Applications of Multivariate Statistics in Honey Bee Research: analysis of metabolomics data from samples of honey bee propolis

Abdulaziz Saleh Alghamdi

Submitted in accordance with the requirements for the degree of Doctor of Philosophy

Department of Mathematics & Statistics
University of Strathclyde, Glasgow

February 28, 2020

This thesis is the result of the author's original research. It has been composed by the author and has not been previously submitted for examination which has led to the award of a degree.

Signed:

Date:

This thesis is dedicated to my family and my supervisors for never doubting my abilities, even when I doubt myself.

Abstract

Honey bees play a significant role both ecologically and economically, through the pollination of flowering plants and crops. Additionally, honey is an ancient food source that is highly valued by different religions and cultures and has been shown to possess a wide range of beneficial uses, including cosmetic treatment, eye disease, bronchial asthma and hiccups. In addition to honey, honey bees also produce beeswax, pollen, royal jelly and propolis. In this thesis, data is studied which comes from samples of propolis from various geographical locations.

Propolis is a resinous product, which consists of a combination of beeswax, saliva and resins that have been gathered by honey bees from the exudates of various surrounding plants. It is used by the bees to seal small gaps and maintain the hives, but is also an anti-microbial substance that may protect them against disease. The appearance and consistency of propolis changes depending on the temperature; it becomes elastic and sticky when warm, but hard and brittle when cold. Furthermore, its composition and colour varies from yellowish-green to dark brown, depending on its age and the sources of resin from the environment. Propolis is a highly biochemically active substance with many potential benefits in health care, which have attracted much attention.

Biochemical analysis of propolis leads to highly multivariate metabolomics data. The main benefit of metabolomics is to generate a spectrum, in which peaks correspond to different chemical components, making possible the detection of multiple substances simultaneously. Relevant spectral features may be used for pattern recognition. The purpose of this research is to study methods used for statistical analysis of biochemical data arising from propolis samples.

We investigate the use of different statistical methods for metabolomics data from chemical analysis of propolis samples using Mass Spectrometry (MS). Methods studied will include pre-treatment methods and multivariate analysis techniques including principal component analysis (PCA), multidimensional scaling (MDS), and clustering methods including hierarchical cluster analysis (HCA), k-means clustering and self organising maps (SOMs). Background material and results of data analysis will be presented from samples of propolis from beehives in Scotland, Libya and Europe. Conclusions are drawn in terms of the data sets themselves as well as the properties of the different methods studied for analysing such metabolomics data.

Contents

List of Figures

List of Tables

Acknowledgements

This thesis would not have been possible without the help of my supervisor. Dr Alison Gray has always been the main source of support and confidence in my study and I have relied on her constant encouragement throughout my academic career. Her kindness and support for me will always be a great source of inspiration in my life. Many thanks go to Dr. David Watson for providing the metabolomics data from propolis samples used in the research in this project, and its pre-processing and general preparation. Thank you to all of my family, for the love, support, understanding, happiness and help in everything in my life. I am grateful to my children, who make me smile wherever I look at them. Finally, I would like to thank King Abdulaziz University through the Department of Mathematics and Statistics in the Faculty of Science for funding this PhD study.

Papers published and posters presented

- Alghamdi, Abdulaziz and Gray, Alison (2017, September). Multivariate statistics for analysis of honey bee propolis. Poster abstract. Proceedings of the Royal Statistical Society Conference 2017, Glasgow.
- Alghamdi, Abdulaziz and Gray, Alison (2019, June). Applications of multivariate statistics in honey bee reasearch. Poster abstract in DSMS19, 1st Doctoral School Multidisciplinary Symposium, University of Strathclyde.
- Alghamdi, Abdulaziz, Gray, Alison and Watson, David (2019). Investigation of metabolomics techniques by analysis of MS propolis data: which pre-treatment method is better? Advances and Applications in Statistics, 58 (1), 13-34.

