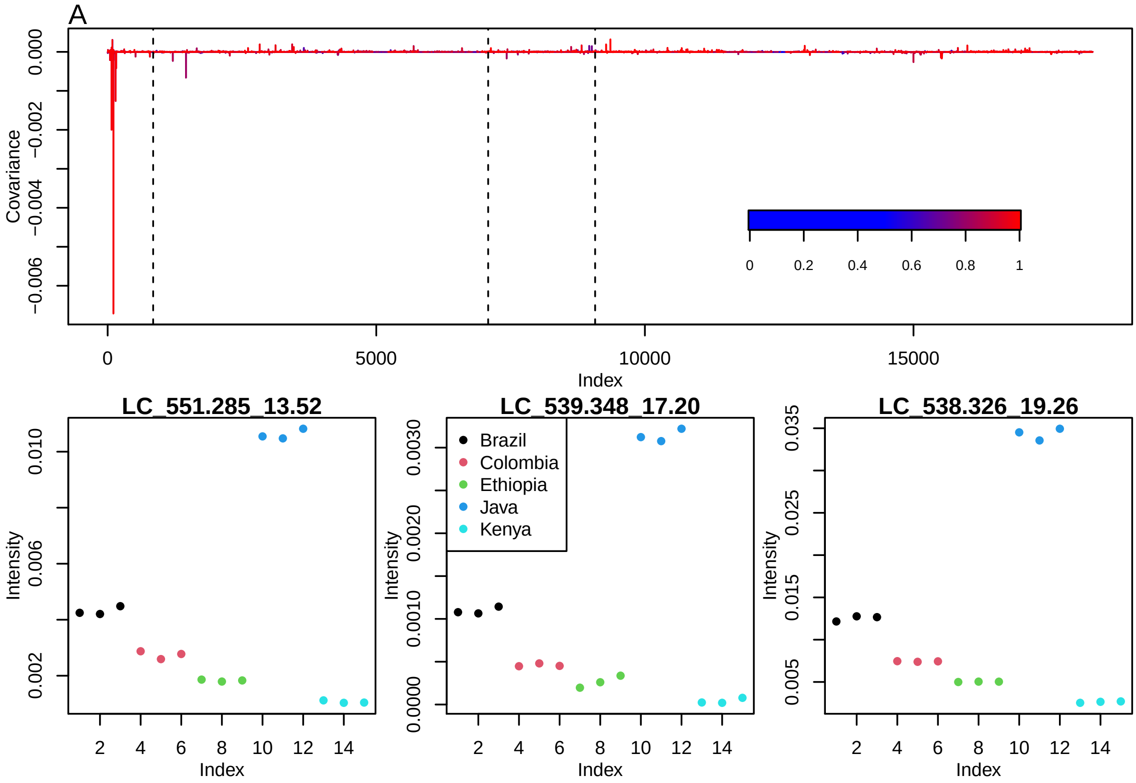


Figure 5.18: Covariance between the driver peak and all other variables, coloured by the corresponding correlation (A), and the relative intensities for the three variables with the highest correlation with the driver peak.

dinucleotide which is formed in roasting of coffee beans [144]. Using Massbank the compound which contributes to the spice flavour with mass 537.342 and retention time 17.20 minutes is also assumed to be Phosphatidylcholine, again found in coffee before [145]

In some cases the STOCSY method is not enough to identify possible compounds, for example the compound with m/z value 455.141 and retention time 4.29 minutes which contributes to nuts, and the same mass at retention time 3.48 which is in the nuts, floral and berry models with low p-values. While these, and other variables are still not identifiable, further methods, such as data fusion, are available and will be discussed in chapter 6.

5.2 Analysis of GC-MS data from coffee

As the retention times and m/z values differ between observations in the GC-MS data set, a 2D binning/interpolation method is used to provide a consistent data matrix for further analysis, and this is performed using bilinear interpolation [146]. This requires a grid formed of m/z values and retention times that encompass the existing range of values, and these ranges contain fixed width intervals determined by the resolution of the grid. Each data point in the raw data observation at retention time t and with a m/z value m, referred to as $x_{t,m}$, contributes to four data points in the new data matrix, with time represented by τ and m/z values by μ.

The interpolation of $x_{t,m}$ onto the four points on the new grid, y, is calculated by

$$\begin{aligned} x_{t,m} = \frac{1}{(\tau_{i+1}-\tau_i)(\mu_{j+1}-\mu_j)} & (y_{\tau_i,\mu_j}(\tau_{i+1}-t)(\mu_{j+1}-m)+ \\ & y_{\tau_{i+1},\mu_j}(t-\tau_i)(\mu_{j+1}-m)+ \\ & y_{\tau_i,\mu_{j+1}}(\tau_{i+1}-t)(m-\mu_j)+ \\ & y_{\tau_{i+1},\mu_{j+1}}(t-\tau_i)(m-\mu_j)), \end{aligned}$$

as illustrated in Figure 5.19. The diagram shows how the data point $x_{t,m}$ contributes to the variables $y_{\tau_i,\mu_j}, y_{\tau_{i+1},\mu_j}, y_{\tau_i,\mu_{j+1}},$ and $y_{\tau_{i+1},\mu_{j+1}}$. For example, with a time resolution of 0.01 minutes and m/z resolution of 0.5 m/z units, a data point at 5.003 minutes and m/z 50.4 would contribute 56% of its intensity to the new data point at 5.00 minutes and 50.5 m/z.

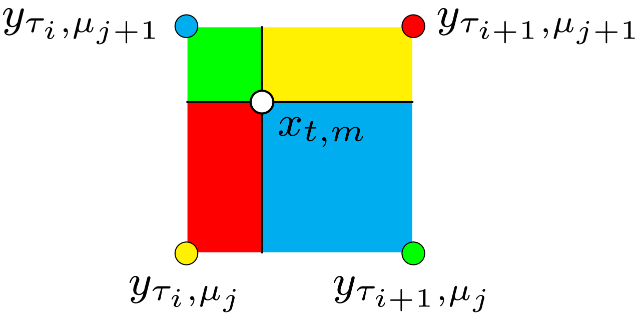


Figure 5.19: Diagram showing the contribution of the data point, $x_{t,m}$, to the new data matrix elements. The area of the coloured squares shows the proportion of the intensity assigned to the correspondingly coloured data points.

5.2.1 Choice of resolution

Clearly, the choice of resolution is vital. If the difference between matrix elements is too large, the resolution of the data could be too low, whereas if is too small, resulting empty data-points could hinder analysis.

In order to determine suitable resolutions for the interpolation various resolutions were compared, making use of the Bhattacharyya distance [147] to compare results. This is a measurement for similarity between two statistical samples, and is calculated as

$$D_B(p,q) = -\ln(BC(p,q)),$$

where $BC(p,q)$ is the Bhattacharyya coefficient, calculated via

$$BC(p,q) = \sum_{x \text{ in } X} \sqrt{p(x)q(x)}$$

for two probability distributions p and q.

The Bhattacharyya distance can also be calculated via the mean and variance of two normal distributions, p and q, as

$$D_B(p,q) = \frac{1}{2} \ln \left(\frac{\sigma_p^2 + \sigma_q^2}{2\sigma_p \sigma_q} \right) + \frac{1}{4} \frac{(\mu_p - \mu_q)^2}{\sigma_p^2 + \sigma_q^2}.$$

To allow direct comparisons to be made between the different resolutions, which obviously result in different sized interpolations, the various dimensions were reduced using PCA with a fixed number of components. As three resolutions are tested for both retention time and m/z values, a total of nine different versions were compared. The retention time intervals tested were 0.003, 0.01 and 0.05 minutes. The first interval was chosen as 0.003 minutes is the average retention time resolution of the GC-MS data for the coffee observations, and therefore provided a lower limit. The other two intervals were chosen arbitrarily to reduce the number of data points in the final data set. The chosen resolutions for m/z values of 0.2, 0.5, and 1.0 m/z units were also chosen arbitrarily. For each data set, three principal components were selected to provide a sufficient proportion of the variance, and the Bhattacharyya distance calculated from these in order to compare the separation of different origins, as this is the major source of variance in the GC-MS data. As separation between groups is used, higher distances are desired.

The distances calculated are shown in Table 5.2, and shown in Figure 5.20. The greatest differences in Bhattacharyya distances are caused by differences in retention time intervals. The highest resolution (0.003 minutes) averages around 80, whereas the coarser data gives a Bhattacharyya distance around 99. Because of this, either of the two coarser time resolutions should be suitable. The Bhattacharyya distances for different m/z intervals do not differ much, so that any of these resolutions should be suitable for this data.

5.2.2 Origin Analysis

For this data medium resolutions were chosen for the interpolation process, and in order to reduce the number of variables the peaks were integrated over to form features. Exploratory analysis of the GC-MS data showed clear separation of coffee origins within the first two principal components (Figure 5.21) for unscaled data, suggesting that this is a key contributor to the variance. One of the Brazilian observations

		m/z spacing		
		1.0	0.5	0.2
	0.05	98.72	98.92	98.97
RT spacing	0.01	97.72	97.96	98.04
	0.003	78.94	79.37	79.82

Table 5.2: Average Bhattacharyya distances for the various time and m/z interval combinations.

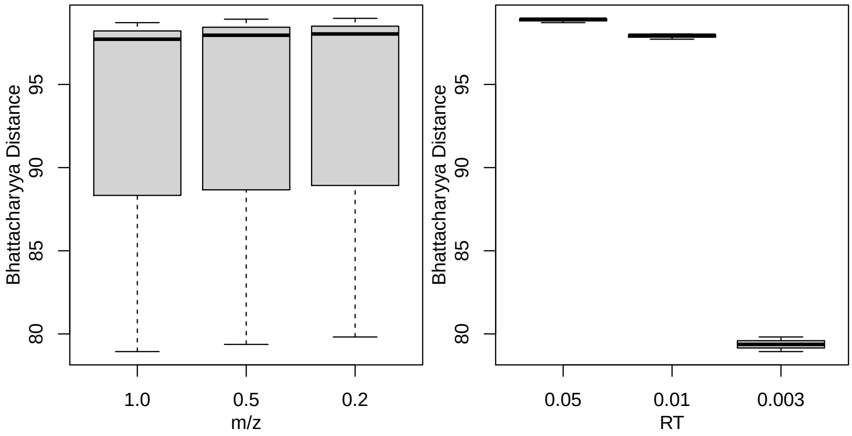


Figure 5.20: Boxplots of the Bhattacharyya distance for different m/z and time intervals.

appears to be an outlier based on the scores plot; most of the observations form a tight cluster however this observation has a significantly higher PC2 score. However, this observation was not removed from the data set. After UV scaling, the separation for some origins decreased considerably, forming three clusters along PC1 with Javan and Ethiopian coffees separated from the other three origins. Due to these results, unscaled data was used for further analysis.

Both decision trees and random forests were used for classification of these data by origin. Two-thirds of the data was used as a balanced training set with a roughly equal proportion from each origin. The remaining data formed an independent test set. A decision tree created from the training set (n=22) achieved 100% accuracy on the training set (Table 5.3) and 54.5% accuracy on the test set (n=11) (Table 5.4). Both Brazilian observations are classified as Kenyan while one Kenyan observation is predicted as Brazilian as is one Colombian. A Kenyan observation is also classified as Colombian. All Ethiopian and Javan observations are correctly classified but the large decrease in accuracy for the test set in comparison to the 100% accuracy for the training set suggests over-fitting.

A random forest was created to determine the variables important for classification, as well as to aggregate the predictions in order to check to see how the accuracy

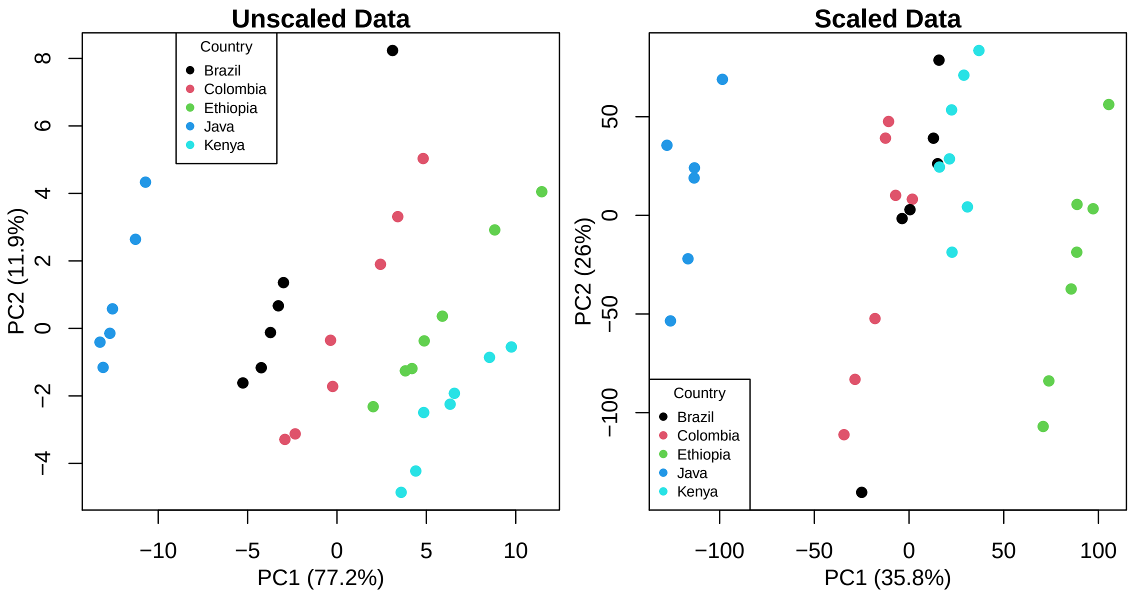


Figure 5.21: Scores for the first two principal components obtained from unscaled and scaled GC-MS data. In unscaled data PC1 contains the majority of the variance (75.2%) and exhibits a clear separation based on the origin whereas scaled data shows notably less separation of origins.

		Reference				
		Brazil	Colombia	Ethiopia	Java	Kenya
Prediction	Brazil	4	0	0	0	0
	Colombia	0	4	0	0	0
	Ethiopia	0	0	5	0	0
	Java	0	0	0	4	0
	Kenya	0	0	0	0	5

Table 5.3: Confusion Matrix of the decision tree model for origin using the GC-MS training set, showing 100% accuracy.

		Reference				
		Brazil	Colombia	Ethiopia	Java	Kenya
Prediction	Brazil	0	1	0	0	1
	Colombia	0	2	0	0	1
	Ethiopia	0	0	2	0	0
	Java	0	0	0	2	0
	Kenya	2	0	0	0	0

Table 5.4: Confusion Matrix of the decision tree model for origin using the GC-MS test set, showing 54.5% accuracy.

changes when multiple trees are used. The average classification of each observation can also be measured to see how the classification changes, and possibly avoid over-fitting. 5000 trees were created using the same training data as used with the decision tree, resulting in a notable decrease in bias. The training set accuracy is still 100% but the accuracy for the test set improved to 90.9% (n = 11) as shown in Table 5.5. The only misclassified observation was an Ethiopian coffee, classified as Brazilian. For each variable the mean decrease in accuracy (MDA) was calculated and the three most important variables by this criterion are shown in Figure 5.22. The first compound shows separation between all origins except Brazilian and Javan whilst the third separates these two origins as well as (to a lesser extent) the other three origins. The second variable shown in Figure 5.22 shows less separation with only Colombian observations clearly separated from most other observations. A combination of just these three is able to achieve 100% accuracy on a decision tree, validated using LOO-CV.

		Reference				
		Brazil	Colombia	Ethiopia	Java	Kenya
Prediction	Brazil	1	1	0	0	0
	Colombia	0	3	0	0	0
	Ethiopia	0	0	2	0	0
	Java	0	0	0	2	0
	Kenya	0	0	0	0	2

Table 5.5: Confusion Matrix of a random forest for the test set of coffee observations showing 91% accuracy.

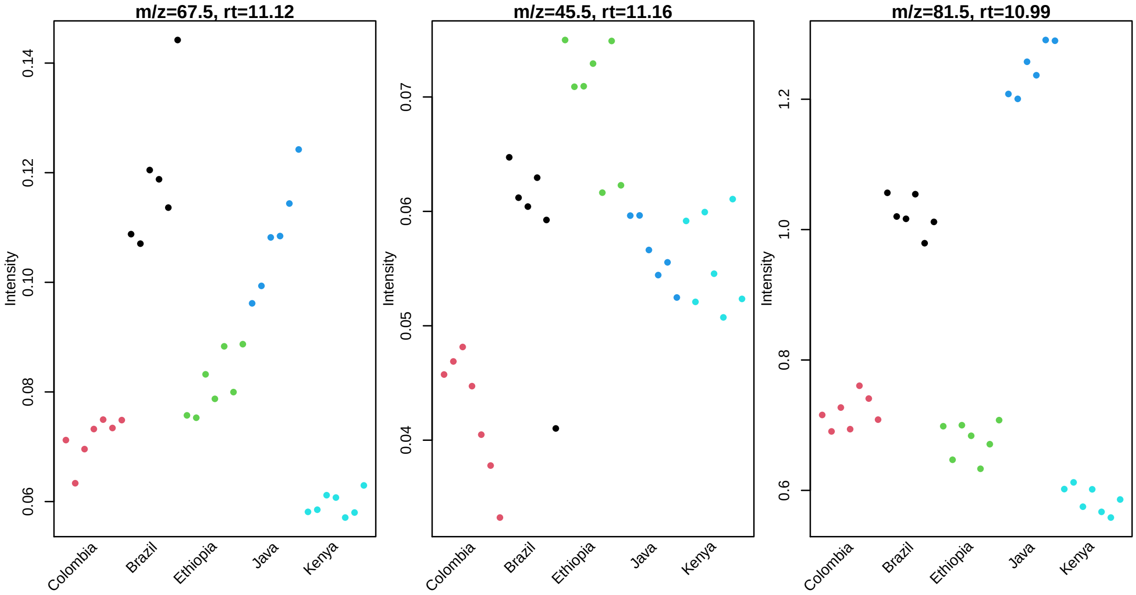


Figure 5.22: Individual intensities for the variables with the greatest mean decrease in accuracy in random forest classification.

5.2.3 Coffee Taste Analysis

PLS-R with taste scores used as responses was applied in the same approach used for the LC-MS analysis. Unscaled and UV-scaled PLS models were compared, and scaled data chosen due to improved clustering in the first two components. Figure 5.23 shows the scores and Y loadings for the scaled GC-MS data. As seen with LC-MS data, most of the Y scores are close to the corresponding X scores, with the exception of Brazilian observations. The loadings for nuts and chocolate are expected to be extremely similar and in fact can be seen to overlap. Again observations cluster along PC1 based on the continent of origin.

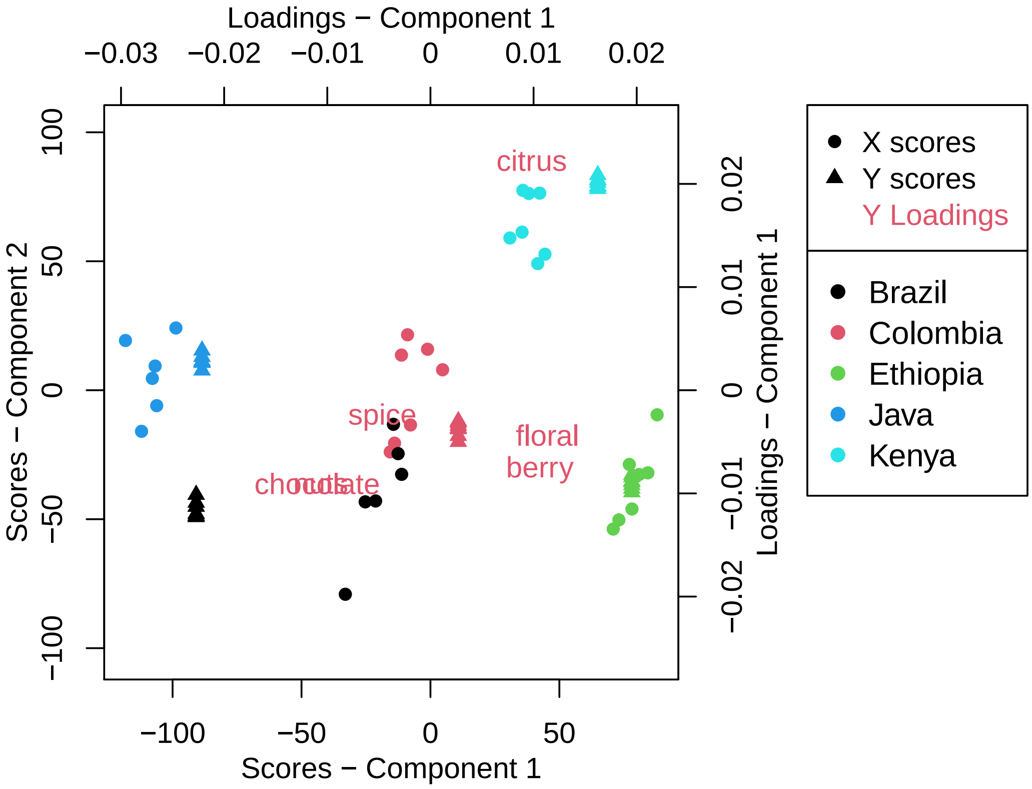


Figure 5.23: PLS scores and Y loadings obtained from unscaled GC-MS data, with the taste intensity as the response.