Chapter 1

Aims and Outline

This chapter has two purposes: (i) introduce the main problem with a brief explanation, and (ii) to outline the thesis.

The purpose of this research is to investigate statistical techniques that can be used in the analysis of metabolomics data. More specifically, the aim is to assess the ability of various clustering techniques and other multivariate methods to analyse metabolomics data by exploring the metabolic profiles of propolis samples. This investigation aims to confirm whether these clustering techniques can be used to identify any natural groupings in data such as those consisting of metabolic profiles.

This thesis is structured as follows. Chapter 2 provides background material on propolis and its chemical analysis. Chapter 2 discusses methods of analysis of propolis samples. In this chapter, mass spectrometry (MS) and liquid chromatography are discussed in detail. In Chapter 3, metabolomics and analysis techniques, and multivariate analysis methods used on metabolomics data are described. In addition, we use two different data sets from propolis samples throughout this thesis, where the first data set contains three sub-sets of data from Scotland (Aberdeenshire, Fort William and Dunblane) and the second data set is from propolis samples from Libya. These will be described in this chapter.

Chapter 4 discusses pre-precessing of raw data. Also, the impact of noise and convolution

are discussed. On the other hand, several methods of pre-treatment are also discussed, such as transformation and scaling methods.

Chapter 5 contains unsupervised techniques, where data reduction is introduced and several applications of PCA in metabolomics are studied. Chapter 6 examines another data-projection method for reducing dimensionality, multidimensional scaling (MDS), with the advantage over PCA that it is flexible and can be used with any dissimilarity measure. It can be also applied to metabolomics data sets as in PCA.

In Chapter 7, clustering similarity or difference techniques are reviewed such as hierarchical clustering. Section 7.3 discusses similarity or difference measures, which this approach reviews, while Section 7.4 covers hierarchical clustering with emphasis on agglomerative nesting algorithms.

In Chapter 8, another category of optimal partitioning methods called hard clustering algorithms are reviewed, and the k-means algorithm is applied to all the propolis data sets.

Competitive learning algorithms are described in Chapter 9, with emphasis on self-organising maps (SOM), a statistical approach, and its application to all the propolis data sets.

Chapter 10 presents a case study, in which the best methods identified are applied to a further data set, of samples from Europe, the results are compared, and conclusions are drown about the data.

Chapter 11 provides a summary and conclusions, including approaches, results and advantages and disadvantages of the methods, as well as suggestions for further work.

Finally, all of these techniques have been applied to all the metabolomics data to assess their effectiveness in reducing the dimensionality of the input space or for the effective and efficient clustering of the data, with a view to determining the best approaches for such metabolomics data, as well as uncovering any interesting patterns in these particular data sets.

Chapter 2

Background: Propolis and Chemical Analysis of Propolis Samples

Writing from over 1400 years ago in the Quran refers to bees that generate the honey as females (the Arabic grammar is in the female mode): [Quran 16, verses 68-69] and your Lord (Allah) revealed to the bees: Build your hives in mountains, trees and in what they build. The Quran used "Kuli" (females).

In the name of of Allah the Merciful

﴿ وَأَوْحَىٰ رَبُّكَ إِلَى النَّحْلِ أَنِ اتَّخِذِي مِنَ الْجِبَالِ بُيُوتًا وَمِنَ الشَّجَرِ وَمِمَّا يَعْرِشُونَ (68) ثُمَّ كُلِي مِن كُلِّ الثَّمَرَاتِ فَاسْلُكِي سُبُلَ رَبِّكِ ذُلُلًا يَخْرُجُ مِن بُطُونِهَا شَرَابٌ مُّخْتَلِفٌ أَلْوَانُهُ فِيهِ شِفَاءٌ لِّلنَّاسِ إِنَّ فِي ذَٰلِكَ لَآيَةً لِّقَوْمٍ يَتَفَكَّرُونَ (69) ﴾

[القران الكريم سورة النحل اية 68-69]

(68) And your Lord inspired the bee: "Set up hives in the mountains, and in the trees, and in what they construct" (69) "Then eat of all the fruits, and go along the pathways of your Lord, with precision. From their bellies emerges a fluid of diverse colors, containing healing

for the people. Surely in this is a sign for people who reflect". [Quran, surah al-nahl 16, verses 68-69].
Bees are well known for producing honey, as well as for pollinating crops, but there are other products of the honey bee colony. In this thesis we focus on propolis, described below.

2.1 Overview

This chapter is divided into four parts: the first part, in Section 2.2, gives background about propolis, and Section 2.3 describes extraction of propolis. Sections 2.4 describes methods of analysis of propolis samples. Finally, conclusions are written in Section 2.5.

2.2 Introduction to propolis and its properties

In economically advanced countries, health professionals favour conventional medicine over natural products, even though the latter can offer a suitable alternative for some conditions. These are of particular value in locations where conventional therapy is not readily available. Bee products are one such natural alternative that have long been used in traditional medicine in some parts of Africa, East Asia and South America (El-Soud, 2012), as well as in Eastern Europe. Bees have been on Earth for millions of years and the fact that their species continues to persist is evidence of their evolutionary success. This success can be attributed to their ability to exploit the chemistry of substances in their environment and use these for their own products: honey, beeswax, pollen, royal jelly, venom and propolis. Because of their biological potential, bee products can be considered as functional foods.

2.2.1 Composition of propolis and its uses in the hive

Propolis is one of the most interesting bee compounds, that honey bees manufacture by combining various quantities of beeswax with resins. Honey bees harvest resins from flower buds, flowers and the bark of particular plants, shrubs and trees, which they mix with beeswax. This multifunctional compound is considered to maintain the hives, by being used as a construction material and sealing open spaces in the hive, as well as being an anti-infective substance offering defence against disease (Bankova, 2005; Bankova et al., 2016; Bertelli et al., 2012; Burdock, 1998). The word propolis is a compound term originating with ancient Greek roots: "pro" is believed to relate to 'defence', and "polis" means 'city'. Therefore, the name propolis describes defending the city or, in this instance, the hive (Ghisalberti, 1979, as reviewed by Burdock, 1998) (Bankova, 2005; Kasiotis et al., 2017). Owing to its glue-like nature, propolis is frequently described as 'bee glue'. Bees are also protected by propolis through its anti-bacterial and anti-fungal properties, conferring the hive with defence against diseases caused by fungi or bacteria.

The appearance of propolis is sticky and gum-like; this is consistent with being highly resinous. It is a hydrophobic compound that undergoes temperature-related changes. When it is cold, propolis becomes inelastic and rigid, but when warm is pliant and gluey (Hausen et al., 1987). Usually, propolis liquefies when it is between 60° C and 70° C, though some specimens are reported to remain solid at these temperatures and not melt until temperatures reach 100° C (Kuropatnicki et al., 2013). The composition of propolis varies between hives, districts and seasons (Toreti et al., 2013). Variation in the composition is also reflected by variations in its colour, which can range from yellowish-green to dark brown, depending on the botanical source of the resins (Kuropatnicki et al., 2013; Marcucci, 1995) (Figure 2.1[1]).

[1]Sources: $(i) www.naturaletz.com/_img/ft_propolis.gif.$

(ii) $www.soorganic.com/blog/a-propolis-buzz-705.html.$

In summary, the benefits of propolis are:

1. To promote the defensibility of the hive by sealing alternative entrances.
2. To protect the colony against disease and parasites; inhibit bacterial and fungal growth (Cremer et al., 2007).
3. Minimise putrefaction within the hive. Usually, bees are fastidious, cleaning waste and removing it away from the hive. Any insect that finds its way into the hive, but fails to find its way out and dies, presents bees with a body that may be too difficult to remove from the hive. In such instances, the bees will instead seal the body in propolis; this form of mummification effectively makes the body scentless and harmless (Qureshi et al., 2014).