The loadings using the first five components were combined according to variance to generate VIP scores and the top 25 selected. These variables were used to create linear models for each taste, optimised to minimise AIC. Boot-strapping was applied in order to produce accurate p-values. The p-values presented in Figure 5.24 show the importance of the variables in each taste model. Citrus has four variables which contribute to the intensity, one of which is also a contributor to chocolate taste. Despite their similarities in the PLS Y loadings chocolate and nut do not have any common components based on the step-wise models.

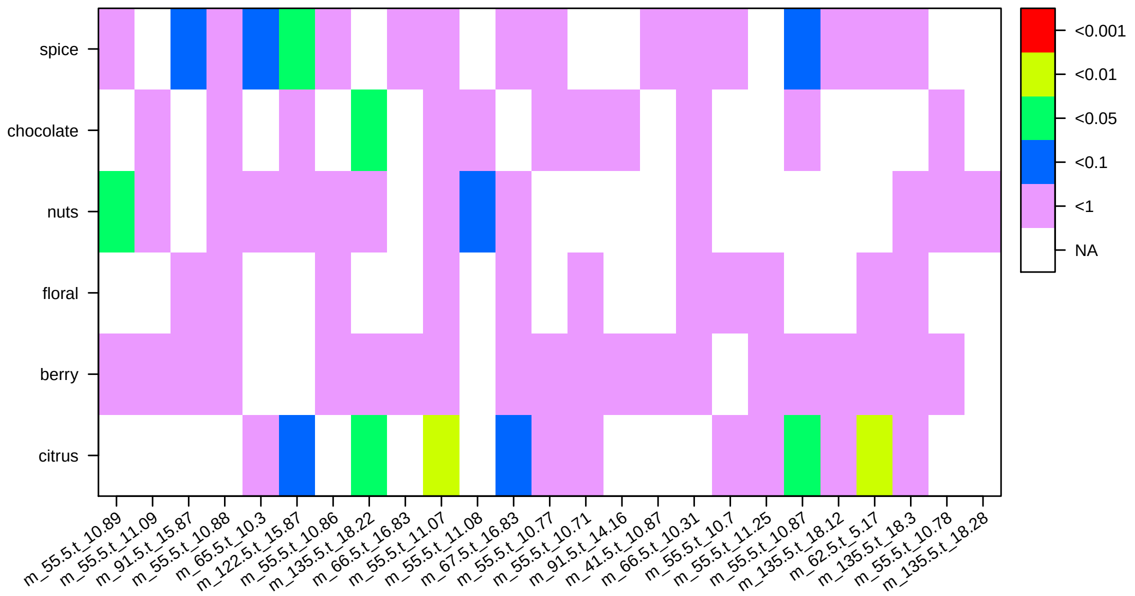


Figure 5.24: Significant levels relating to the p-values from t-tests for linear models for taste. The 25 GC-MS variables with the highest pseudo-VIP scores were used in stepwise regression. Significance levels are represented by colours, with red being the highest level of significance and pink the lowest. White is used to show variables that are not included in the final model.

5.2.4 STOCSY analysis to determine key features

STOCSY was used to determine any compounds related to two of the features identified. Firstly, the feature identified in random forest classification as having the greatest MDA is chosen as the driver. The STOCSY plot in Figure 5.25(A) shows seven features with correlation coefficients above 0.95. All of these have the same m/z value, and are in a close time range (10.78 to 10.89 minutes) and are likely to results from the same compound. The second driver region used in STOCSY has m/z 135.5 and retention time 18.22 minutes. This feature has low p-values for both the citrus and chocolate taste models. For the features which have a correlation coefficient greater than 0.99 24 have the same m/z value, and many of them elute at a similar time.

5.2.5 Tentative compound identification

NIST MS Search database [30] was used to provide potential identifications for the compounds highlighted as important for different coffee flavours. Of the three compounds with the greatest MDAs in random forest classification of coffee origin, the first has m/z value of 67.5, and this is found to likely be a pyrrole known to give off a ‘green’ aroma [148]. The second compound with m/z 45.5 and retention time 11.16 minutes was not able to be identified, whilst the third, with m/z 81.5 and retention time 10.99 minutes could be a furfuryl mercaptan fragment, 1-methylpyrrole.

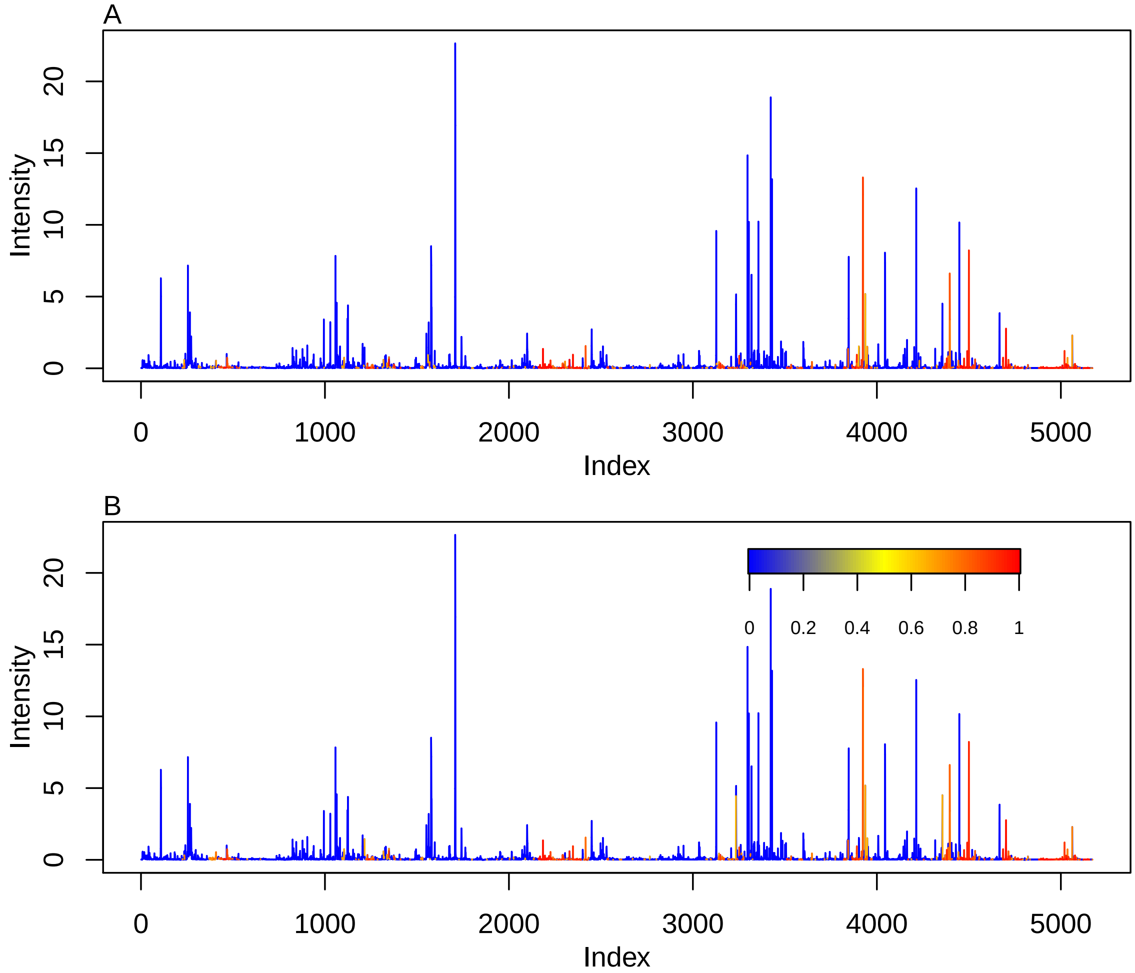


Figure 5.25: STOCSY plots for two features. (A) used the feature with the greatest MDA in the random forest. (B) used the feature with a p-value < 0.001 in five of the six stepwise taste models.

Further testing is however required to verify the identification of these compounds.

5.3 Discussion

These three data sets, both positive and negative mode LC-MS and the GC-MS data, have been analysed using similar methods to the ^{1}H NMR data (Chapter 4.2) in order to find further features of interest for both taste and origin. In order to perform this analysis, two alternative approaches for pre-processing needed to be applied.

For the positive-mode LC-MS data poor experimental preparation led to an issue with solvent levels, and caused significant differences between half of the observations. Rather than using known groups for group standardisation, instead the observations were standardised based on when the measurements were run relating to solvent top-up. This was able to correct the problem, however it resulted in the data being scaled which limited the analysis and hindered with finding some key features. The negative mode data did not have this problem, and so the analysis of this was relatively straightforward.

The GC-MS analysis was straight forward, making use of bilinear interpolation in order to standardise the m/z values and retention times collected for each observation. A random forest for predicting the origin was able to achieve 91% accuracy for these observations. The mean decrease in accuracy was calculated for each variable in the random forest, and the three variables with the greatest MDA were used in a model which was able to predict the origin with 100% accuracy. Analysis of the taste intensity identified a number of potential markers for four of the flavours, however none of these were able to be matched in a database.

Chapter 6

Data Fusion

Within this chapter both high-level and medium-level data fusion have been applied to the coffee data to compare the results from both levels, as well as to try to aid identification of compounds which have not been able to be identified yet. The different level of data fusion is in reference to the processing which has already been applied to the data. For medium-level data fusion the data has undergone feature extraction, for example binning. For high-level analysis these features are analysed, and the features which are found to already contribute to the response is used.

It was not possible to use the Wet Ethiopian, Natural Brazilian or Wet Rwandan observations in this analysis as they were not measured by all methods and therefore observation labels will be based solely on origin and not include the coffee processing method. For the ^{1}H NMR data, only the MSFE Class 1 and Class 2 features were used in the analyses described here. Due to the solvent issue for LC-MS only the negative mode data is used.

6.1 High-level data-fusion analysis

High-level data fusion is the least computationally expensive method to combine data sets, as the features used have already been selected from the analysis of individual data sets. This required analysis however can mean that high-level is the most computationally expensive method, depending on the approach used. This results in analysis which should be able to achieve high accuracies, and correlated features across data sets can provide support for compound identification of important features.

6.1.1 Coffee Origin

6.1.1.1 Feature Extraction

The features discussed in previous chapters were identified via various methods, and so key feature extraction was repeated using a uniform approach for consistency across the data sets. The mean decrease in accuracy (MDA) from random forests was chosen as the metric for selection. For each data set 500 trees were applied, with $\sqrt{N}$ features in each tree as suggested by Breiman [110]. The MDAs were plotted (Figure 6.1) and the threshold to choose the top variables set according to where the scores plateaued. For example, after the top 66 variables in the NMR data set, the decreases in MDA are small and the threshold here provided a balance between the number of variables and classification accuracy. In a similar way, for GC-MS 84 variables were chosen for both the LC-MS and GC-MS data sets.

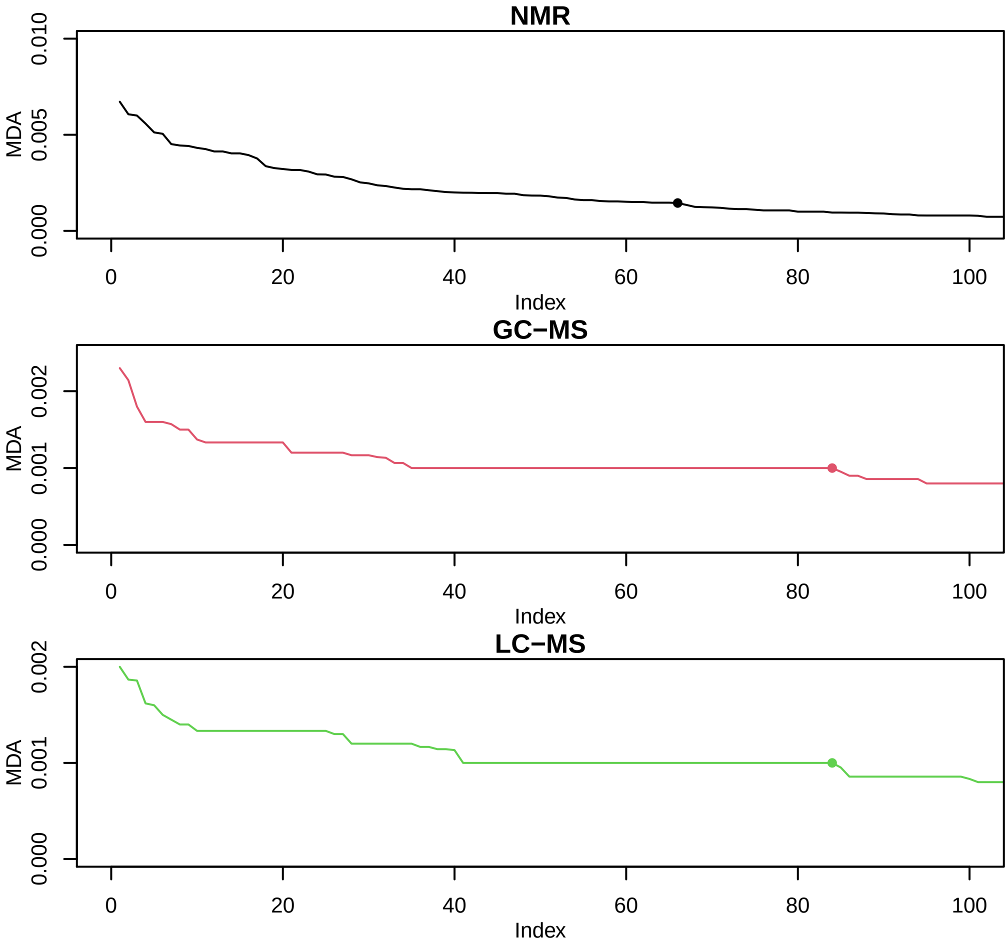


Figure 6.1: Line plots showing the mean decrease in accuracy (MDA) for each of the three data sets. Points represent the threshold for choosing variables for each data set, being 66 for NMR, 84 for LC-MS and 84 for GC-MS. Only the top 100 MDA scores are plotted here to show subtle changes.

6.1.1.2 Correlations between high level features

Before performing the high-level analysis, pairwise correlations between features within the data sets were inspected. The heatmap in Figure 6.2 shows the correlations for the NMR and LC-MS features. The majority of correlations here are negative, as shown by the blue in the bottom left and top right. The dendrograms also highlight the similarities within each data set, with many of the features having similar correlations between the data sets.

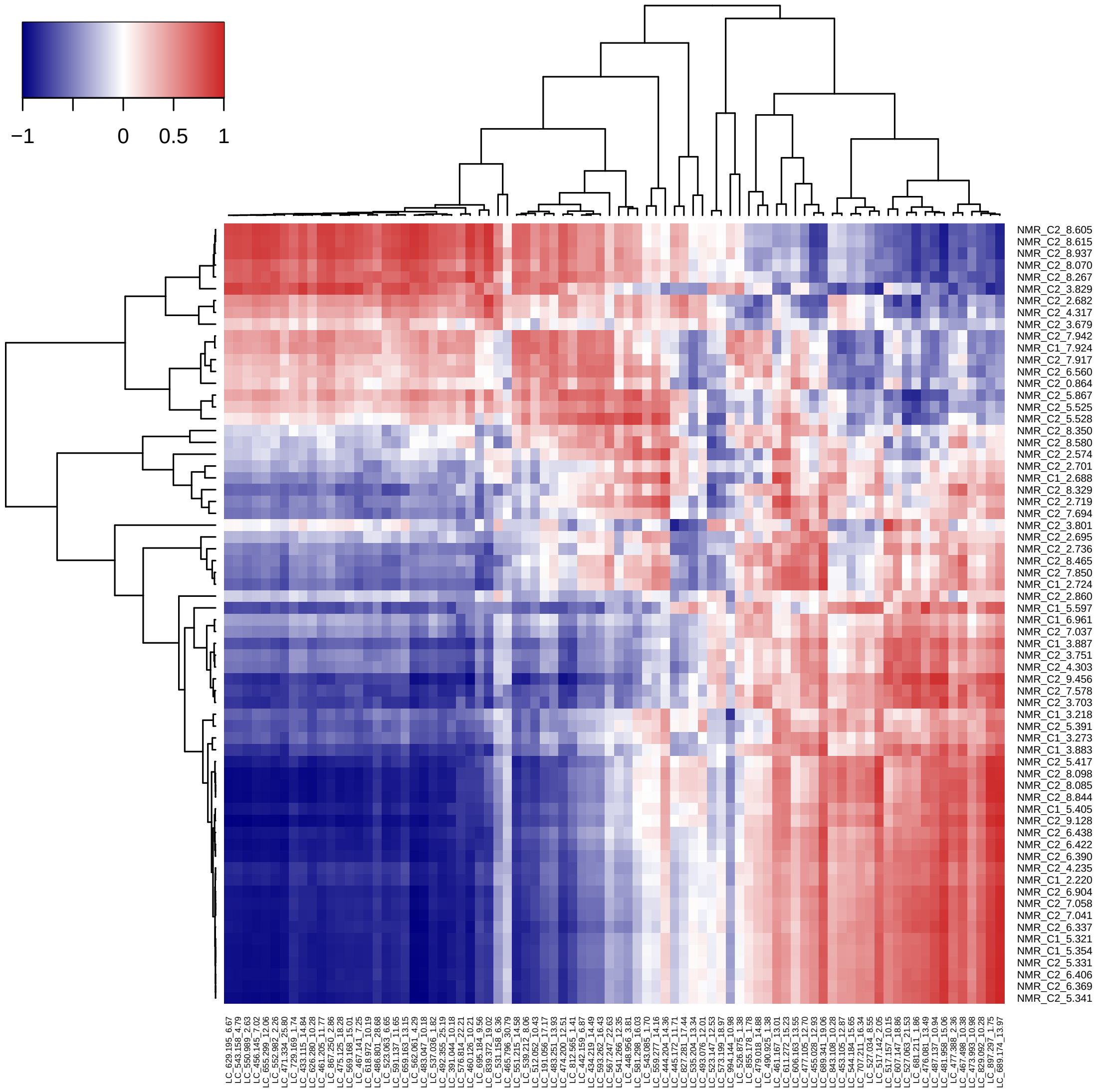


Figure 6.2: Heatmap showing the correlation between the ^{1}H NMR and LC-MS features.

A similar pattern can be seen for GC-MS and LC-MS correlations (Figure 6.3). Many of the correlations between the LC-MS and GC-MS features are weak, which could suggest that they complement each other. However, the dendrogram for the GC-MS features does show that there are many similar features within this data set and the LC-MS data set also has one very large cluster of similar features which could result in less information gained.

The NMR and GC-MS show similar trends as depicted by the last two pairings, and so the plot is omitted. Overall this correlation analysis suggest that the three sets of data can complement each other, because of a large number of weak correlations, and so data fusion is likely to perform well.

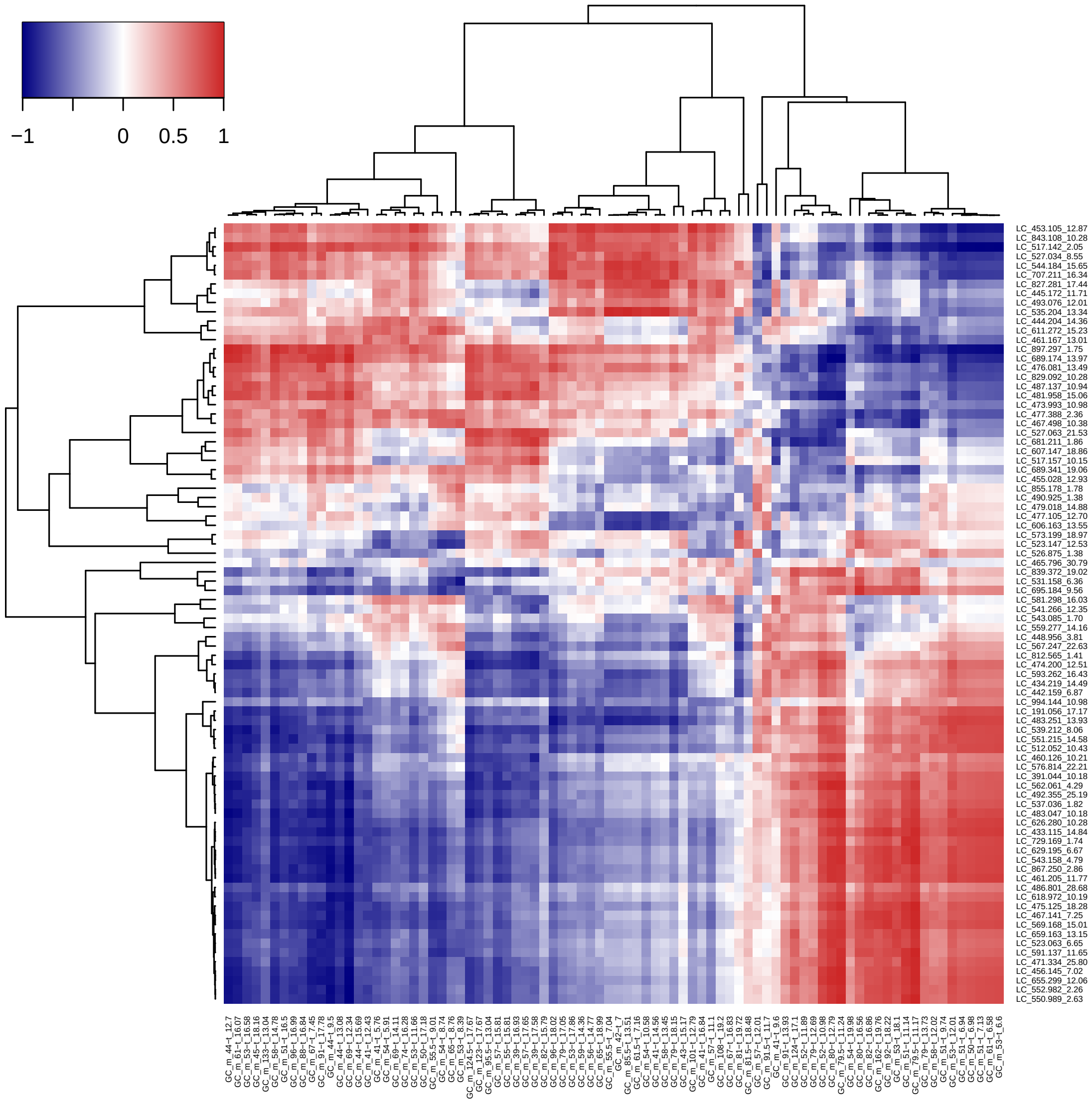


Figure 6.3: Heatmap showing the correlation between the LC-MS and GC-MS features.