2.2.2 Medical and other uses of propolis

Since ancient times, propolis has been used by humans as a remedy for various aliments. Today, in the Balkan states, propolis continues to be used as a traditional treatment for burns, dental caries, sore throats, stomach ulcers, wounds and other ailments (Wollenweber et al., 1990). In Europe and North Africa, this is demonstrated by the very Greek name of propolis. The ancient Egyptians were familiar with the unique wound healing properties of propolis and used it to preserve their dead (Lotfy, 2006). Also, the Arabs, Incas and Romans used propolis to treat fever. Four centuries ago, it was defined in the London Pharmacopoeias as being a certified drug (Sforcin and Bankova, 2011).

In the writings of Bin Sina (Avicenna), two forms of beeswax are described; the first he noted was 'clean' and the other he referred to as 'black wax', which is assuredly propo-

(iii) $http://www.mofaid.com/an/images/propolis.jpg$.

(iv) $http://commons.wikimedia.org/wiki/File:Propolis$.

(v) $www.made-in-china.com/image/2f0j00PBsaQLdGJRgKM/Propolis.jpg$.

(i)

(ii)

(iii)

(iv)

(v)

Figure 2.1: Different propolis morphology and colours: (i) Light brown hard propolis, (ii) Yellow hard propolis, (iii) Fragile brown propolis, (iv) Dark brown propolis, (v) Black hard propolis.

lis (Lotfy, 2006). In contrast to some customary preservatives, the effect of propolis upon human health is generally regarded as being beneficial. The constituents of propolis are usually standard constituents of food and/or food additives, which are typically considered safe to humans. Results of the research undertaken by Burdock (1998) support earlier proposals to use propolis in the food industry. Mizuno (1989) suggested the germicide and insecticide qualities of propolis be exploited, by using it as a preservative component in the protective food material. Furthermore, Han and Park (1995) proposed propolis be used in the preservation of meat products.

Since ancient times, propolis has been applied for medical purposes. Banskota et al. (2001) report that as far back as 300 BC, propolis was used as a traditional medicine to treat wound healing and inflammation; it was also used as a cosmetic. Because of its antibacterial, antifungal and antiviral properties, folk medicine applies it both internally and externally to kill bacteria, fungi and viruses, and to treat inflammation and ulcers (Banskota et al., 2001; Lotfy, 2006). Propolis is also reputed to lower blood pressure and be an immune system stimulant. According to Cuesta et al. (2012), propolis products are used by the body to produce energy and maintain. For more than 2000 years, cultures including Asian, European and Middle Eastern have used propolis to kill microbes and to treat aggravated wounds like bedsores and diabetic ulcers.

2.2.3 Biochemical properties and study of propolis

The propolis most widely researched over the past ten years is temperate propolis. Table 2.1 shows the typical constituents of temperate propolis; approximately 55% of the total is made of balsams and resinous substances, wax constitutes about 30%, 10% are aromatic and essential oils, 5% is pollen, and the remaining 5% is made up of organic debris (Burdock, 1998). However, as indicated earlier, the actual composition of propolis is variable and reflects the diversity of the plants, the season of harvesting and geographical location

in which the honey bees forage to gather the materials for propolis (Crane et al., 1990).