6.1.1.3 High-level fused modelling

The use of a random forest classifier with the combined features allowed comparisons of the MDAs before and after fusion to be made. It also allows for the features which act as drivers over all three data sets to be identified and compared to, as well as to see which data set contains these features. 500 trees with $\sqrt{N}$ of the features in each were created, with origin used as the response. The results were validated using LOO-CV and showed 100% accuracy. This was to be expected as high accuracies were achieved in individual data set analyses, and high-level data fusion makes use of the key features from individual analysis modelling.

Figure 6.4 shows the MDAs calculated from the random forest classifier using the chosen variables from all three data sets. The average MDAs for each data set (horizontal dashed lines) show that, although all are comparable, the LC-MS features have a greater average MDA, followed by GC-MS. The feature with the greatest MDA is a LC-MS feature, with m/z 515.142 and a retention time of 8.97 minutes. Comparison of the MDA scores from the individual models with the combined forest shows that many NMR features appear less important in the model, whereas for both the GC-MS and LC-MS many features appear more important. There is no noticeable correlation between MDAs in individual and combined analyses.

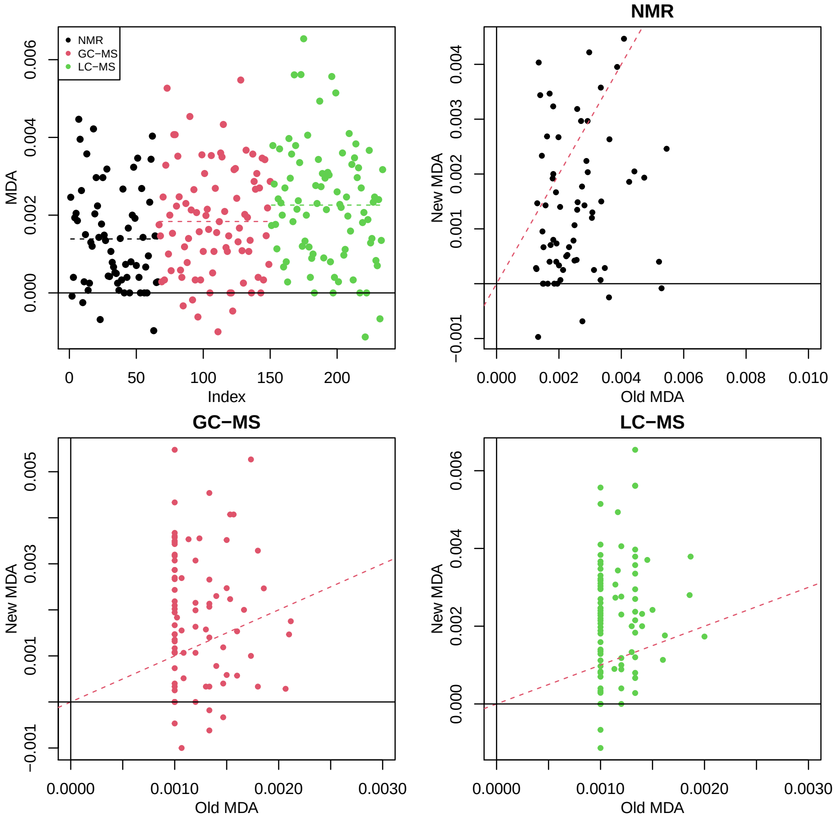


Figure 6.4: The mean decrease in accuracy (MDA) scores for each feature used in the high-level data fusion classifier calculated from 500 trees. The first plot shows the new MDA scores coloured by data source, with the average for each shown by the horizontal dashed lines. The remaining three plots compare MDA scores from the individual random forest classifiers (x-axis) with the new scores in the combined analysis (y-axis). The red diagonal line shows $y = x$, to allow any improvement to be assessed.

6.1.2 Taste profiles

In order to compare the contributions to taste the three data sets were fused. The features to be used in the high level analysis with taste scores as responses model were chosen according to VIP scores in the models used in individual analyses. For each data set the appropriate number of components was chosen based on the RMSEP, and 30 features with the greatest VIP scores were extracted.

The features were UV-scaled and combined to provide a PLS-DA model with the taste scores as responses. Figure 6.5 shows the scores and Y-loadings for the first two components obtained for this model. Within the scores there is clear clustering occurring based on origin, with the Ethiopian and Brazilian coffees separating in the first latent variable, and Java and Kenya in the second. Despite the lower intensity for all of the origins, the Y-scores resemble the X-scores suggesting a good approximation and also that just two components are adequate to model tastes. In fact, the RMSEP suggests that only floral and spice require further components to be able to model this, the rest having sufficiently low RMSEPs after two. This is a notable decrease as was seen with the individual analysis, with the NMR model requiring eight components, the five a required for the positive mode LC-MS, three for negative mode LC-MS, and five for the GC-MS analysis.

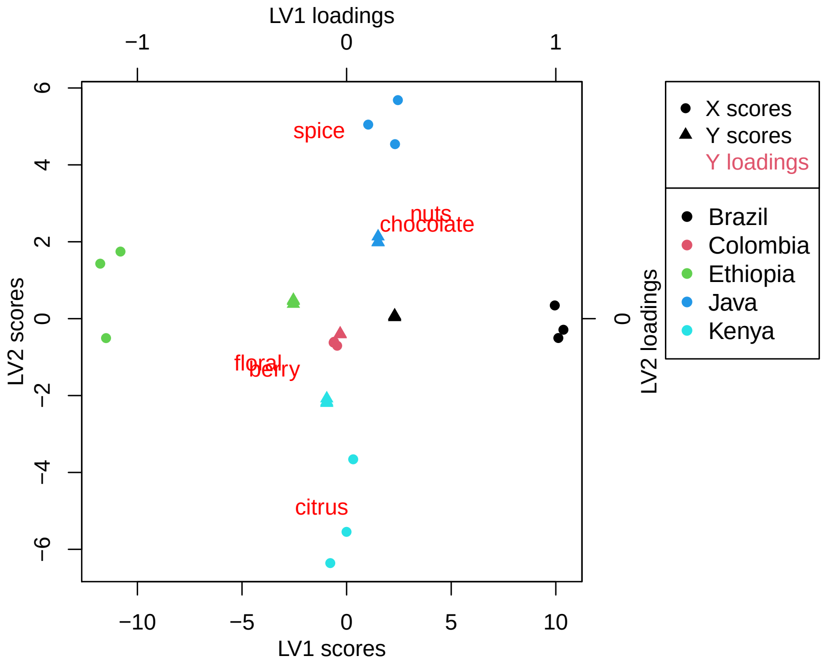


Figure 6.5: Triplot showing the scores and Y-loadings for the first two components from a PLS-DA model for the high-level fused coffee data with taste scores as responses.

The VIP scores for the PLS-DA model obtained for the high-level fused coffee data were calculated using three latent variables. As only 15 observations are available, this should avoid over-fitting, while also avoiding under-fitting. Inspection of the VIP scores (Figure 6.6), showed an NMR feature (Class 2, 9.456 ppm) had the highest score, with nine of the top 10 also being NMR Class 2 and seen previously in the NMR analysis for taste. The GC-MS feature at m/z 55.5 and retention time 11.09 minutes (previously seen in the GC-MS analysis for taste), has the next greatest VIP score. The VIP scores for the LC-MS features are all approximately the same, and this data set has the lowest average score of all three.

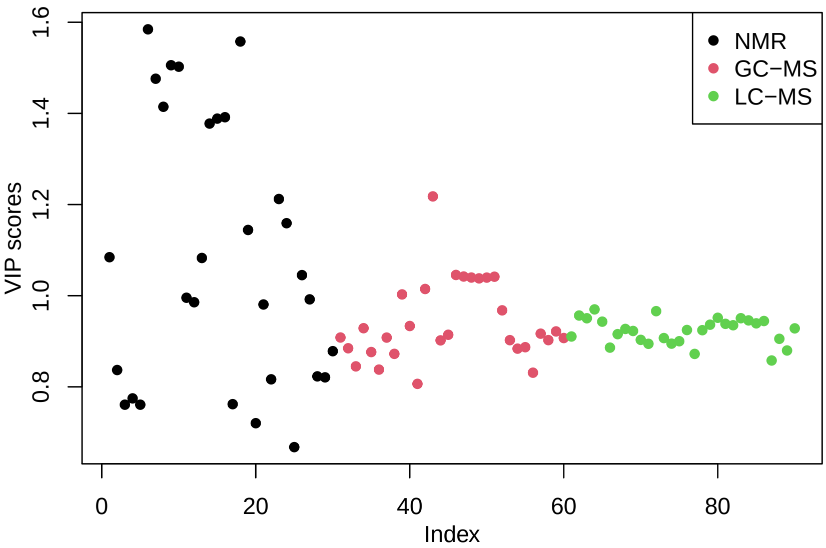


Figure 6.6: VIP scores from the PLS-LDA model for the high-level features and taste as the response.

6.1.3 Results found from high-level data fusion

High-level analysis uses features already known to perform well in modelling the response, and this analysis showed that when combined they still performed well. By fusing the data and comparing the MDAs created from a random forest modelling the origin, the LC-MS features were found to have the greatest MDAs om the resulting model, which suggests that they are the most appropriate for modelling the origin. In the individual analysis LC-MS data also had the greatest accuracy for predicting origin, and so this suggests that LC-MS data is the most suitable set of data to use when modelling coffee origin.

When modelling the taste intensities the NMR data was found to be the most important, and this corresponds with the results also seen in the individual analysis, with the NMR having the lowest average residuals from the step-wise models. New features were identified using data fusion to aid with the profiling of the taste intensity, such as the Class 2 NMR feature found at 9.456 ppm, however many of the features had already been identified in the individual analysis.

6.2 Medium-level data fusion analysis

To compare the results from different levels of data fusion, mid-level data fusion techniques were used. Here data sets of extracted features are combined, regardless of if they contribute to any analysis. All three coffee data sets were already pre-processed and no further processing was necessary prior to starting mid-level analysis.

6.2.1 Medium-level data fusion relating to geographical origin

6.2.1.1 PCA scores data fusion

Mid-level data fusion analysis of the coffee samples was carried out using the approach of Spiteri et. al. [94]. For each of the three data sets, the number of principal components to include was chosen to account for at least 95% of the variance. Using this criteria six principal components were needed for NMR, three for GC-MS, and two for LC-MS. These eleven components were then standardised and combined to form a new data block after multiplying the components by the amount of variance each component contributed. In order to visualise this new data block PCA was performed and the scores from the first four components in this fused model (Figure 6.7) show that this has lower separation than seen for the individual data sets. The variables with the greatest loadings for PC1 are the first principal components from GC-MS and PC2 from LC-MS, and for PC2 the GC-MS second component has the greatest value. Considering these loadings it is surprising that the separation between origins has decreased. For both MS data sets the separation along PC1 is similar, however the GC-MS shows no separation along PC2 (Figure 5.8 and 5.21), which could explain why the clustering is not as tightly grouped here. With only 12 components in total this still requires 5 components to account for 95% of the total variance showing that the components appear to not be complementing each other. If scaled data is used instead for the individual PCAs approximately 10 components are required from each data set, and the combined PCA scores plots show no improvement in separation within the first four components.

6.2.1.2 PLS-DA medium-level data fusion

Supervised medium-level analysis was carried out using PLS-DA with the three UV-scaled data sets combined and the coffee origin as the response. The X and Y scores from this PLS-DA model (Figure 6.8) show clear clustering with origin for both X and Y-scores. Within the second latent variable there is clear separation resulting from the Javan observations, with all other origins having a similar score. Due to the clear cluster which is present in the first to LVs it suggests that these

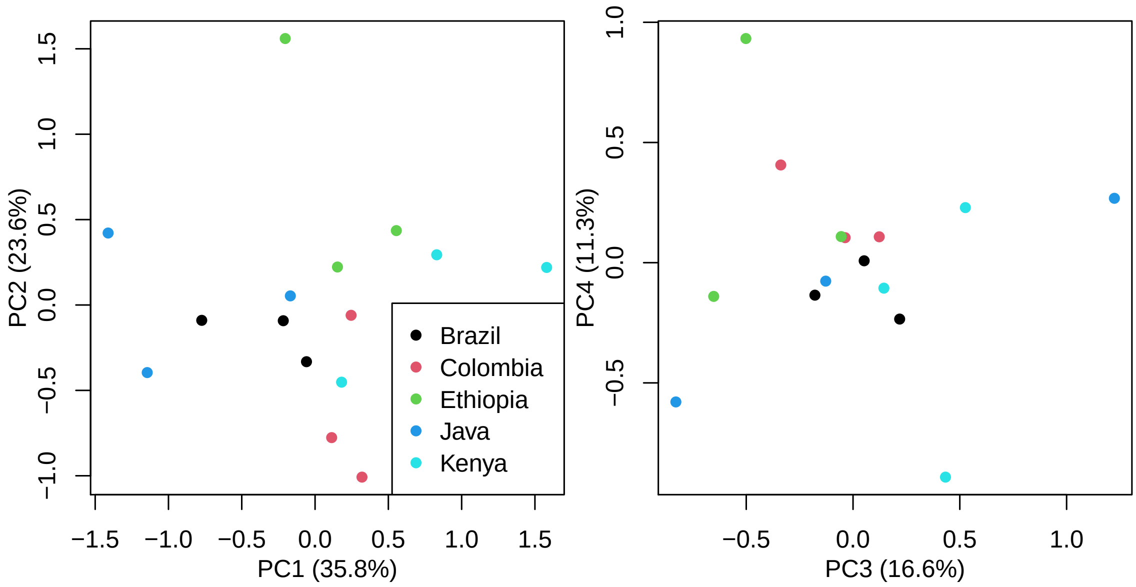


Figure 6.7: PCA scores plots showing the first four components from the mid-level data fusion to combining PC scores from individual data block analyses. Observations are coloured by origin.

alone should be able to model the origins. Inspection of the RMSEP also supports this, with it plateauing after 2 components. This supervised technique shows a clear improvement to the unsupervised PCA approach.

The two-component PLS-LDA model was used with LOO-CV and achieved 100% accuracy, as could be expected from the separation seen in the scores plot. VIP scores were calculated from the model using these two components, presented in Figure 6.9. Due to the scaling necessary for this analysis no singular feature stands out, however it is still clear to see that the NMR features are less important for discrimination between origins than features from either of the two MS data sets.

6.2.2 Analysis at mid-level data fusion relating to taste

The three UV-scaled data sets combined were also used in PLS-DA with tastes scores as responses. Figure 6.10 shows the X and Y-scores, as well as the Y-loadings for the first two components of the PLS-DA model. The clustering is significantly more tightly grouped than seen in the high level data fusion PLS-DA (Figure 6.5), with the exception of the Brazilian observations. There are clear similarities between the X and Y scores for the first two latent variables, and for the Y-loadings these cluster somewhat well based on the measured intensity. For citrus the Y-loading is similar to the scores for Colombia, despite its low taste intensity (2). As such this suggests that a greater number of LVs are required to be able to predict this taste, and the RMSEP calculated for each taste suggested that four components would be required

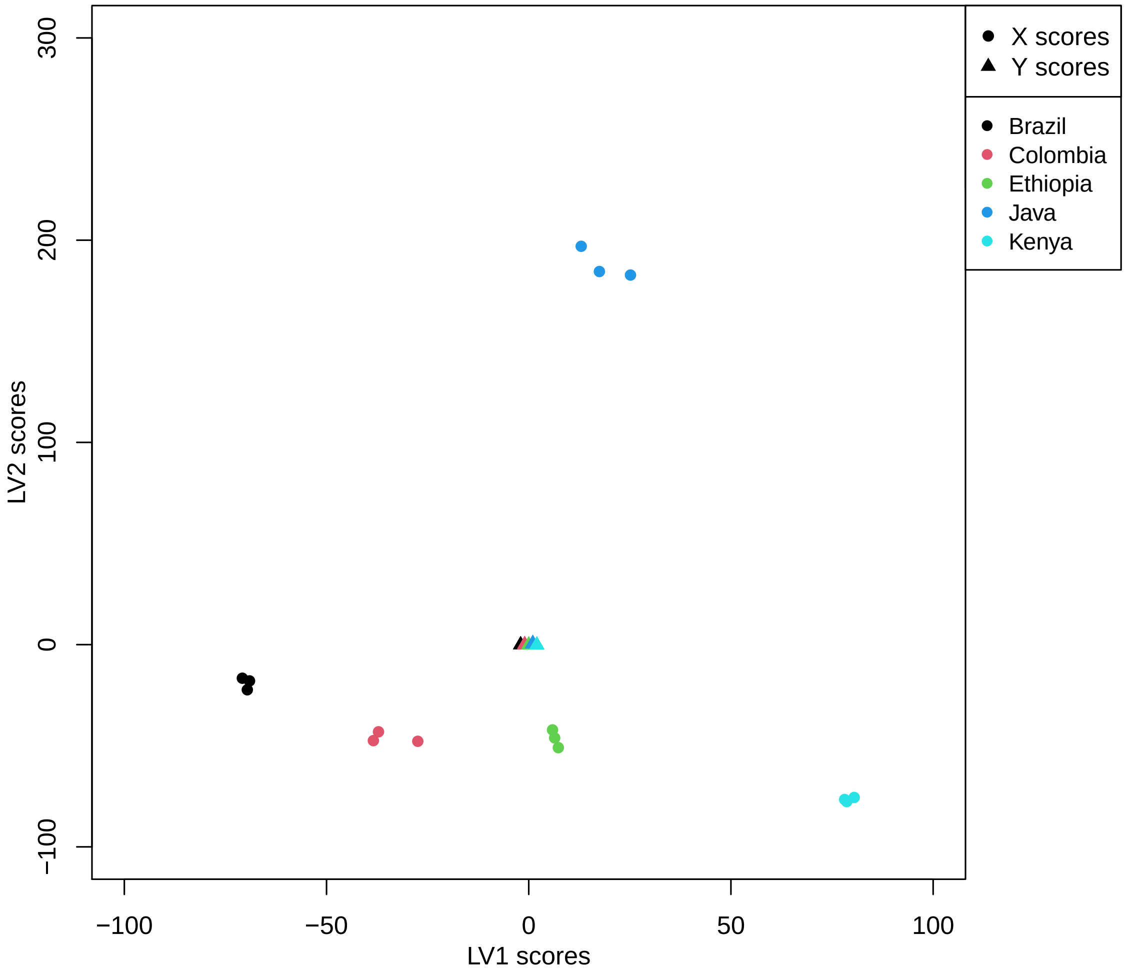


Figure 6.8: *X* and *Y* scores for the first two latent variables from a PLS-DA model using mid-level data fusion and origin as the response. Clear clustering can be seen between the two sets of scores.

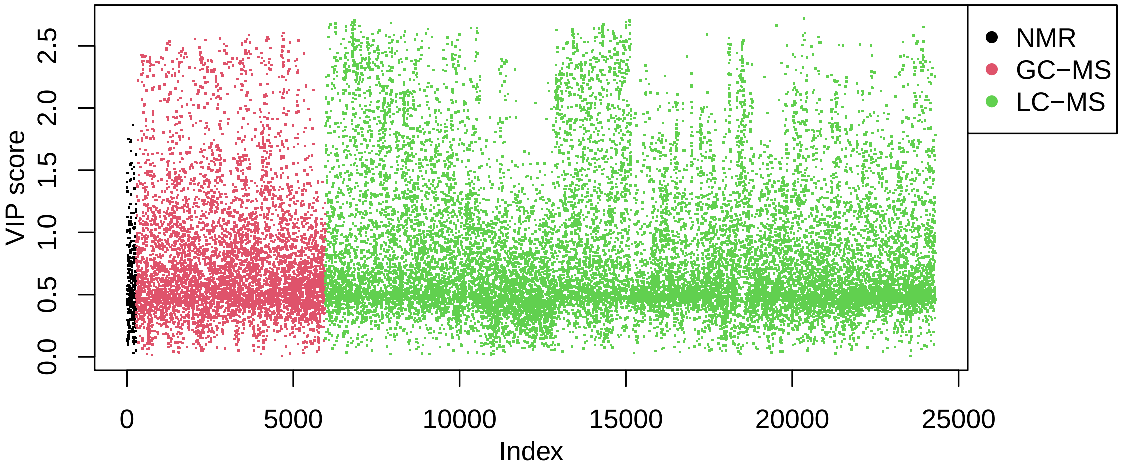


Figure 6.9: VIP scores calculated from the PLS-DA model for the mid-level data fusion analysis with origin as the response. Scores are coloured by data source.

to model all taste scores.

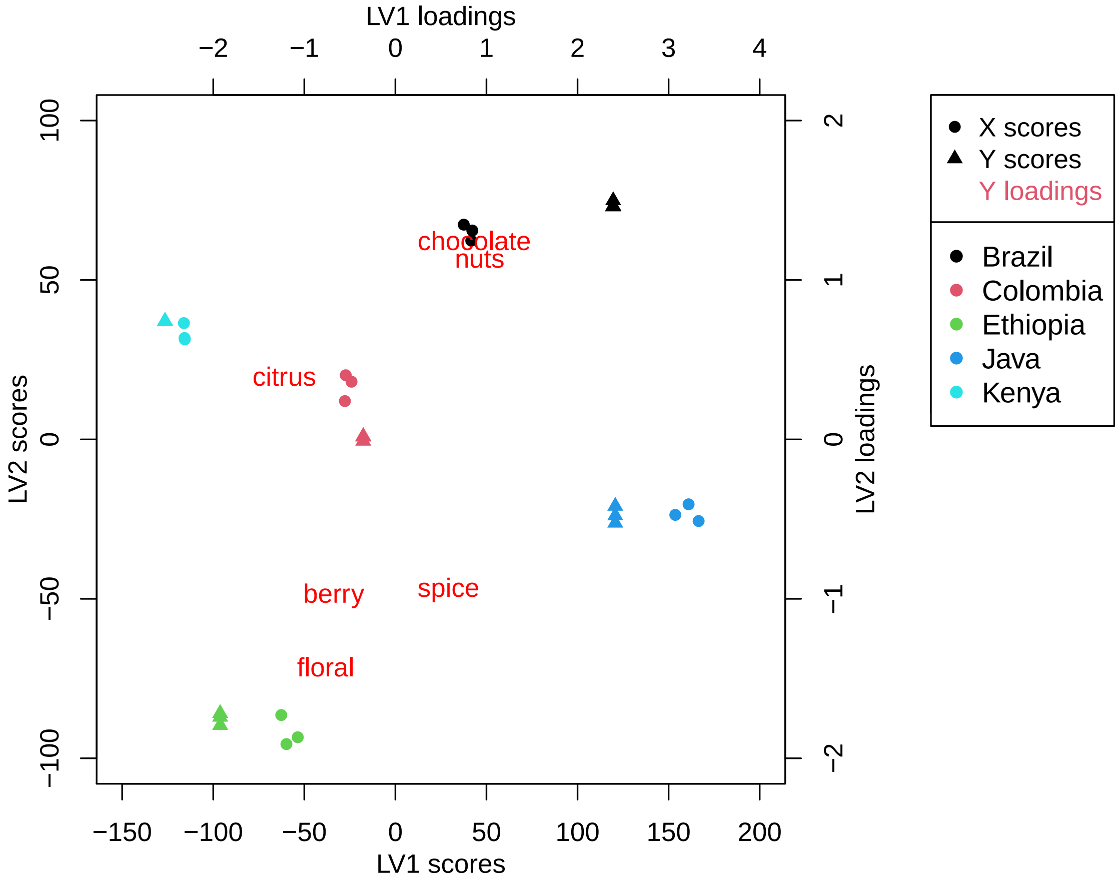


Figure 6.10: Triplot showing the scores and Y-loadings for the first two latent variables the PLS-DA for the mid-level data fusion analysis with taste scores as responses.

VIP scores for a four-component model show that the most discriminatory variables are all from LC-MS. The majority of the variables here were also found in the individual model for LC-MS and so no further analysis was performed.

6.2.3 Statistical Heterospectrosopy

Performing SHY on the NMR and LC-MS features with a correlation threshold of 0.95 showed that there were few such correlations. Figure 6.11 shows a region of the heatmap, with only those features that have an absolute correlation above the threshold. As well as needed to have a sufficiently high correlation between the features, the null hypothesis that there is no relationship between the observed features is tested, and p-values are calculated for this. If the p-value is greater than 0.001, then the correlation is also ignored. The most noticeable observation is that a significant number of LC-MS variables are all correlated with a singular NMR Class 2 feature at 9.128 ppm, the majority of these being negatively correlated. For the NMR features around 6.4 ppm, which contains 6 peaks discussed previously in

Chapter 4.3.3.1, there are a number of low intensity LC-MS features with strong positive correlation. These features could potentially be derived from a reaction that occurs, such as the breakdown of a compound, and a combination of the NMR ppm and the parameters from LC-MS are used in database searches at the end of this chapter.

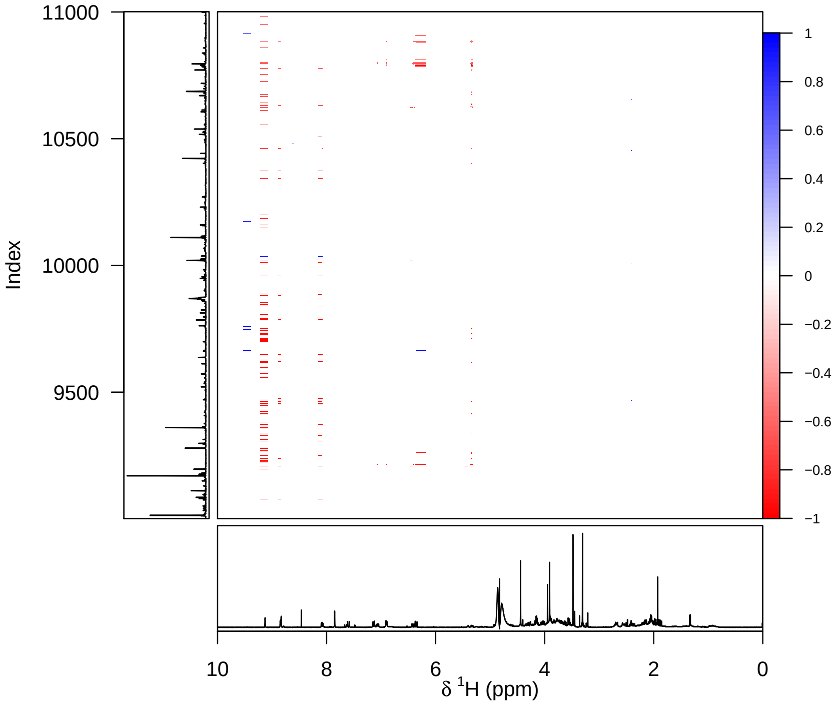


Figure 6.11: Heatmap showing a section of the greatest correlations between the NMR (x-axis) and LC-MS (y-axis) features, calculated via SHY. The corresponding region of the spectra are also shown.

Using the two chromatography methods within SHY reveals a greater proportion of correlated features than seen with the NMR / LC-MS analysis, using the same thresholds. The majority of the correlations are strong positive correlations, which suggest that the compounds could be relating to fragments of a large common compound, rather than derivatives of a reaction. For many of the compounds there are multiple correlated features across the data sets, and so when combining the parameters for compound identification features with a greater coefficient should be preferred.

Finally, the NMR and GC-MS features were compared. Of the correlations found via SHY many of these are negatively correlated, and are found between 8 and 9 ppm in the NMR spectrum. Inspection of the GC-MS m/z values and retention times reveals that many of these have a m/z value of 57.5, suggesting that they could all be from the related compounds.

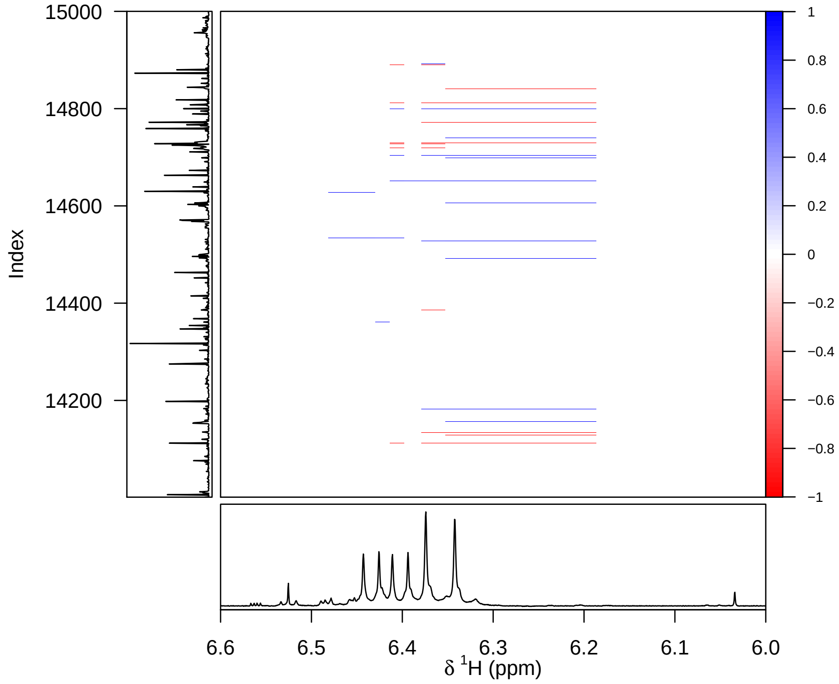


Figure 6.12: Heatmap showing a section of the greatest correlations between the NMR (x-axis) and LC-MS (y-axis) features, calculated via SHY. It has been expanded to show the positively correlated regions around 6.4ppm in the ^{1}H NMR. The corresponding region of the spectra are also shown.

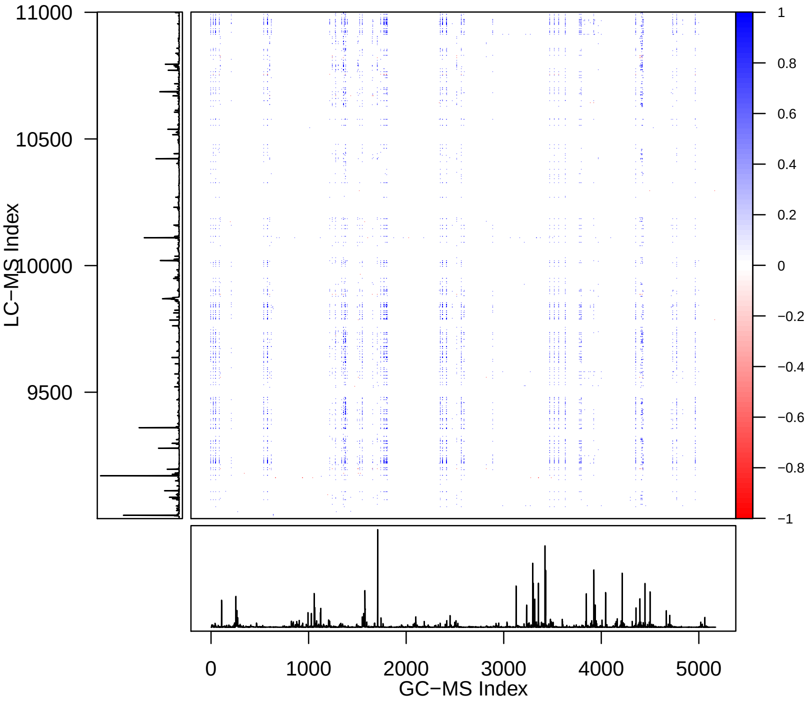


Figure 6.13: Heatmap showing the full GC-MS spectrum (x-axis) and a section of the LC-MS (y-axis), with correlations calculated via SHY. The corresponding region of the spectra are also shown.

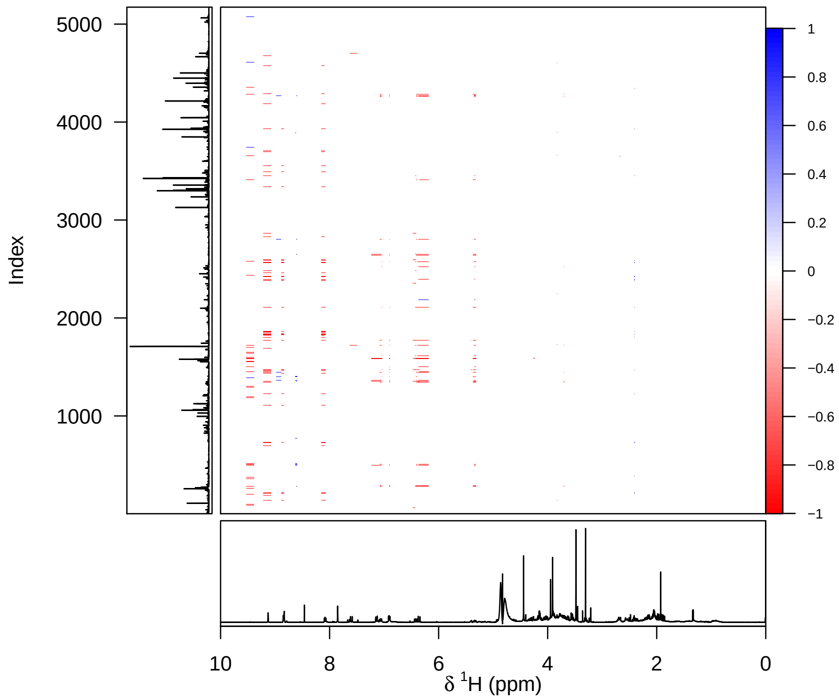


Figure 6.14: Heatmap showing the correlations between the NMR and GC-MS key features, calculated via SHY.

6.3 Comparison between high and medium level data fusion

For the high and medium level data fusion analyses both of these were able to achieve improved classification rates for the geographical origin of the 15 coffees, both attaining 100% accuracy via LOO-CV via different methods of analysis. In the individual analysis of the three data sets, only the negative-mode LC-MS data was able to achieve this. For the high-level modelling, this was reflected in the MDAs, with the LC-MS features having the highest average MDA. By using a small subset of the features from each for the taste, chosen based on their VIP score, the high-level data-fusion was able to also model the tastes better. Via PLS-DA only two of the latent variables are needed to be able to suitable predict the tastes for four of them, whereas the medium-level model required four for each taste. The clustering of the X and Y scores was more tightly grouped than seen in the medium level model within the first two LVs. The VIP scores for the PLS-DA model on taste also suggest different data sets as the most important. The NMR features in the high-level model generate the highest VIP scores, whereas the mid-level LC-MS features had the greatest. There is the potential for this to be related to there being a greater number of features present in the LC-MS data set at medium-level analysis than both the NMR and GC-MS sets. As the NMR VIP scores are notably lower on average this is unlikely to be the case. In the high-level analysis as each data set was restricted to

30 variables this does not apply there.

The medium-level analysis does have the benefit of not limiting the features when it comes to correlation. At the expense of greater computational time as well as verification, this level can find a greater number of features which are correlated with any features of interest.

6.4 Tentative compound identification

Using the information gained from data fusion of the three UV-scaled data sets combined and database searches, tentative compound identification was carried out.

High level data fusion revealed the feature with the greatest VIP score as the NMR feature at 8.326 ppm, which is potentially 2,4-dimethylpyridine previously found by GC-MS analysis of coffee and is associated with roasting [149]. Also in the high level fusion, the feature with the greatest MDA for origin was the LC-MS feature with m/z 515.142 and RT 8.97 minutes. This could possibly be 1,5-Dicaffeoylquinic acid, a polyphenolic compound which occurs in plants such as fennel and coffee.

6.5 Driver-based Statistical Heterospectroscopy

While methods such as Statistical Heterospectroscopy (SHY) can only be applied with two data sets, it is possible that a greater number of data sources are available. In this section a method is proposed which can be applied to any number of data sets in order to find similarities between data points. The proposed method, termed Driver-based Statistical Heterospectroscopy (DRY), works in a similar way to SHY in that correlation based analysis is applied but, as with STOCSY, driver peaks are employed. The psuedocode for this is shown in Appendix A.4 Algorithm 3.

Rather than correlations between each possible pair in N data sets, the aim is to find relationships which span all N dimensions. For example, with 3 data sets X, Y and Z, we require the combination (X_i, Y_j, Z_k) such that $\text{cor}(X_i, Y_j) > \lambda$, $\text{cor}(Y_j, Z_k) > \lambda$ and $\text{cor}(X_i, Z_k) > \lambda$ for some threshold λ. One critical assumption is the presence of compounds which are present in all N data sets so that, when comparing data from various techniques, it is important to take into consideration the properties of the compounds that can be measured by those methods. For instance, if both LC-MS and GC-MS were to be considered, the range of masses would be different, and any high correlations with GC-MS compounds likely to be fragments in LC-MS. It would also be likely that fragments of the same compound would be strongly correlated, so groups of fragments will therefore be formed.

The DRY method adopts the driver-based approach used in STOCSY, in which variables already identified as beneficial to the analysis are used to find other related variables. The proposed method uses PLS-DA on the individual data sets and then selects drivers using the VIP scores calculated for a suitable number of components. Thus, any variable whose VIP score is greater than a given threshold, recommended to be 1, can be considered as a driver.

6.5.1 Clustering

The second process in the DRY algorithm is to cluster the driver peaks in each data set prior to checking for inter-dataset relationships. Auto-correlation is applied to the drivers identified within a dataset and any highly correlated driver peaks are grouped. This clustering reduces the number of variables to be used as drivers, by selecting a single ‘key driver’ variable from each cluster. Several methods to determine this key variable were explored.

The simplest method to select a ‘key driver’ from each cluster would be to select one at random, as all variables in the cluster are strongly correlated with each other. Thus, any variable from a different data set should be similarly correlated with any variable in the cluster. For consistency, a rule can be imposed, such as selecting the feature with the greatest intensity. This feature should also have the lowest noise interference, which should be beneficial. Alternatively, by choosing the feature which has the greatest average correlation coefficient should be the most representative of the cluster. All methods were tested to determine the most appropriate and to identify any significant differences between them.

6.5.2 Results of DRY applied to data sets obtained from coffee by NMR, LC-MS and GC-MS

Initial features were chosen for each data set based on the VIP scores calculated from a PLS model with origin as the response. Using the variable with the greatest average correlation coefficient as the driver for each cluster reduces the number of NMR components from 280 to 36, the binned GC-MS data from 5685 to 211, and negative-mode LC-MS variables from 18535 to 165 (29 billion potential combinations down to 1 million). This does however come at the cost of potentially missing variables where the compound is not found in all reduced data sets. The NMR features were chosen to provide the drivers, resulting in 27 clusters. The DRY algorithm was applied with the 27 driver variables and identified six correlated groupings of varying sizes using a correlation threshold of 0.95. The smallest grouping consists of a single feature from each set (NMR feature 3.887 ppm, GC m/z 57.0 and RT 9.495 mines, and LC m/z

671.163 and RT 13.32 mins) which was identified in all three individual analyses as potentially being 1,3-butanediol.

Figure 6.15 shows pairwise correlations for the NMR driver at 9.128 ppm over a limited range of the LC-MS and GC-MS data. Here red shows a positive correlation above the threshold of 0.95, blue shows negative correlations and green shows positive correlations with the driver for all data sets. Seven LC-MS features are found to be correlated, including one of the three features (m/z 707.184) used for classification in Section 5.1. A single GC-MS feature (m/z = 57, RT=9.805)is identified above the threshold. Several other variables with strong correlations are found just below the threshold on at least one pairing, but do not appear in the final combination. Cross-referencing identified features in chemometric databases suggests this could be nicotinic acid, related to the roasting of the coffee beans [150].

Part of a second grouping is shown in Figure 6.16. Here, the single driver peak (at 2.414 ppm) resulted in a total of 636 combinations, with 49 GC-MS features and 14 LC-MS features found to be highly correlated. This grouping contains extremely strong correlation coefficients within some of the combinations, with 16 having correlation coefficients above 0.99 for each of the three pairings. Within these top combinations, they all contain the LC-MS variable m/z = 471.130, RT = 14.26. All GC-MS features have a very similar retention time of approximately 16.4 minutes, with the exception of a single feature with m/z 43 and a retention time of 5.08 minutes. The additional information allowed potential identification of this compound as 2-Furanmethanol, acetate [151].