A number of studies have explored the biological activity, chemical data and molecu-

Ingredient	Group of Ingredients	Amount
Resins	Flavonoids, Esters and Phenolic Acids	45-55%
Waxes and Fatty Acids	Beeswax and Plant Origin	23-35%
Essential Oils	Volatiles	10%
Pollen	Proteins (16 free amino acids>1%)	5%
Other Organics and Minerals	Arginine and Proline, together 46% of total 14 traces of minerals, zinc and iron most common; lactones, quinines, ketenes, steroids, benzoic acid, sugar and vitamins	5%

Table 2.1: Common compounds occurring in raw propolis (Bankova et al., 2016).

lar structures of different components isolated from propolis. Modern researchers have applied advanced technology such as mass spectrometry (MS) and nuclear magnetic resonance (NMR) to identify components, in addition to using gas chromatography (GS) and medium pressure liquid chromatography (MPLC) to separate, analyse and purify the individual constituents of propolis. Some of the chemical features that have been characterised include hydrocarbons, flavonoids, phenolics, terpenes as well as mineral elements (Inui et al., 2012; Oliveira et al., 2010; Petrova et al., 2010).

Propolis manufactured by honey bees, which is valued for its antimicrobial qualities and has been used as a traditional pharmaceutical, is highly biochemically active (Sforcin et al., 2000). Therefore, propolis and extracts from it have been the subject of several studies in which researchers have analysed the compound's antibacterial action against Gram-positive and Gram-negative pathogens. The results of these studies reveal that propolis is active against a wide variety of Gram-positive bacteria; however, it only ex-

erts a limited reaction against Gram-negative bacilli (Cuesta et al., 2012; Lu et al., 2005; Sforcin et al., 2000). The findings obtained by Seidel et al. (2008) reinforced earlier results, although the study focused on comparing the antibacterial activity of propolis collected from different countries spanning tropical, subtropical and temperate climate zones. The diverse propolis was tested for effectiveness against six Gram-positive and two Gram-negative microorganisms. What the study revealed was that propolis manufactured by bees from windy, tropical locations exerted the most significant antibacterial effect. Subsequent research into some of the propolis constituents of the samples which Seidel et al. (2008) collected in the Solomon Islands showed inhibitory effects against methicillin-resistant Staphylococcus aureus (MRSA) (Raghukumar et al., 2010).

Other research has examined the cytotoxic effects of propolis and its constituents upon inhibiting tumour cell growth. Constituent analysis of propolis reveals that the active compounds include caffeic acid phenethyl ester, flavonoids and terpenes (Wagh, 2013). It has also been demonstrated that using the propolis extracts in treating cancer can reduce the overall costs of cancer treatment (Banskota et al., 2001). Because of the antineoplastic constituents of propolis, which are capable of killing cancerous cells, these propolis extracts have been used to reduce the tumour activity of cancerous cells (Lu et al., 2005). Several propolis compounds have also been evaluated for effectiveness in being used for chemopreventive purposes. Research into the scope and effectiveness of propolis extracts for use as anti-cancer interventions is still ongoing.

The benefits of propolis to various areas of medicine, such as its use as an antioxidant, a molecule that can protect other molecules from the damage of oxidation radicals, have been described by da Cunha et al. (2013). Because propolis has an antioxidative capability, it and its derivatives offer potential in being used as preservatives (Banskota et al., 2001). As mentioned previously, propolis and its constituent components of caffeic acid, cinnamic acid and ferulic acid have antifungal capabilities that able to exert effects against

Candida albicans. Even parasitic diseases are responsive to propolis, with reports of the compound effectively curing such infections.

However, this wonder compound is not without drawbacks, and some people may exhibit allergic reactions to propolis, presenting with irritation of mucous membranes and skin (Lotfy, 2006). Nonetheless, the antibacterial and antifungal capacity of propolis (Banskota et al., 2001; Burdock, 1998; Marcucci, 1995) and its ability to promote tissue re-modelling have led to high demand for propolis extracts to be regularly incorporated into dermatological and cosmetic treatments (Gulcin et al., 2010). It was also recorded that a level of sunlight protection can be supplied by ethanolic extracts of propolis and its components in addition to strong antioxidant activity (Gregoris et al., 2011). Together, these characteristics identify propolis as a potential active ingredient to include in cosmetics.