6.6 Tentative compound identification

Although some of the compounds found to be discriminatory between coffee origins could not be identified by individual analyses, combining information using DRY and other data fusion methods could allow tentative compound identification. For example, several compounds found to be associated with taste in the ^{1}H NMR analysis could not be identified (Section 4.2). By cross referencing the ppm values with information on the masses of correlated variables several of these have now been tentatively identified.

The compound around 7.118 ppm with a low p-value for both berry and citrus could potentially be pyridine 1-oxide which increases with roasting [152]. For nuts the feature with the lowest p-value has now been tentatively assigned as a cinnamate derivative [153].

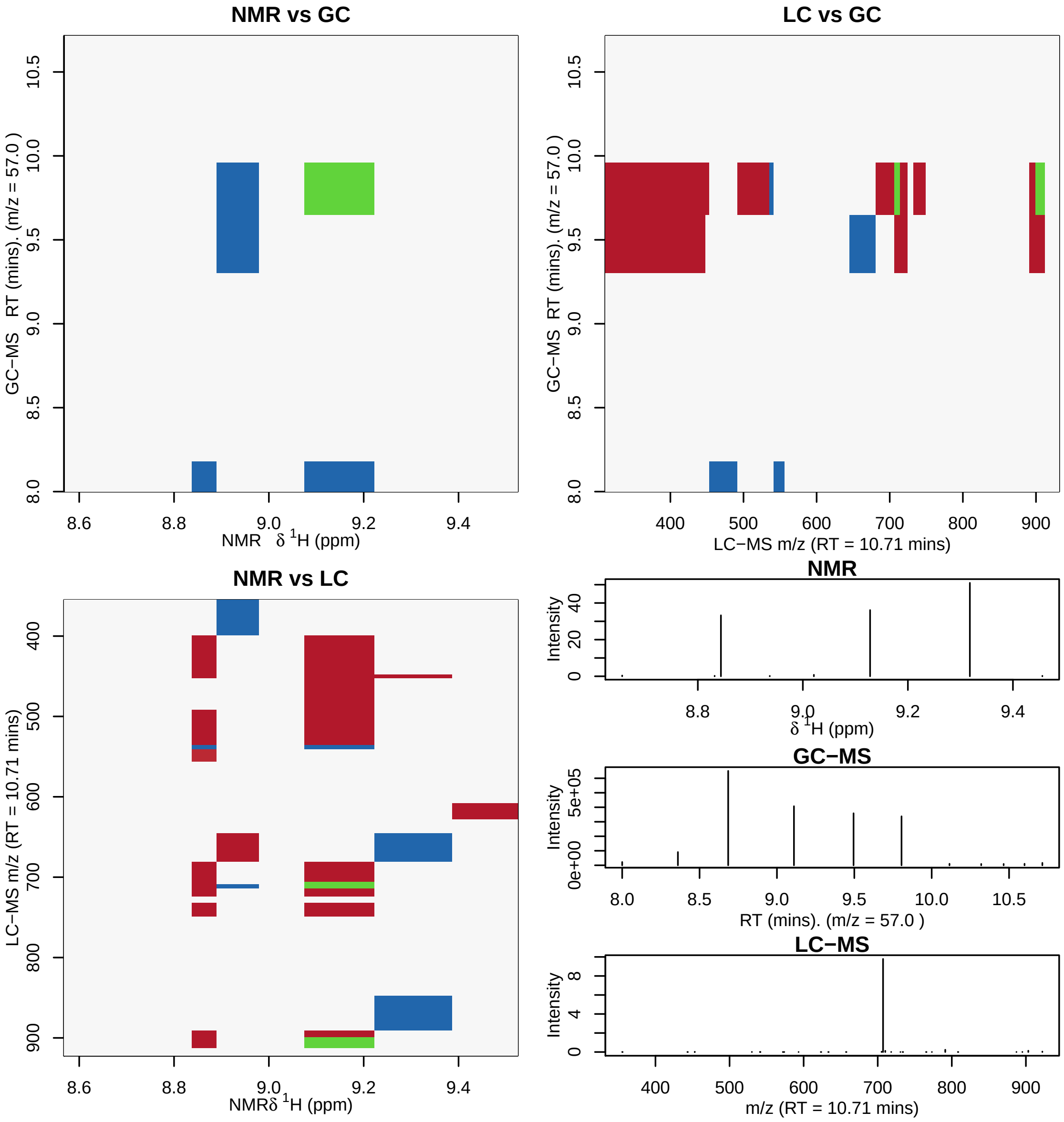


Figure 6.15: Correlation plots between three data sets showing a small region of the spectra in each case. Red shows correlations above 0.95, blue below -0.95, and green shows strong positive correlations between all three sets. The spectra for each region are also plotted.

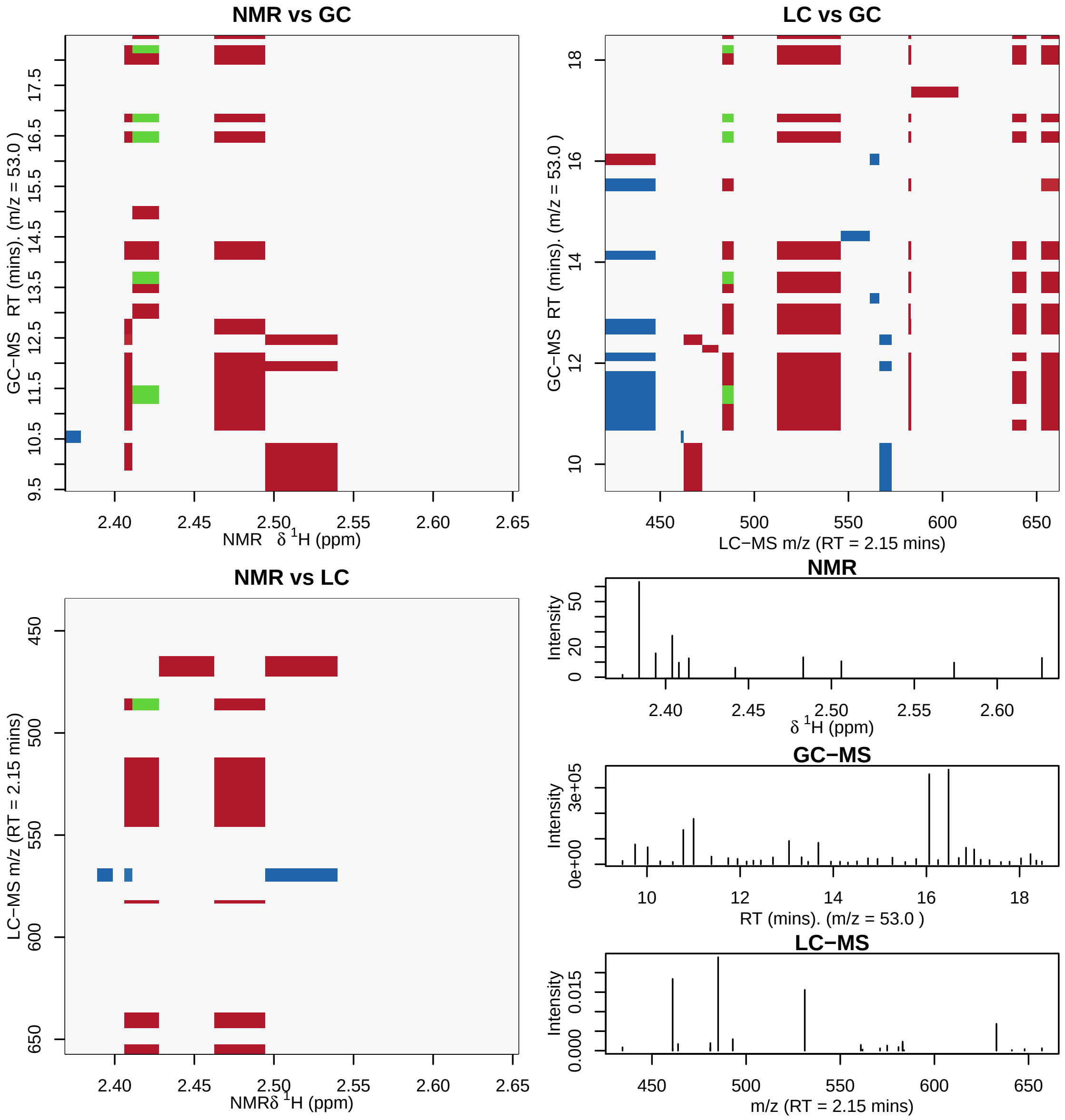


Figure 6.16: Correlation plots between three data sets showing a small region of the spectra. Red shows correlations above 0.95, blue below -0.95, and green shows the strong positive correlations which form a triple pairing between the three sets. The spectra for each region are also plotted.

6.7 Discussion

The use of data fusion has enabled features which were previously unable to be identified to find corresponding features in complementary data sets and then using this information they have been able to have compounds matched with them.

Initially, data sets were combined at high and mid-level data fusion using a variety of techniques. Concatenation of features was able to find which data sets contribute best to origin and taste. For the high-level analysis, the NMR features were the most important for the taste, however for origin the negative-mode LC-MS features on average increased the accuracy the most. For mid-level analysis, the Spiteri et al. approach, where PCA is performed on the data and the scores are then combined resulted in poor clustering via PCA, and instead, PLS was utilised. SHY was able to find many correlated features between the complementary data sets, however, this approach is limited to just two data sets at a time. A generalised approach was derived, called Driver-based Statistical Heterospectroscopy (DRY) which cross-references the correlated features over any number of data sets, and was able to find several features in each set relating to given drivers. This allowed for more parameters to be used with compound identification and was used to tentatively identify more features of interest.

Chapter 7

Conclusion

Throughout this thesis, chemometric methods have been explored with the aims of comparing and evaluating existing methods for use with ^{1}H NMR, GC-MS and LC-MS data, and then developing new methods to pre-process data, offering improvements on existing methods, as well as comparing data sets in data fusion. Analysis of these data sets has also been presented, with the aims of identifying features which can be used as classifiers for geographical origin, taste intensity, and floral origin over multiple data sets.

Within chemometrics, numerous methods are used for both the pre-processing of the data, as well as the statistical analysis of the observational data. A review of these and recent advancements and approaches for use with ^{1}H NMR, GC-MS and LC-MS data was presented in Chapter 2. Various parameters such as the temperature of the sample, pH, and ionic strength [56] can cause unwanted shifts within a spectrum, as well as interference with neighbouring peaks. Common methods for binning and alignment to resolve this were presented in Chapter 2 and applied to both ^{1}H NMR data sets.

The honey data provided an example of how an existing technique, adaptive binning, can make use of a fixed wavelet scale to account for small shifts, and useful features can be extracted from the data. Whereas Chinese honey is considered to not conform to the Codex Alimentarius [17], which describes the food standards set out by FAO/WHO, due to the difference with harvesting processes, exploratory analysis of ^{1}H NMR data revealed that there were no significant sources of variation which supported this claim. While the classification of Chinese honey was still possible with 97.2% accuracy, compound identification of the features with the greatest VIP scores did not suggest these are related to moisture content. While certain honey observations were easily classified, such as the manuka honey from New Zealand, this was not possible for other monofloral honeys. This lower accuracy is caused by several monofloral honeys having a low number of observations making it harder to identify consistent differences between floral types.

For the ^{1}H NMR coffee data, various levels of wavelet transforms were applied and were shown to either be too low of a scale to accommodate these shifts or when set as the lowest level to account for them, became general and over-smoothed peaks. Correlation optimised warping (COW) and Speaq 2.0 were also applied to this data set, with similarly poor results. To find and extract the features within this data set with large shifts throughout, the novel method of multi-stage feature extraction (MSFE) was devised, presented in Chapter 4. This utilises different wavelet scales to create bins for ^{1}H NMR spectra which are at an appropriate scale, while also adapting the bins based on shifting occurring within the spectrum. Classifying the bins based on the resolution of the data allows insight into the quality of the features, while also providing control over which class of data to be used in the analysis.

The Class 1 features are selected based on the range corresponding peak maximums appear for each of the spectra. Where the range is below a threshold, chosen based on the resolution of the data, these regions are considered to be not exhibiting detrimental shifting, and so therefore no alignment would be required to be able to appropriately extract features. Ideally, a significant proportion of the peaks within the spectrum fall under this category, suggesting that there is high repeatability between the observations. Where this is not the case, for the regions the magnitude of the shifting was taken into consideration. For each region first, all the peaks within a region were clustered in such a way to account for shifting over the entire region, and then the shifting that occurs within each cluster over the remaining spectrum was correlated. This allowed for peaks with similar shifting to be identified and have their shifts corrected using one of the features as a driver. This forms an iterative step, where increasing heights in a dendrogram are tested until there is a decrease in the quality of the correction, at which point it is terminated for the regions of interest. An example of this was presented for the ^{1}H NMR coffee data, where four peaks that caused issues with existing methods were found to be correlated based on their shifts. This allowed compound identification to occur, and the tentative assumption for the compound which causes these peak is caffeine. After correcting each region the process is repeated at different wavelet transformation scales, and the results are compared. For each region, the optimal wavelet is selected, thus avoiding the problem of having either a too low or too high scale. For regions that contain peaks, but is too noisy for correction by this method, these were classified as the lowest tier, Class 3, and left as unaltered regions.

To verify this approach the features from each class were used in PLS-DA analysis after being applied to the ^{1}H NMR coffee data. The models were validated using LOO-CV because of the low number of observations available in the data set, and this was able to achieve accuracies around 60% for the Class 1 data alone, and when combined with the Class 2 data this increased to around 90%, which is similar to the accuracies achieved using the GC-MS and LC-MS positive and negative mode data. When compared to existing methods, such as Speaq 2.0 and adaptive binning, it was highlighted how much this approach can improve the quality of the data. Many of these methods failed to achieve over 50% accuracy in the same analysis, with many of them attaining accuracies on par with random classification. This shows that this technique can extract good quality features from otherwise troublesome data. This means that, while a strict process should still be in place for sample homogeneity and handling, it is possible to use data of this low quality. This also avoids the need of having to repeat analysis, which in turn can save costs.

The features extracted from MSFE were analysed to identify which can be used as markers for the samples' geographical origin, as well as finding markers for taste intensity in six different taste profiles. For the origin, PLS-DA was applied with 87.5% accuracy achieved using 5 components. For two of the six tastes, features were found that also act as markers to determine intensity, and for a further two there are markers identified at a lower confidence level that might be able to be used. This analysis and the identification of these potential markers was only possible due to the application of the MSFE algorithm.

Similar analysis was carried out for the LC-MS and GC-MS data. The positive-mode LC-MS data exhibited significant batch issues which required correction, by group standardisation with groups defined by when the solvent top-up was performed. After correction separation between the origins became the predominant source of variance, so much so that PCA-LDA applied to just the first component was able to achieve 80% accuracy, and including the second PC, this increased further to 97.8%. A PLS-DA model was created for the taste, and using the features with the greatest overall VIP scores step-wise regression models were formed for each of the tastes. These were shown to have low residual values for each taste, and the p-values provided insight into which taste each feature contributes to. The negative mode data had no significant batch issues and displayed clear separation based on origin. Three compounds, believed to be Nicotinamide dinucleotide, Apigenin C-glucoside malonylated and an oligosaccharide, had large loadings for the first two PCs, and these alone were able to classify the origin. A similar PLS-DA model was formed for taste, this time with negligible residual values for the step-wise models and was able to be used to identify more compounds relating to taste intensity.

For the GC-MS data set, a random forest was formed from the bilinear interpolated data, and this allowed for metrics such as the mean decrease in accuracy to identify key features relating to origin to be found within the data set. For instance, a pyrrole was found to be a discriminant for origin, as well as what is believed to be a different pyrrole, 1-methylpyrrole.

Data fusion was used to both find common compounds in the complementary data sets, as well as to further improve and compare classification rules. To find features which span all data sets available, in this case three, a method devised called driver-based Statistical Heterospectroscopy (DRY) was presented. Rather than finding correlated features from just two data sets, as is done with SHY, this instead uses any number of data sets and finds common features relating to chosen driver peaks, such as features found to be useful in previous analysis. To limit the compounds found it also restricts each data set to only correlate features shown to be above average in importance relating to the analysis the drivers were chosen from. This provides a more concise set of features which are all exhibiting the same distribution between observations, and then therefore potentially are all responses from the same compound. The use of DRY was able to find multiple sets of features for compounds which were previously unable to be identified, and then utilising these several of these were able to be identified.

7.1 Future Research

The main aim of this research was to evaluate and develop methods for the preprocessing and analysis of chemometric data. For 1D NMR data a method was devised which finds the optimal wavelet level for each peak present in the spectrum, however, with further data and research this method could be applied to 2D NMR spectra instead. Comparisons of shifts within a second dimension will allow for stronger relationships to be utilised, as well as aid with compound identification. The 1D version of method is complete, with the aims to publish this with the algorithm being publicly available as a complete R package.

For the creation of the MSFE technique only the coffee data set was available in order to validate the method. Prior to publishing this method additional data sets which exhibit this severity of shift should be analysed in order to validate the approach. As data of this quality is not usually analysed and published, this is not widely available. This analysis provides an interesting and difficult problem which needs to be overcome.

The SHY algorithm is also completed, with the same aims to publish these findings along side the code as an R package. Again, as only the three data sets presented in this thesis we used when designing the algorithm, further data sets need to be identified and this approach applied to them prior to publishing.

For all of the compounds which were matched with online databases, standards would need to be measured to verify the compounds. Along with this, they would need to have stability trials performed to see if they can be used as suitable markers for their corresponding class. The features of interest found via the mass spectrometry methods can be further split via tandem mass spectrometry, where two mass analyses are used simultaneously. As the ions exit the first they are fragmented before being measured by the second analyser, and this can provide greater information about the chemical composition as well as aid with compound identification.

One limitation of the coffee data was the low number of individual samples. Many of the observations were analytical or biological replicates, and so any discriminant function is limited to just the individual type of coffee, rather than being indicative of the origin. If further testing and analysis of these is to occur, then a wider variety of coffees and a greater range of coffee cherry processing methods available would be beneficial. The roasting intensity for the coffees was also unknown, as well as how fine the beans were ground. Both of these are known to affect the extraction which occurs when brewing, and so adds a new variable to study. The low number of observations within groups is also true for the honey data, with many origins and flower types having very few observations. With sufficient observations from a wide range of origins and floral types a database of honey could be created, and this would allow for rapid and reliable comparisons between unknown honeys to match them. For adulterated honey it could also be used to try to identify which honey was used as a base, as well as proving its inauthenticity.

Appendix A

Appendices

A.1 Chemometric Methods Appendix

k	C6	C12	C18	C24	C30
-10					-0.00030
-9					0.00051
-8				0.00126	0.00308
-7				-0.00230	-0.00588
-6			-0.00536	-0.01039	-0.01433
-5			0.01101	0.02272	0.03310
-4		0.02318	0.03317	0.03773	0.03984
-3		-0.05864	-0.09302	-0.11493	-0.13000
-2	-0.10286	-0.09528	-0.08644	-0.07931	-0.07361
-1	0.47786	0.54604	0.57301	0.58733	0.59619
0	1.20572	1.14936	1.12257	1.10625	1.09502
1	0.54428	0.58973	0.60597	0.61431	0.61940
2	-0.10286	-0.10817	-0.10154	-0.09423	-0.08773
3	-0.02214	-0.08405	-0.11639	-0.13608	-0.14929
4		0.03349	0.04887	0.05563	0.05839
5		0.00794	0.02246	0.03547	0.04621
6		-0.00258	-0.01274	-0.02151	-0.02794
7		-0.00102	-0.00364	-0.00800	-0.01295
8			0.00158	0.00531	0.00956
9			0.00066	0.00179	0.00344
10			-0.00010	-0.00083	-0.00235
11			-0.00005	-0.00037	-0.00090
12				0.00009	0.00043
13				0.00004	0.00020
14				0.00000	-0.00006
15				0.00000	-0.00003
16					0.00001
17					0.00000
18					0.00000
19					0.00000

Table A.1: Coefficients of coiflet wavelets, normalised to have a sum of 2, and rounded to 5dp. Where there is no support no coefficient is shown.