2.3 Extraction of Propolis

In its crude state, propolis is considered unsuitable for use in cosmetics, foods or medicine; it needs to go through a number of purification processes that remove unwanted components, such as wax. The residual material is a flavonoid-rich concentrate and the flavonoids are responsible for the biological activity of propolis (Vaher and Koel, 2003). Due to the wax component, experimental analyses of propolis also need to be done using solvents to separate the raw materials. The preferred extraction procedure uses ethanol to isolate the waxy materials, leaving a residue of polyphenolic compounds. Extraction often uses 70% or 80% ethanol, as it is particularly effective in separating propolis to isolate the extracts that are rich in polyphenol compounds (Vaher and Koel, 2003).

The next section addresses the methods used to analyse samples of propolis.

2.4 Methods of Analysis of Propolis Samples

2.4.1 Introduction

As described earlier, propolis is a complex substance manufactured by honey bees using wax and secondary metabolites of a plant root. To identify the composition of propolis, it has been subjected to several high-tech chemical analyses, including high-performance liquid chromatography (HPLC), mass spectrometry (MS) and liquid chromatography-mass spectrometry (LC-MS). Due to the diversity of propolis compositions arising from the geographical variations in the raw materials used in its manufacture, it is not practical to use a single-instrument strategy for analysis. Therefore, different chemical techniques are used, such as LC-MS, as it is effective to analyse the various flavonoid components of propolis. This technology is considered a reliable and most adaptable procedure for analysing the quality of different propolis samples (Ivanauskas et al., 2008). Some propolis samples with compounds, such as terpenoids, that have been challenging to characterise using Ultraviolet-Visible (UV) spectrophotometry, have been examined using gas chromatography-mass spectrometry (GC-MS) in conjunction with HPLC (Gardana et al., 2007; Vaher and Koel, 2003). To identify propolis's phenolic compounds, thin layer chromatography (TLC) has been used. Other techniques that have been used include atmospheric pressure ionisation and electrospray ionisation mass spectrometry (ESI-MS) as this enables the typical 'fingerprints' of complex materials to be characterised (Volpi, 2004). However, analyses of the composition of propolis are generally performed using traditional phytochemical methods, such as chromatographic and spectroscopic techniques, which are capable of isolating and identifying the individual constituents. Mass spectrometry is employed for structural determination, and is described below.

The purpose of mass spectrometry is to measure atoms and molecules to calculate their molecular weight. The information of mass or weight data is sometimes enough, often

essential, and always valuable in identifying species. Using mass spectrometers, the chemical structure of the mass can be established. Typically, these instruments are employed in industrial and academic research settings. A mass spectrometer creates charged particles from molecules. These particles are investigated for the purpose of determining the molecular weight of the mass and its chemical structure. Mass spectrometry has applications across diverse domains including biotechnology, clinical, environment and geology and pharmaceutics. It is also useful for metabolite fingerprinting, or metabolome analysis, by creating a spectral 'fingerprint' of the metabolites produced by a sample; this enables particular metabolites, such as sulfides, hormones and vitamins to be explored.

2.4.2 Introduction of Mass Spectrometry(MS)

Mass spectrometry was invented by the English physicist and physics Nobel laureate, Joseph John Thomson, in 1897 (Go et al., 2007). Working at Cambridge University, he explored electrical discharges in gases. During the 20th century, the concept of Thomson's initial mass spectrometer was advanced by Aston, to enable analysis of isotopes; this technology was further advanced by Dempster who developed the modern mass spectrometer that used an electron beam as the ion source to ionise volatile molecules. Between 1946 and 1953, four different contributions were made (Borman et al., 2003). In 1946, at the University of Pennsylvania, Stephens introduced the concept of Time of Flight (TOF) MS. A Time of Flight analyser is used to verify the mass of biomolecules as it has relatively boundless mass range. Ion Cyclotron Resonance (ICR) was first described in 1949 by Hipple, Sommer and Thomas (Hipple et al., 1949); this technique enables ions to be detected sequentially. M. B. Comisarow and A. G. Marshall went on to combine ICR with Fourier Transformations (FT) to arrive at FT-ICR MS, which facilitated the simultaneous measurement of multiple ions. In 1953, Nier and Johnson developed the double-focusing instrument that made investigating isotopes easier. At a similar time, the quadropole mass analyser was presented by Paul and Steinwedel (Borman et al.,