A.2 ^{1}H NMR Analysis Appendix

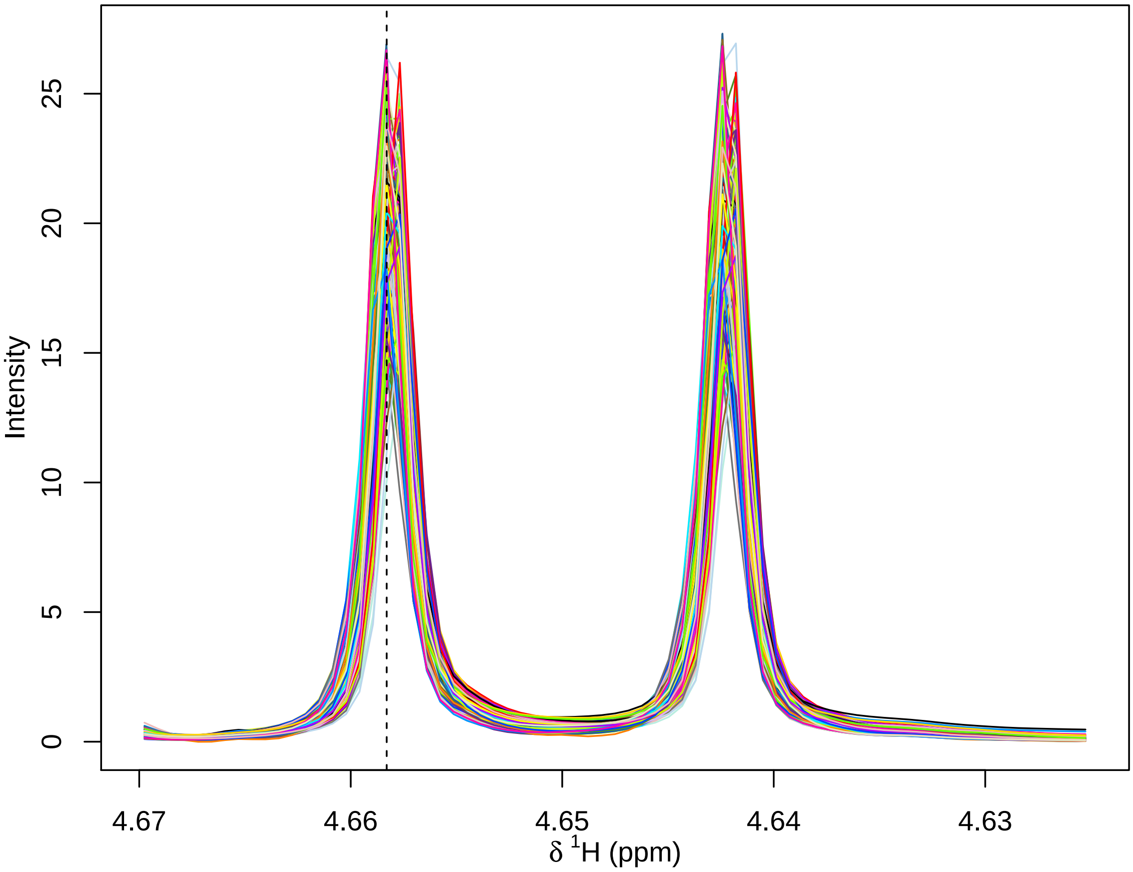


Figure A.1: Region of ^{1}H NMR spectra of the honey data which the greatest loading in the first latent variable for the PLS-DA model on origin.

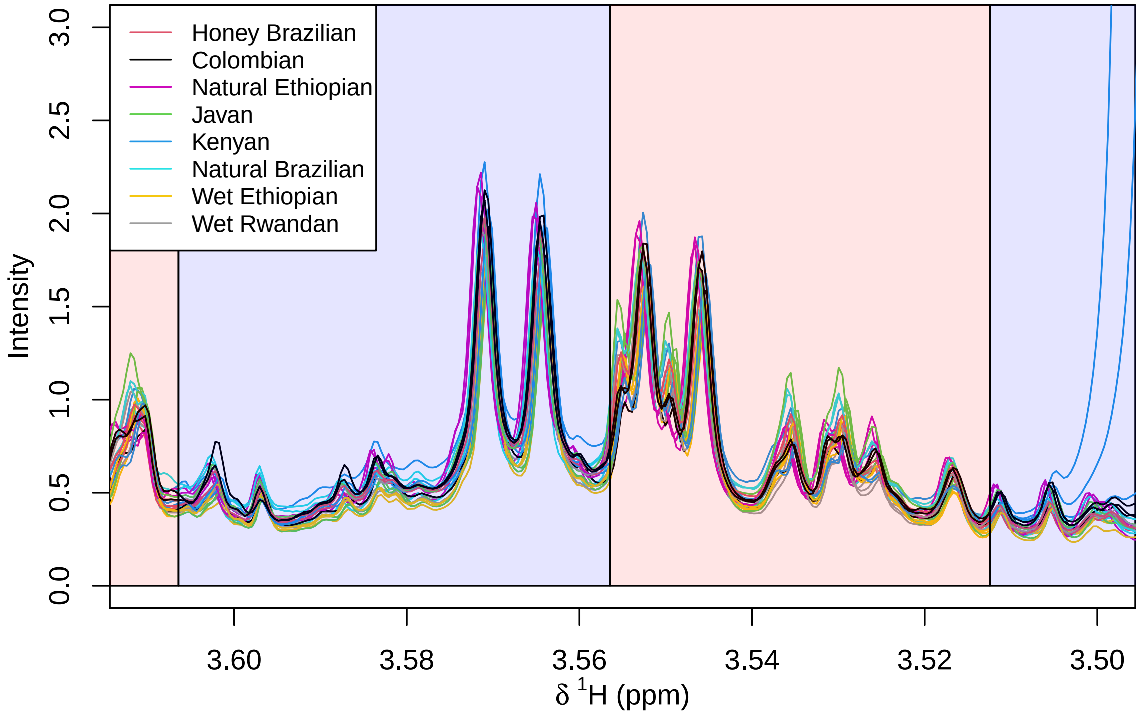


Figure A.2: Region of ^{1}H NMR spectra with bins obtained by adaptive binning with sixth level wavelet smoothing overlaid. Alternating colours represent different bins, with black vertical lines at the boundaries.

Algorithm 1: Pseudo-code for Multi-Stage Feature Extraction

```
Data: Spectra ready for pre-processing
Parameters: Region, Clustering, and Correction parameters
begin
    Apply Adaptive Binning to the data at the desired level to find bins and
      noise;
    Discard regions which are found to only contain noise;
    Calculate reference spectrum;
    Count number of peaks within a width from the peaks in the reference
      spectrum;
    for each region with one peak in a small window w do
        Categorise regions as Class 1;
    end
    for remaining regions do
        Perform Multi-resolution Alignment method;
    end
    if silhouette scores are above the threshold then
        Categorise regions as Class 2;
    end
    else
        Categorise regions as Class 3;
    end
end
```

Algorithm 2: Pseudo-code for Multi-Resolution Alignment algorithm.

```
Data: Regions in spectra containing shift
begin
    for each wavelet transformation scale (decreasing) do
        Apply a wavelet transformation at the current scale;
        Calculate a reference spectrum;
        for each section with a varying number of peaks do
            Identify the observations with the minimum number of peaks in
              this section ($k$);
            Perform correlation-based clustering to resolve shifts in these
              spectra;
            for observations with more peaks in this section than the minimum
              ($k$) do
                Perform $k$-means clustering on peaks with $k$ clusters;
                Form $k$ features from this section by using the minima in the
                  reference spectra;
            end
        end
        Calculate silhouette scores for each section;
        Correlate peak orders with all other sections using Spearman's rank;
        Form a dendrogram from the correlation matrix;
        for each branch point in the dendrogram do
            Start at lowest branch;
            for nodes in branch do
                Pick driver region based on silhouette scores;
                Correct remaining sections in branch based on master;
                Calculate new silhouette scores;
                if silhouette scores have improved then
                    Move up to next branch;
                end
                else
                    Step down branch;
                    Correct shifts using features in this branch;
                    Remove corrected sections from dendrogram;
                end
            end
        end
    end
    for each region do
        Select scale with best silhouette score;
        Extract variables at this level (checking for overlaps);
    end
end
```

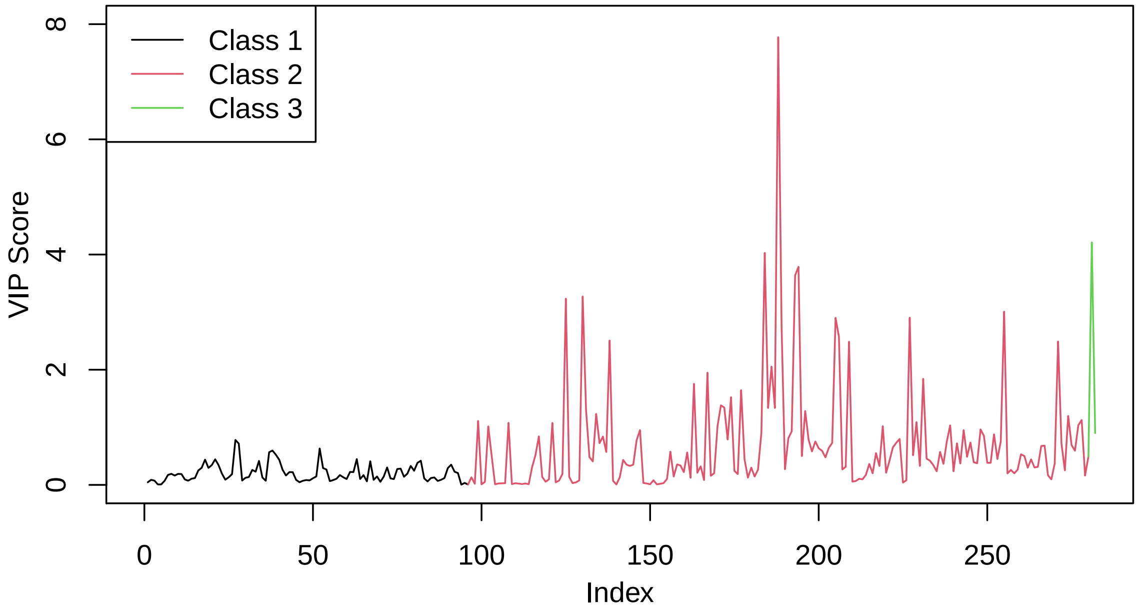


Figure A.3: VIP scores calculated from the five component PLS-DA model using all three classes of data from the MSFE algorithm. The scores are coloured according to their class. The second greatest feature belongs to Class 3, which affects the accuracy of the model.

A.3 LC-MS Analysis Appendix

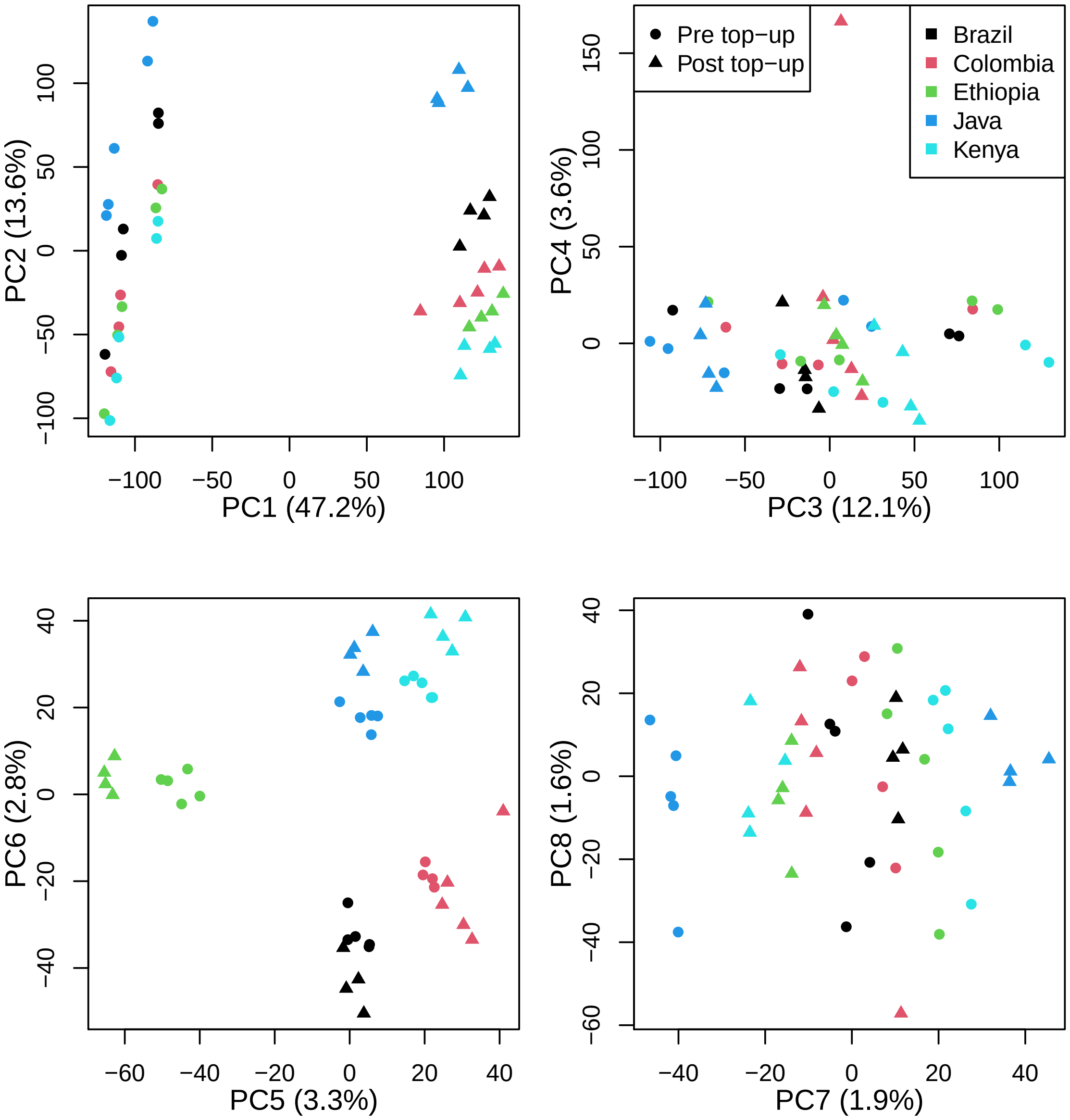


Figure A.4: PCA scores for the first eight components of the LC-MS positive-mode data. Observations are coloured by origin, and the symbol represents if the observation was measured before or after solvent top-up.

A.4 Data Fusion Analysis Appendix

Algorithm 3: Pseudocode for Driver based Statistical Heterospectroscopy

```
Data: Multiple data sets of the same observations
begin
    Perform PLS on the three data sets;
    Calculate VIP scores for each of the features at the optimal number of
      latent variables;
    Identify the features with above average importance from the VIP scores;
    Identify driver variables from these features;
    for each pairing of the data sets do
        Calculate Pearson's correlation coefficient;
        Remove correlations which are below the given threshold;
    end
    Identify the correlations which correspond to the drivers;
    Form the combinations to see which correlations are common in each pair
      of data sets;
end
```

Bibliography

[1] Roma Tauler, Ed., *Chemometrics and Intelligent Laboratory Systems*. Elsevier. [Online]. Available: https://www.journals.elsevier.com/chemometrics-and-intelligent-laboratory-systems (visited on 20/03/2022).

[2] S. Kamiloglu, ‘Authenticity and traceability in beverages’, *Food Chemistry*, vol. 277, no. June 2018, pp. 12–24, 2019, ISSN: 18737072. DOI: 10.1016/j.foodchem.2018.10.091.

[3] M. J. Walker, S. Cowen, K. Gray, P. Hancock and D. T. Burns, ‘Honey authenticity: the opacity of analytical reports - part 1 defining the problem’, *npj Science of Food*, vol. 6, no. 1, pp. 1–9, 2022. DOI: 10.1038/s41538-022-00126-6.

[4] M. Spiteri, E. Jamin, F. Thomas, A. Rebours, M. Lees, K. M. Rogers and D. N. Rutledge, ‘Fast and global authenticity screening of honey using 1 H-NMR profiling’, *Food chemistry*, vol. 189, pp. 60–66, 2015, ISSN: 18737072. DOI: 10.1016/j.foodchem.2014.11.099.

[5] E. Holmes, A. W. Nicholls, J. C. Lindon, S. C. Connor, J. C. Connelly, J. N. Haselden, S. J. Damment, M. Spraul, P. Neidig and J. K. Nicholson, ‘Chemometric models for toxicity classification based on NMR spectra of biofluids’, *Chemical Research in Toxicology*, vol. 13, no. 6, pp. 471–478, Jun. 2000, ISSN: 0893228X. DOI: 10.1021/tx990210t.

[6] S. Dietrich, A. Floegel, M. Troll *et al.*, ‘Random Survival Forest in practice: A method for modelling complex metabolomics data in time to event analysis’, *International Journal of Epidemiology*, vol. 45, no. 5, pp. 1406–1420, 2016, ISSN: 14643685. DOI: 10.1093/ije/dyw145.

[7] F. H. Larsen, F. Van Den Berg and S. B. Engelsen, ‘An exploratory chemometric study of 1H NMR spectra of table wines’, *Journal of Chemometrics*, no. June, pp. 198–208, 2006. DOI: 10.1002/cem.991.

[8] V. Kaškoniene and P. R. Venskutonis, ‘Floral Markers in Honey of Various Botanical and Geographic Origins: A Review’, *Comprehensive Reviews in*

Food Science and Food Safety, vol. 9, no. 6, pp. 620–634, 2010, ISSN: 15414337. DOI: 10.1111/j.1541-4337.2010.00130.x.

[9] F. Bianchi, M. Careri and M. Musci, 'Volatile norisoprenoids as markers of botanical origin of Sardinian strawberry-tree (Arbutus unedo L.) honey: Characterisation of aroma compounds by dynamic headspace extraction and gas chromatography–mass spectrometry', *Food Chemistry*, vol. 89, no. 4, pp. 527–532, 2005, ISSN: 03088146. DOI: 10.1016/j.foodchem.2004.03.009.

[10] F. N. Miros, S. J. Murch and P. R. Shipley, 'Exploring feature selection of St John's wort grown under different light spectra using 1H-NMR spectroscopy', *Phytochemical Analysis*, vol. 31, no. 5, pp. 670–680, 2020, ISSN: 10991565. DOI: 10.1002/pca.2932.

[11] Y. Kato, R. Fujinaka, A. Ishisaka, Y. Nitta, N. Kitamoto and Y. Takimoto, 'Plausible authentication of manuka honey and related products by measuring leptosperin with methyl syringate', *Journal of agricultural and food chemistry*, vol. 62, no. 27, pp. 6400–6407, 2014.

[12] M.-E. Dumas, E. C. Maibaum, C. Teague, H. Ueshima, B. Zhou, J. C. Lindon, J. K. Nicholson, J. Stamler, P. Elliott, Q. Chan *et al.*, 'Assessment of analytical reproducibility of 1H NMR spectroscopy based metabonomics for large-scale epidemiological research: the INTERMAP Study', *Analytical chemistry*, vol. 78, no. 7, pp. 2199–2208, 2006. DOI: 10.1021/ac0517085.Assessment.

[13] S. Mitra, *Sample preparation techniques in analytical chemistry*. Wiley Online Library, 2003, ISBN: 3175723993.

[14] R. Wehrens, J. A. Hageman, F. van Eeuwijk *et al.*, 'Improved batch correction in untargeted MS-based metabolomics', *Metabolomics*, vol. 12, no. 5, 2016, ISSN: 15733890. DOI: 10.1007/s11306-016-1015-8.

[15] R. A. Davis, A. J. Charlton, S. Oehlschlager and J. C. Wilson, 'Novel feature selection method for genetic programming using metabolomic 1H NMR data', *Chemometrics and Intelligent Laboratory Systems*, vol. 81, no. 1, pp. 50–59, Mar. 2006, ISSN: 01697439. DOI: 10.1016/j.chemolab.2005.09.006.

[16] FAO, *Value of Agricultural Production*, Rome., 2021. [Online]. Available: http://www.fao.org/faostat/en/%7B%5C#%7Ddata/QV/ (visited on 20/07/2021).

[17] J. F. C. A. Commission, *Codex alimentarius*. Food \& Agriculture Org., 1992.

[18] A. Dübecke, J. Meulen, B. Schütz, D. Tanner, G. Beckh and C. Lüllmann, 'NMR profiling a defense against honey adulteration', *American Bee Journal*, vol. 158, pp. 83–86, 2018.

[19] J. K. Bond, C. Hitaj, D. Smith, K. Hunt, A. Perez and G. Ferreira, ‘Economic Research Service Economic Research Report Number 290 Honey Bees on the Move: From Pollination to Honey Production and Back’, no. 290, 2021.

[20] K.-P. Raezke, E. Jamin and M. Lees, ‘Honey’, in *FI Handbook on Food Authenticity Issues and Related Analytical Techniques*, J. F. Morin and M. Lees, Eds., 2018, ch. 04. Honey, pp. 43–60, ISBN: 9782956630302.

[21] *A Guide to Coffee Processing from the Experts at Taylors – Taylors of Harrogate*. [Online]. Available: https://discover.taylorsofharrogate.co.uk/blogs/news/coffee-processing (visited on 17/03/2022).

[22] R. A. Fisher, ‘The use of multiple measurements in taxonomic problems’, *Annals of eugenics*, vol. 7, no. 2, pp. 179–188, 1936. DOI: 10.1111/J.1469-1809.1936.TB02137.X.

[23] O. Cloarec, M. E. Dumas, A. Craig *et al.*, ‘Statistical total correlation spectroscopy: An exploratory approach for latent biomarker identification from metabolic 1H NMR data sets’, *Analytical Chemistry*, vol. 77, no. 5, pp. 1282–1289, Mar. 2005, ISSN: 00032700. DOI: 10.1021/ac048630x.

[24] D. J. Crockford, E. Holmes, J. C. Lindon, R. S. Plumb, S. Zirah, S. J. Bruce, P. Rainville, C. L. Stumpf and J. K. Nicholson, ‘Statistical heterospectroscopy, an approach to the integrated analysis of NMR and UPLC-MS data sets: Application in metabonomic toxicology studies’, *Analytical Chemistry*, vol. 78, no. 2, pp. 363–371, Jan. 2006, ISSN: 00032700. DOI: 10.1021/ac051444m.

[25] W. D. Van Horn, A. J. Beel, C. Kang and C. R. Sanders, ‘The impact of window functions on NMR-based paramagnetic relaxation enhancement measurements in membrane proteins’, *Biochimica et Biophysica Acta - Biomembranes*, vol. 1798, no. 2, pp. 140–149, 2010, ISSN: 00052736. DOI: 10.1016/j.bbamem.2009.08.022.

[26] P. Giraudeau and S. Akoka, ‘Sensitivity and lineshape improvement in ultrafast 2D NMR by optimized apodization in the spatially encoded dimension’, *Magnetic Resonance in Chemistry*, vol. 49, no. 6, pp. 307–313, 2011, ISSN: 07491581. DOI: 10.1002/mrc.2746.

[27] R. A. Davis, A. J. Charlton, J. Godward, S. A. Jones, M. Harrison and J. C. Wilson, ‘Adaptive binning: An improved binning method for metabolomics data using the undecimated wavelet transform’, *Chemometrics and Intelligent Laboratory Systems*, vol. 85, no. 1, pp. 144–154, Jan. 2006, ISSN: 01697439. DOI: 10.1016/j.chemolab.2006.08.014.

[28] Chemical Book, *Caffeine(58-08-2) 1H NMR*. [Online]. Available: https://www.chemicalbook.com/SpectrumEN%7B%5C_%7D58-08-2%7B%5C_%7D1HNMR.htm (visited on 22/03/2022).

[29] F. W. McLafferty, F. Tureček and F. Turecek, *Interpretation of mass spectra*. University science books, 1993.

[30] NIST Mass Spectrometry Data Center and William E. Wallace, 'Mass Spectra', in *NIST Chemistry WebBook, NIST Standard Reference Database Number 69*, Eds. P.J. Linstrom and W.G. Mallard, Eds., National Institute of Standards and Technology, Gaithersburg MD, 20899.

[31] D. Bier, R. Hartmann and M. Holschbach, 'Collision-induced dissociation studies of caffeine in positive electrospray ionisation mass spectrometry using six deuterated isotopomers and one N1-ethylated homologue', *Rapid Communications in Mass Spectrometry*, vol. 27, no. 8, pp. 885–895, 2013, ISSN: 09514198. DOI: 10.1002/rcm.6520.

[32] N. Okahashi, K. Maeda, S. Kawana, J. Iida, H. Shimizu and F. Matsuda, 'Sugar phosphate analysis with baseline separation and soft ionization by gas chromatography-negative chemical ionization-mass spectrometry improves flux estimation of bidirectional reactions in cancer cells', *Metabolic Engineering*, vol. 51, no. August 2018, pp. 43–49, 2019, ISSN: 10967184. DOI: 10.1016/j.ymben.2018.08.011.

[33] J. C. Cocuron, Z. Ross and A. P. Alonso, 'Liquid chromatography tandem mass spectrometry quantification of13C-labeling in sugars', *Metabolites*, vol. 10, no. 1, 2020, ISSN: 22181989. DOI: 10.3390/metabo10010030.

[34] A. K. Smilde, M. J. Van Der Werf, S. Bijlsma, B. J. Van Der Werff-Van Der Vat and R. H. Jellema, 'Fusion of mass spectrometry-based metabolomics data', *Analytical Chemistry*, vol. 77, no. 20, pp. 6729–6736, 2005, ISSN: 00032700. DOI: 10.1021/ac051080y.

[35] C. Beirnaert, P. Meysman, T. N. Vu, N. Hermans, S. Apers, L. Pieters, A. Covaci and K. Laukens, 'speaq 2.0: A complete workflow for high-throughput 1D NMR spectra processing and quantification', *PLoS Computational Biology*, vol. 14, no. 3, pp. 1–25, 2018, ISSN: 15537358. DOI: 10.1371/journal.pcbi.1006018.

[36] P. Du, W. A. Kibbe and S. M. Lin, 'Improved peak detection in mass spectrum by incorporating continuous wavelet transform-based pattern matching', *Bioinformatics*, vol. 22, no. 17, pp. 2059–2065, Sep. 2006, ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btl355.

[37] I. Daubechies, 'Ten Lectures on Wavelets', *Ten Lectures on Wavelets*, 1992. doi: 10.1137/1.9781611970104.

[38] S. G. Mallat, 'A theory for multiresolution signal decomposition: the wavelet representation', *IEEE Transactions on Pattern Analysis \& Machine Intelligence*, no. 7, pp. 674–693, 1989.

[39] A. Haar, 'On the Theory of Orthogonal Function Systems (translation)', *Mathematische Annalen*, vol. 69, no. 3, pp. 331–371, 1909.

[40] P. Subramani, R. Sahu and S. Verma, 'Feature selection using Haar wavelet power spectrum', *BMC Bioinformatics*, vol. 7, 2006, issn: 14712105. doi: 10.1186/1471-2105-7-432.

[41] D. L. Donoho and I. M. Johnstone, 'Ideal spatial adaptation by wavelet shrinkage', *Biometrika*, vol. 81, no. 3, pp. 425–455, Sep. 1994, issn: 0006-3444. doi: 10.1093/biomet/81.3.425.

[42] S. Mallat, *A Wavelet Tour of Signal Processing*. Elsevier, 2009, isbn: 978-0-12-374370-1. doi: 10.1016/B978-0-12-374370-1.X0001-8.

[43] S. Halouska and R. Powers, 'Negative impact of noise on the principal component analysis of NMR data', *Journal of Magnetic Resonance*, vol. 178, no. 1, pp. 88–95, Jan. 2006, issn: 10907807. doi: 10.1016/j.jmr.2005.08.016.

[44] A. Derome, 'Modern NMR Techniques for Chemistry Reserach', in *Organic Chemistry Series*, J. Baldwin, Ed., vol. 6, Pergamon Press, 1987, pp. 22–24.

[45] B. Walczak and D. L. Massart, 'Noise suppression and spinal compression using the wavelet packet transform', *Chemometrics and Intelligent Laboratory Systems*, vol. 36, no. 2, pp. 81–94, Apr. 1997, issn: 01697439. doi: 10.1016/S0169-7439(96)00077-9.

[46] C. Zheng and Y. Zhang, 'Low-field pulsed NMR signal denoising based on wavelet transform', in *2007 IEEE 15th Signal Processing and Communications Applications, SIU*, IEEE, Jun. 2007, isbn: 1424407192. doi: 10.1109/SIU.2007.4298696.

[47] J. A. Cadzow, 'Signal Enhancement—A Composite Property Mapping Algorithm', *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 36, no. 1, pp. 49–62, 1988, issn: 00963518. doi: 10.1109/29.1488.

[48] Y. Li, M. Jiang and F. Liu, 'Time fractional super-diffusion model and its application in peak-preserving smoothing', *Chemometrics and Intelligent Laboratory Systems*, vol. 175, no. February 2018, pp. 13–19, 2018, issn: 18733239. doi: 10.1016/j.chemolab.2018.02.005.

[49] G. Laurent, W. Woelffel, V. Barret-Vivin, E. Gouillart and C. Bonhomme, ‘Denoising applied to spectroscopies – part I: concept and limits’, *Applied Spectroscopy Reviews*, vol. 0, no. 0, pp. 1–29, 2019, ISSN: 0570-4928. DOI: 10.1080/05704928.2018.1523183.

[50] C. Tang, ‘An Analysis of Baseline Distortion and Offset in NMR Spectra’, *Journal of Magnetic Resonance, Series A*, vol. 109, no. 2, pp. 232–240, Aug. 1994, ISSN: 10641858. DOI: 10.1006/jmra.1994.1160.

[51] D. G. Davis, ‘Elimination of baseline distortions and minimization of artifacts from phased 2D NMR spectra’, *Journal of Magnetic Resonance (1969)*, vol. 81, no. 3, pp. 603–607, Feb. 1989, ISSN: 00222364. DOI: 10.1016/0022-2364(89)90100-5.

[52] W. Dietrich, C. H. Rüdel and M. Neumann, ‘Fast and precise automatic baseline correction of one- and two-dimensional nmr spectra’, *Journal of Magnetic Resonance (1969)*, vol. 91, no. 1, pp. 1–11, Jan. 1991, ISSN: 00222364. DOI: 10.1016/0022-2364(91)90402-F.

[53] J. Carlos Cobas, M. A. Bernstein, M. Martín-Pastor and P. G. Tahoces, ‘A new general-purpose fully automatic baseline-correction procedure for 1D and 2D NMR data’, *Journal of Magnetic Resonance*, vol. 183, no. 1, pp. 145–151, Nov. 2006, ISSN: 10907807. DOI: 10.1016/j.jmr.2006.07.013.

[54] M. Sawall, E. von Harbou, A. Moog, R. Behrens, H. Schröder, J. Simoneau, E. Steimers and K. Neymeyr, ‘Multi-objective optimization for an automated and simultaneous phase and baseline correction of NMR spectral data’, *Journal of Magnetic Resonance*, vol. 289, pp. 132–141, 2018, ISSN: 10960856. DOI: 10.1016/j.jmr.2018.02.012.

[55] S. Sokolenko, T. Jézéquel, G. Hajjar, J. Farjon, S. Akoka and P. Giraudeau, ‘Robust 1D NMR lineshape fitting using real and imaginary data in the frequency domain’, *Journal of Magnetic Resonance*, vol. 298, no. 2019, pp. 91–100, Jan. 2019, ISSN: 10907807. DOI: 10.1016/j.jmr.2018.11.004.

[56] J. T. M. Pearce, T. J. Athersuch, T. M. D. Ebbels, J. C. Lindon, J. K. Nicholson and H. C. Keun, ‘Robust algorithms for automated chemical shift calibration of 1D 1H NMR spectra of blood serum’, *Analytical Chemistry*, vol. 80, no. 18, pp. 7158–7162, Sep. 2008, ISSN: 00032700. DOI: 10.1021/ac8011494.

[57] M. Defernez and I. J. Colquhoun, ‘Factors affecting the robustness of metabolite fingerprinting using 1H NMR spectra’, *Phytochemistry*, vol. 62, no. 6, pp. 1009–1017, Mar. 2003, ISSN: 00319422. DOI: 10.1016/S0031-9422(02)00704-5.

[58] O. Beckonert, J. Monnerjahn, U. Bonk and D. Leibfritz, 'Visualizing metabolic changes in breast-cancer tissue using 1H-NMR spectroscopy and self-organizing maps', *NMR in Biomedicine*, vol. 16, no. 1, pp. 1–11, 2003, ISSN: 09523480. DOI: 10.1002/nbm.797.

[59] M. Čuperlović-Culf, *NMR metabolomics in cancer research.* Elsevier, 2012.

[60] T. D. Meyer, D. Sinnaeve, B. V. Gasse, E. Tsiporkova, E. R. Rietzschel, M. L. D. Buyzere, T. C. Gillebert, S. Bekaert, J. C. Martins and W. V. Criekinge, 'NMR-Based Characterization of Metabolic Alterations in Hypertension Using an Adaptive, Intelligent Binning Algorithm', *Analytical Chemistry*, vol. 80, no. 10, pp. 3783–3790, May 2008, ISSN: 0003-2700. DOI: 10.1021/ac7025964.

[61] B. Worley and R. Powers, 'Generalized adaptive intelligent binning of multiway data', *Chemometrics and Intelligent Laboratory Systems*, vol. 146, pp. 42–46, 2015, ISSN: 18733239. DOI: 10.1016/j.chemolab.2015.05.005.

[62] N. Trbovic, F. Dancea, T. Langer and U. Günther, 'Using wavelet de-noised spectra in NMR screening', *Journal of Magnetic Resonance*, vol. 173, no. 2, pp. 280–287, Apr. 2005, ISSN: 10907807. DOI: 10.1016/j.jmr.2004.11.032.

[63] F. Puig-Castellví, Y. Pérez, B. Piña, R. Tauler and I. Alfonso, 'Compression of multidimensional NMR spectra allows a faster and more accurate analysis of complex samples', *Chemical Communications*, vol. 54, no. 25, pp. 3090–3093, 2018, ISSN: 1364548X. DOI: 10.1039/c7cc09891j.

[64] A. Kassidas, J. F. MacGregor and P. A. Taylor, 'Synchronization of Batch Trajectories Using Dynamic Time Warping', *AIChE Journal*, vol. 44, no. 4, pp. 864–875, Apr. 1998, ISSN: 00011541. DOI: 10.1002/aic.690440412.

[65] N. P. V. Nielsen, J. M. Carstensen and J. Smedsgaard, 'Aligning of single and multiple wavelength chromatographic profiles for chemometric data analysis using correlation optimised warping', *Journal of Chromatography A*, vol. 805, no. 1-2, pp. 17–35, May 1998, ISSN: 00219673. DOI: 10.1016/S0021-9673(98)00021-1.

[66] V. Pravdova, B. Walczak and D. L. Massart, 'A comparison of two algorithms for warping of analytical signals', *Analytica Chimica Acta*, vol. 456, no. 1, pp. 77–92, Apr. 2002, ISSN: 00032670. DOI: 10.1016/S0003-2670(02)00008-9.

[67] R. J. O. Torgrip, M. Åberg, B. Karlberg and S. P. Jacobsson, 'Peak alignment using reduced set mapping', *Journal of Chemometrics*, vol. 17, no. 11, pp. 573–582, Nov. 2003, ISSN: 08869383. DOI: 10.1002/cem.824.

[68] Z. Wang and S. B. Kim, 'Automatic alignment of high-resolution NMR spectra using a Bayesian estimation approach', in *Proceedings - International Conference on Pattern Recognition*, vol. 4, IEEE, 2006, pp. 667–670. DOI: 10.1109/ICPR.2006.295.

[69] F. Savorani, G. Tomasi and S. B. Engelsen, 'icoshift: A versatile tool for the rapid alignment of 1D NMR spectra', *Journal of Magnetic Resonance*, vol. 202, no. 2, pp. 190–202, Feb. 2010, ISSN: 10907807. DOI: 10.1016/j.jmr.2009.11.012.

[70] K. Kumar, 'Optimizing the process of reference selection for correlation optimised warping (COW) and interval correlation shifting (icoshift) analysis: Automating the chromatographic alignment procedure', *Analytical Methods*, vol. 10, no. 2, pp. 190–203, 2018, ISSN: 17599679. DOI: 10.1039/c7ay02340e.

[71] K. Wang, G. A. Barding and C. K. Larive, 'Peak alignment of one-dimensional NMR spectra by means of an intensity fluctuation frequency difference (IFFD) segment-wise algorithm', *Analytical Methods*, vol. 7, no. 22, pp. 9673–9682, 2015, ISSN: 1759-9660. DOI: 10.1039/C5AY01079A.

[72] V. Maus, G. CĂmara, R. Cartaxo, A. Sanchez, F. M. Ramos and G. R. De Queiroz, 'A Time-Weighted Dynamic Time Warping Method for Land-Use and Land-Cover Mapping', *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 9, no. 8, pp. 3729–3739, 2016, ISSN: 21511535. DOI: 10.1109/JSTARS.2016.2517118.

[73] M. Belgiu and O. Csillik, 'Sentinel-2 cropland mapping using pixel-based and object-based time-weighted dynamic time warping analysis', *Remote Sensing of Environment*, vol. 204, no. January 2017, pp. 509–523, 2018, ISSN: 00344257. DOI: 10.1016/j.rse.2017.10.005.

[74] R. Varatharajan, G. Manogaran, M. K. Priyan and R. Sundarasekar, 'Wearable sensor devices for early detection of Alzheimer disease using dynamic time warping algorithm', *Cluster Computing*, vol. 21, no. 1, pp. 681–690, 2018, ISSN: 15737543. DOI: 10.1007/s10586-017-0977-2.

[75] T. Syeda-Mahmood, D. Beymer and F. Wang, 'Shape-based matching of ECG recordings', *Annual International Conference of the IEEE Engineering in Medicine and Biology - Proceedings*, pp. 2012–2018, 2007, ISSN: 05891019. DOI: 10.1109/IEMBS.2007.4352714.

[76] T. Giorgino, 'Computing and Visualizing Dynamic Time Warping Alignments in R: The dtw Package', *Journal of Statistical Software*, vol. 31, no. 7, 2009.

[77] L. Rabiner and B.-H. Juang, *Fundamentals of Speech Recognition*. Upper Saddle River, NJ, USA: Prentice-Hall, Inc., 1993, ISBN: 0-13-015157-2.

[78] Y. Xi, J. S. Ropp, M. R. Viant, D. L. Woodruff and P. Yu, 'Automated screening for metabolites in complex mixtures using 2D COSY NMR spectroscopy', *Metabolomics*, vol. 2, no. 4, pp. 221–233, Nov. 2006, ISSN: 15733882. DOI: 10.1007/s11306-006-0036-0.

[79] J. Hao, W. Astle, M. De iorio and T. M. D. Ebbels, 'Batman-an R package for the automated quantification of metabolites from nuclear magnetic resonance spectra using a bayesian model', *Bioinformatics*, vol. 28, no. 15, pp. 2088–2090, 2012, ISSN: 13674803. DOI: 10.1093/bioinformatics/bts308.

[80] T. Padayachee, T. Khamiakova, E. Louis, P. Adriaensens and T. Burzykowski, 'The impact of the method of extracting metabolic signal from 1 H-NMR data on the classification of samples: A case study of binning and BATMAN in lung cancer', *PLoS ONE*, vol. 14, no. 2, pp. 1–17, 2019, ISSN: 19326203. DOI: 10.1371/journal.pone.0211854.

[81] C. Ludwig and M. R. Viant, 'Two-dimensional J-resolved NMR spectroscopy: Review of a key methodology in the metabolomics toolbox', *Phytochemical Analysis*, vol. 21, no. 1, pp. 22–32, Jan. 2010, ISSN: 09580344. DOI: 10.1002/pca.1186.

[82] T. Vu, E. Riekeberg, Y. Qiu and R. Powers, 'Comparing normalization methods and the impact of noise', *Metabolomics*, vol. 14, no. 8, pp. 1–10, 2018, ISSN: 15733890. DOI: 10.1007/s11306-018-1400-6.

[83] M. E. Dumas, E. C. Maibaum, C. Teague *et al.*, 'Assessment of analytical reproducibility of 1H NMR spectroscopy based metabonomics for large-scale epidemiological research: The INTERMAP study', *Analytical Chemistry*, vol. 78, no. 7, pp. 2199–2208, Apr. 2006, ISSN: 00032700. DOI: 10.1021/ac0517085.

[84] M. M. Hendriks, F. A. Eeuwijk, R. H. Jellema, J. A. Westerhuis, T. H. Reijmers, H. C. Hoefsloot and A. K. Smilde, 'Data-processing strategies for metabolomics studies', *TrAC - Trends in Analytical Chemistry*, vol. 30, no. 10, pp. 1685–1698, 2011, ISSN: 01659936. DOI: 10.1016/j.trac.2011.04.019.

[85] M. Rusilowicz, M. Dickinson, A. Charlton, S. O'Keefe and J. C. Wilson, 'A batch correction method for liquid chromatography–mass spectrometry data that does not depend on quality control samples', *Metabolomics*, vol. 12, no. 3, pp. 1–11, 2016, ISSN: 15733890. DOI: 10.1007/s11306-016-0972-2.

[86] F. Dieterle, A. Ross, G. Schlotterbeck and H. Senn, 'Probabilistic quotient normalization as robust method to account for dilution of complex biological mixtures. Application in 1H NMR metabonomics.', *Analytical chemistry*,

vol. 78, no. 13, pp. 4281–90, Jul. 2006, ISSN: 0003-2700. DOI: 10.1021/ac051632c.

[87] H. C. Keun, T. M. Ebbels, H. Antti, M. E. Bollard, O. Beckonert, E. Holmes, J. C. Lindon and J. K. Nicholson, 'Improved analysis of multivariate data by variable stability scaling: Application to NMR-based metabolic profiling', *Analytica Chimica Acta*, vol. 490, no. 1-2, pp. 265–276, Aug. 2003, ISSN: 00032670. DOI: 10.1016/S0003-2670(03)00094-1.

[88] P. V. Purohit, D. M. Rocke, M. R. Viant and D. L. Woodruff, 'Discrimination Models Using Variance-Stabilizing Transformation of Metabolomic NMR Data', *OMICS: A Journal of Integrative Biology*, vol. 8, no. 2, pp. 118–130, Jul. 2004, ISSN: 1536-2310. DOI: 10.1089/1536231041388348.

[89] R. Stoyanova, A. W. Nicholls, J. K. Nicholson, J. C. Lindon and T. R. Brown, 'Automatic alignment of individual peaks in large high-resolution spectral data sets', *Journal of Magnetic Resonance*, vol. 170, no. 2, pp. 329–335, Oct. 2004, ISSN: 10907807. DOI: 10.1016/j.jmr.2004.07.009.

[90] Y. Liu and S. D. Brown, *Wavelet multiscale regression from the perspective of data fusion: New conceptual approaches*, 2004. DOI: 10.1007/s00216-004-2776-x.

[91] D. L. Hall, J. Llinas, C. L. Bowman, A. N. Steinberg, E. Waltz, R. R. Brooks, L. Grewe and J. W. Carl, *Mathematical Techniques in Multisensor Data Fusion 1.Introduction to Multisensor Data Fusion.* Artech House, 2001.

[92] M. Hohmann, Y. Monakhova, S. Erich, N. Christoph, H. Wachter and U. Holzgrabe, 'Differentiation of Organically and Conventionally Grown Tomatoes by Chemometric Analysis of Combined Data from Proton Nuclear Magnetic Resonance and Mid-infrared Spectroscopy and Stable Isotope Analysis', *Journal of Agricultural and Food Chemistry*, vol. 63, no. 43, pp. 9666–9675, 2015, ISSN: 15205118. DOI: 10.1021/acs.jafc.5b03853.

[93] S. Erich, S. Schill, E. Annweiler, H. U. Waiblinger, T. Kuballa, D. W. Lachenmeier and Y. B. Monakhova, 'Combined chemometric analysis of 1H NMR, 13C NMR and stable isotope data to differentiate organic and conventional milk', *Food Chemistry*, vol. 188, pp. 1–7, 2015, ISSN: 18737072. DOI: 10.1016/j.foodchem.2015.04.118.

[94] M. Spiteri, E. Dubin, J. Cotton, M. Poirel, B. Corman, E. Jamin, M. Lees and D. Rutledge, 'Data fusion between high resolution 1 H-NMR and mass spectrometry: a synergetic approach to honey botanical origin characterization', *Analytical and Bioanalytical Chemistry*, vol. 408, no. 16, pp. 4389–4401, 2016, ISSN: 16182650. DOI: 10.1007/s00216-016-9538-4.

[95] R. Ríos-Reina, R. M. Callejón, F. Savorani, J. M. Amigo and M. Cocchi, 'Data fusion approaches in spectroscopic characterization and classification of PDO wine vinegars', *Talanta*, vol. 198, no. October 2018, pp. 560–572, 2019, ISSN: 00399140. DOI: 10.1016/j.talanta.2019.01.100.

[96] N. Cavallini, F. Savorani, R. Bro and M. Cocchi, 'Fused adjacency matrices to enhance information extraction: The beer benchmark', *Analytica Chimica Acta*, vol. 1061, pp. 70–83, 2019, ISSN: 18734324. DOI: 10.1016/j.aca.2019.02.023.

[97] S. Beteinakis, A. Papachristodoulou, G. Gogou, S. Katsikis, E. Mikros and M. Halabalaki, 'NMR-based metabolic profiling of edible olives-determination of quality parameters', *Molecules*, vol. 25, no. 15, 2020, ISSN: 14203049. DOI: 10.3390/molecules25153339.

[98] B. Worley and R. Powers, 'Multivariate analysis in metabolomics', *Current Metabolomics*, vol. 1, no. 1, pp. 92–107, 2013.

[99] K. Pearson, 'LIII. On lines and planes of closest fit to systems of points in space', *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, vol. 2, no. 11, pp. 559–572, 1901, ISSN: 1941-5982. DOI: 10.1080/14786440109462720.

[100] H. Hotelling, 'Analysis of a complex of statistical variables into principal components', *Journal of Educational Psychology*, vol. 24, no. 6, pp. 417–441, 1933, ISSN: 00220663. DOI: 10.1037/h0071325.

[101] X. Shao and Y. Li, 'Classification and Prediction by LF NMR', *Food and Bioprocess Technology*, vol. 5, no. 5, pp. 1817–1823, Nov. 2012, ISSN: 19355130. DOI: 10.1007/s11947-010-0455-9.

[102] H. Wold, 'Estimation of principal components and related models by iterative least squares', *Multivariate analysis*, pp. 391–420, 1966.

[103] M. Sjöström, S. Wold and B. Söderström, 'PLS discriminant plots', in *Pattern recognition in practice*, Elsevier, 1986, pp. 461–470.

[104] S. Wold, M. Sjöström and L. Eriksson, 'PLS-regression: A basic tool of chemometrics', *Chemometrics and Intelligent Laboratory Systems*, vol. 58, no. 2, pp. 109–130, 2001, ISSN: 01697439. DOI: 10.1016/S0169-7439(01)00155-1. arXiv: S0169-74390100155-1.

[105] S. Rännar, P. Geladi, F. Lindgren and S. Wold, 'A PLS KERNEL ALGORITHM FOR DATA SETS WITH MANY VARIABLES AND FEWER OBJECTS. PART 1: THEORY AND ALGORITHM', *Journal of Chemometrics*, vol. 9, no. 6, pp. 459–470, 1995, ISSN: 1099128X. DOI: 10.1002/cem.1180090604.

[106] F. Lindgren, P. Geladi and S. Wold, 'The kernel algorithm for PLS', *Journal of Chemometrics*, vol. 7, no. 1, pp. 45–59, Jan. 2005, ISSN: 0886-9383. DOI: 10.1002/cem.1180070104.

[107] S. Wold, A. Ruhe, H. Wold and W. J. Dunn III, 'THE COLLINEARITY PROBLEM IN LINEAR REGRESSION. THE PARTIAL LEAST SQUARES (PLS) APPROACH TO GENERALIZED INVERSES*', *SIAM J. Sci. and Stat. Comput.*, vol. 5, no. 3, pp. 735–743, 1984. arXiv: arXiv:1308.0863v1.

[108] J. Trygg and S. Wold, 'Orthogonal projections to latent structures (O-PLS)', *Journal of Chemometrics*, vol. 16, no. 3, pp. 119–128, 2002, ISSN: 08869383. DOI: 10.1002/cem.695.

[109] M. Barker and W. Rayens, 'Partial least squares for discrimination', *Journal of Chemometrics*, vol. 17, no. 3, pp. 166–173, 2003, ISSN: 08869383. DOI: 10.1002/cem.785.

[110] L. Breiman, 'Random forests', *Machine learning*, pp. 5–32, 2001.

[111] X. Song, S. She, M. Xin, L. Chen, Y. Li, Y. V. Heyden, K. M. Rogers and L. Chen, 'Detection of adulteration in Chinese monofloral honey using 1H nuclear magnetic resonance and chemometrics', *Journal of Food Composition and Analysis*, vol. 86, no. June 2019, 2020, ISSN: 08891575. DOI: 10.1016/j.jfca.2019.103390.

[112] X. Wang, Y. Chen, Y. Hu, J. Zhou, L. Chen and X. Lu, 'Systematic Review of the Characteristic Markers in Honey of Various Botanical, Geographic, and Entomological Origins', *ACS Food Science & Technology*, vol. 2, no. 2, pp. 206–220, 2022, ISSN: 2692-1944. DOI: 10.1021/acsfoodscitech.1c00422.

[113] J. A. Donarski, S. A. Jones, M. Harrison, M. Driffield and A. J. Charlton, 'Identification of botanical biomarkers found in Corsican honey', *Food Chemistry*, vol. 118, no. 4, pp. 987–994, Feb. 2010, ISSN: 03088146. DOI: 10.1016/j.foodchem.2008.10.033.

[114] *Manuka Lab Monofloral Manuka Honey 700 MGO 500g | Holland & Barrett.* [Online]. Available: https://www.hollandandbarrett.com/shop/product/manuka-lab-monofloral-manuka-honey-700-mgo-60050503?skuid=050507 (visited on 18/03/2022).

[115] R. J. Weston, L. K. Brocklebank and Y. Lu, 'Identification and quantitative levels of antibacterial components of some New Zealand honeys', *Food Chemistry*, vol. 70, no. 4, pp. 427–435, 2000, ISSN: 03088146. DOI: 10.1016/S0308-8146(00)00127-8.

[116] R. J. Weston, K. R. Mitchell and K. L. Allen, 'Antibacterial phenolic components of New Zealand manuka honey', *Food chemistry*, vol. 64, no. 3, pp. 295–301, 1999, ISSN: 03088146. DOI: 10.1016/S0308-8146(98)00100-9.

[117] B. G. Visavadia, J. Honeysett and M. H. Danford, 'Manuka honey dressing: An effective treatment for chronic wound infections', *British Journal of Oral and Maxillofacial Surgery*, vol. 46, no. 1, pp. 55–56, 2008, ISSN: 02664356. DOI: 10.1016/j.bjoms.2006.09.013.

[118] National Institute of Advanced Industrial Science and Technology, *SDBSWeb*. [Online]. Available: https://sdbs.db.aist.go.jp (visited on 03/09/2021).

[119] B. Olas, 'Honey and its phenolic compounds as an effective natural medicine for cardiovascular diseases in humans?', *Nutrients*, vol. 12, no. 2, pp. 1–14, 2020, ISSN: 20726643. DOI: 10.3390/nu12020283.

[120] Y. Cheung, M. Meenu, X. Yu and B. Xu, 'Phenolic acids and flavonoids profiles of commercial honey from different floral sources and geographic sources', *International Journal of Food Properties*, vol. 22, no. 1, pp. 290–308, 2019, ISSN: 15322386. DOI: 10.1080/10942912.2019.1579835.

[121] G. Aronne, M. Giovanetti, R. Sacchi and V. De Micco, 'From flower to honey bouquet: Possible markers for the botanical origin of robinia honey', *Scientific World Journal*, vol. 2014, 2014, ISSN: 1537744X. DOI: 10.1155/2014/547275.

[122] P. J. Wood, I. R. Siddiqui and J. Weisz, 'Determination of Glucose and Fructose in Honey', *Journal of Apicultural Research*, vol. 14, no. 1, pp. 41–45, 1975. DOI: 10.1080/00218839.1975.11099799.

[123] M. S. Nurul Syazana, S. H. Gan, A. S. Halim, N. S. M. Shah and H. A. Sukari, 'Analysis of volatile compounds of Malaysian Tualang (Koompassia Excelsa) honey using gas chromatography mass spectrometry', *African Journal of Traditional, Complementary and Alternative Medicines*, vol. 10, no. 2, pp. 180–188, 2013, ISSN: 01896016. DOI: 10.4314/AJTCAM.V10I2.2.

[124] O. Sids and G. Acid, 'Introduction Gluconic Acid and Its Derivatives', *History*, 2020.

[125] G. W. Bruns and R. A. Currie, 'Determination of 2-chloroethanol in honey, beeswax, and pollen', *Journal of the Association of Official Analytical Chemists*, vol. 66, no. 3, pp. 659–662, 1983.

[126] H. A. Ghramh, E. H. Ibrahim and M. Kilany, 'Study of anticancer, antimicrobial, immunomodulatory, and silver nanoparticles production by Sidr honey from three different sources', *Food Science and Nutrition*, vol. 8, no. 1, pp. 445–455, 2020, ISSN: 20487177. DOI: 10.1002/fsn3.1328.

[127] C. E. Manyi-Loh, R. N. Ndip and A. M. Clarke, 'Volatile compounds in honey: A review on their involvement in aroma, botanical origin determination and potential biomedical activities', *International Journal of Molecular Sciences*, vol. 12, no. 12, pp. 9514–9532, 2011, ISSN: 14220067. DOI: 10.3390/ijms12129514.

[128] C. Auguistine Awasum, 'Gas Chromatography-Mass Spectroscopy Analysis and Chemical Composition of Ngaoundere, Cameroon Honey', *American Journal of Bioscience and Bioengineering*, vol. 3, no. 5, p. 33, 2015, ISSN: 2328-5885. DOI: 10.11648/j.bio.20150305.11.

[129] L. M. Nijssen, C. A. Visscher, H. Maarse, L. C. Willemsens and M. H. Boelens, *Volatile compounds in food : qualitative and quantitative data*, Zeist Netherlands. (visited on 26/04/2022).

[130] M. Pontes, J. C. Marques and J. S. Câmara, 'Screening of volatile composition from Portuguese multifloral honeys using headspace solid-phase microextraction-gas chromatography-quadrupole mass spectrometry', *Talanta*, vol. 74, no. 1, pp. 91–103, 2007, ISSN: 00399140. DOI: 10.1016/j.talanta.2007.05.037.

[131] T. Skov, F. Van Den Berg, G. Tomasi and R. Bro, 'Automated alignment of chromatographic data', *Journal of Chemometrics*, vol. 20, no. 11-12, pp. 484–497, 2006, ISSN: 08869383. DOI: 10.1002/cem.1031.

[132] L. Finos and L. Salmaso, 'Weighted methods controlling the multiplicity when the number of variables is much higher than the number of observations', *Journal of Nonparametric Statistics*, vol. 18, no. 2, pp. 245–261, 2006, ISSN: 10485252. DOI: 10.1080/10485250600720803.

[133] M. O. Okuom, M. V. Wilson, A. Jackson and A. E. Holmes, 'Intermolecular Interactions between Eosin Y and Caffeine Using 1 H-NMR Spectroscopy', *International Journal of Spectroscopy*, vol. 2013, pp. 1–6, Nov. 2013, ISSN: 1687-9449. DOI: 10.1155/2013/245376.

[134] C. Severini, A. Derossi, I. Ricci, R. Caporizzi and A. Fiore, 'Roasting Conditions, Grinding Level and Brewing Method Highly Affect the Healthy Benefits of A Coffee Cup', *International Journal of Clinical Nutrition & Dietetics*, vol. 4, no. 1, 2018. DOI: 10.15344/2456-8171/2018/127.

[135] R. A. Fadri, K. Sayuti, N. Nazir and I. Suliansyah, 'Analysis of Caffeine Levels in the Beverages of Roasted Arabica Coffee Balango in Bukik Apik with the Method of Spectroscopic', *IOP Conference Series: Earth and Environmental Science*, vol. 515, no. 1, 2020, ISSN: 17551315. DOI: 10.1088/1755-1315/515/1/012071.

[136] C. Petisca, T. Pérez-Palacios, A. Farah, O. Pinho and I. M. Ferreira, 'Furans and other volatile compounds in ground roasted and espresso coffee using headspace solid-phase microextraction: Effect of roasting speed', *Food and Bioproducts Processing*, vol. 91, no. 3, pp. 233–241, 2013, ISSN: 09603085. DOI: 10.1016/j.fbp.2012.10.003.

[137] M. Gancarz, B. Dobrzański, U. Malaga-Toboła *et al.*, 'Impact of Coffee Bean Roasting on the Content of Pyridines Determined by Analysis of Volatile Organic Compounds', *Molecules*, vol. 27, no. 5, pp. 1–12, 2022, ISSN: 14203049. DOI: 10.3390/molecules27051559.

[138] G. Greco, E. Núñez-Carmona, M. Abbatangelo, P. Fava and V. Sberveglieri, 'How coffee capsules affect the volatilome in espresso coffee', *Separations*, vol. 8, no. 12, 2021, ISSN: 22978739. DOI: 10.3390/separations8120248.

[139] K. Król, M. Gantner, A. Tatarak and E. Hallmann, 'The content of polyphenols in coffee beans as roasting, origin and storage effect', *European Food Research and Technology*, vol. 246, no. 1, pp. 33–39, 2020, ISSN: 14382385. DOI: 10.1007/s00217-019-03388-9.

[140] G. N. Jham, S. A. Fernandes, C. F. Garcia and A. Araujo Da Silva, 'Comparison of GC and HPLC for the quantification of organic acids in coffee', *Phytochemical Analysis*, vol. 13, no. 2, pp. 99–104, 2002, ISSN: 09580344. DOI: 10.1002/pca.629.

[141] H. Horai, M. Arita, S. Kanaya *et al.*, 'MassBank: a public repository for sharing mass spectral data for life sciences', *Journal of Mass Spectrometry*, vol. 45, no. 7, pp. 703–714, 2010. [Online]. Available: https://analyticalsciencejournals.onlinelibrary.wiley.com/doi/abs/10.1002/jms.1777 (visited on 17/09/2021).

[142] T. Tian, S. Freeman, M. Corey, J. B. German and D. Barile, 'Effect of Roasting on Oligosaccharide Abundance in Arabica Coffee Beans', *Journal of Agricultural and Food Chemistry*, vol. 66, no. 38, pp. 10 067–10 076, 2018, ISSN: 15205118. DOI: 10.1021/acs.jafc.8b02641.

[143] R. B. Gurung, E. H. Kim, T. J. Oh and J. K. Sohng, 'Enzymatic synthesis of apigenin glucosides by glucosyltransferase (YjiC) from Bacillus licheniformis DSM 13', *Molecules and Cells*, vol. 36, no. 4, pp. 355–361, 2013, ISSN: 10168478. DOI: 10.1007/s10059-013-0164-0.

[144] L. Buckel, J. I. Kremer, S. Stegmüller and E. Richling, 'Fast, Sensitive and Robust Determination of Nicotinic Acid (Vitamin B3) Contents in Coffee Beverages Depending on the Degree of Roasting and Brewing Technique', *Proceedings*, vol. 11, no. 1, p. 13, 2019, ISSN: 2504-3900. DOI: 10.3390/proceedings2019011013.

[145] L. Zhou, A. Khalil, F. Bindler, M. Zhao, C. Marcic, S. Ennahar and E. Marchioni, 'Effect of heat treatment on the content of individual phospholipids in coffee beans', *Food Chemistry*, vol. 141, no. 4, pp. 3846–3850, 2013, ISSN: 18737072. DOI: 10.1016/j.foodchem.2013.06.056.

[146] E. J. Kirkland, 'Bilinear interpolation', in *Advanced Computing in Electron Microscopy*, Springer, 2010, pp. 261–263.

[147] A. Bhattacharyya, 'On a Measure of Divergence between Two Multinomial Populations', *Sankhyā: The Indian Journal of Statistics (1933-1960)*, vol. 7, no. 4, pp. 401–406, May 1946, ISSN: 00364452.

[148] J. Zapata, V. Londoño, M. Naranjo, J. Osorio, C. Lopez and M. Quintero, 'Characterization of aroma compounds present in an industrial recovery concentrate of coffee flavor', *CYTA - Journal of Food*, vol. 16, no. 1, pp. 367–372, 2018, ISSN: 19476345. DOI: 10.1080/19476337.2017.1406995.

[149] W. Baltes and G. Bochmann, 'Model reactions on roast aroma formation', *Zeitschrift für Lebensmittel-Untersuchung und Forschung 1987 185:1*, vol. 185, no. 1, pp. 5–9, Jul. 1987, ISSN: 1438-2385. DOI: 10.1007/BF01083331.

[150] E. B. Hughes and R. F. Smith, 'The nicotinic acid content of coffee', *Journal of the Society of Chemical Industry*, vol. 65, no. 10, pp. 284–286, 1946.

[151] A. Yakugaku Zasshi, M. Abstracts of Papers, W. Baltes and G. Bochmann, 'Model reactions on roast aroma formation. 1. Reaction of serine and threonine with sucrose under the conditions on coffee roasting and identification of new coffee aroma compounds', *Journal of Agricultural and Food Chemistry*, vol. 35, no. 3, pp. 340–346, 1987.

[152] M. Gancarz, B. Dobrzański, U. Malaga-Toboła *et al.*, 'Impact of Coffee Bean Roasting on the Content of Pyridines Determined by Analysis of Volatile Organic Compounds', *Molecules*, vol. 27, no. 5, 2022, ISSN: 14203049. DOI: 10.3390/molecules27051559.

[153] S. Adisakwattana, 'Cinnamic acid and its derivatives: Mechanisms for prevention and management of diabetes and its complications', *Nutrients*, vol. 9, no. 2, 2017, ISSN: 20726643. DOI: 10.3390/nu9020163.

List of Abbreviations

AB Adaptive Binning.

COW Correlation Optimised Warping.

CWT Continuous Wavelet Transformation.

DRY Driver-based Statistical Heterospectroscopy.

DTW Dynamic Time Warping.

DWT Discrete Wavelet Transformation.

FID Free Induction Delay.

GC Gas Chromatography.

GC-MS Gas Chromatography - Mass Spectrometry.

HPLC High Performance Liquid Chromatography.

LC Liquid Chromatography.

LC-MS Liquid Chromatography - Mass Spectrometry.

LDA Linear Discriminant Analysis.

LOO Leave-One-Out.

LV Latent Variables.

MDA Mean Decrease in Accuracy.

MDG Mean Decrease in Gini Index.

MS Mass Spectrometry.

MSFE Multistage Feature Extraction.

NIPALS Non-Linear iterative partial least squares.

NMR Nuclear Magnetic Resonance.

PC Principal Component.

PCA Principal Component Analysis.

PLS Partial Least Squares / Projection on Latent Spaces.

QC Quality Control.

RF Random Forest.

RMSEP Root-mean-square Error of Prediction.

SHY Statistical Heterospectroscopy.

SNR Signal-to-Noise Ratio.

STOCSY Statistical Total Correlation Spectroscopy.

SVD Singular Value Decomposition.

TSP Trimethylsilylpropanoic acid.

UV Unit Variance.

VIP Variable Importance in Projection.

WT Wavelet Transformation.