2003); this technology offered considerable stability over a dynamic range, conferring the technology with particular value in quantitatively assessing medications and drugs. The field of molecular investigation has benefitted from two key developments; the first is the Electrospray ionisation (ESI) strategy, which was described in 1968 by M. Dole. In spite of this reality, it was J. B. Fenn who connected this strategy out of the blue, in 1984, in biomolecular analysis. The second development was the creation of the matrix-assisted laser desorption/ionisation technique (MALDI). MALDI was first described in 1983 by two different research groups, Hillenkamp and Karas at the University of Frankfurt, and Tanaka at Shimadzu Corp (Xavier and Rauter, 2008).

Figure 2.2[2] depicts the key accomplishments in the development of mass spectrometry technology over the past century. In the last part of the 20th century, these MS technologies have been refined further, enabling coordinated exploration of pharmacokinetics, including analysis of small drug molecules, identifying proteins and mapping peptide mass. More recently mass spectrometry has been used in clinical investigations to screen neonates for over thirty diseases (Borman et al., 2003).

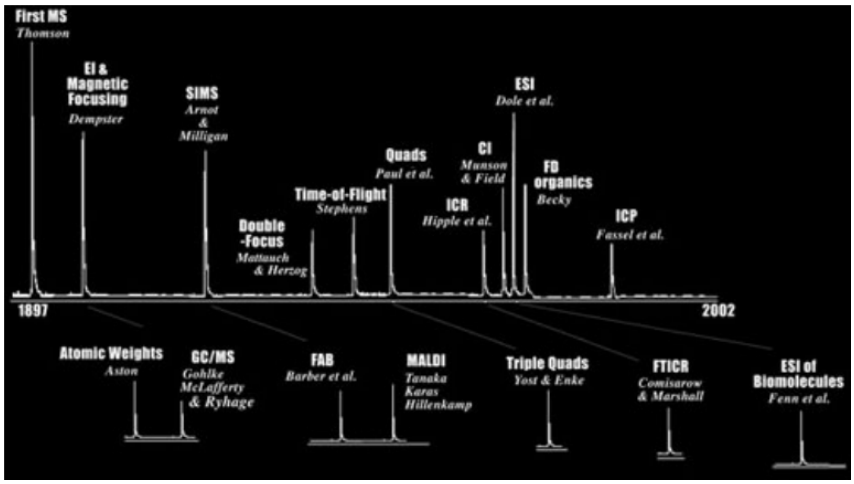


Figure 2.2: Important timelines and contributions to Mass Spectrometry

[2]Source: Scripps Centre for Metabolomics and Mass Spectrometry.

2.4.3 Description of Mass Spectrometers

There are three prime components of mass spectrometers; these are the ionisation source, the ion analyser and the detector. The first step in the analysis process is to embed the sample into the instrument's ioniser, ready for the ionisation of the molecules in the sample. Ions are easier to work with than uncharged molecules. The ions are collected by the mass spectrometer's ion analyser, which then separates them according to their mass (m) - to - charge (z) ratios. Then the ions that have been isolated are recorded by the instrument's detector, which generates a signal that is transmitted to an information system that stores the ratios of ions and their relative abundance. These data are presented in the form of an $\frac{m}{z}$ spectrum (Kang, 2012). The components of the mass spectrometer are typically maintained in a high vacuum to promote the potential of the ions travelling through the instrument without encountering air molecules, as they present an obstruction. Figure 2.3 shows the main components of a mass spectrometer.

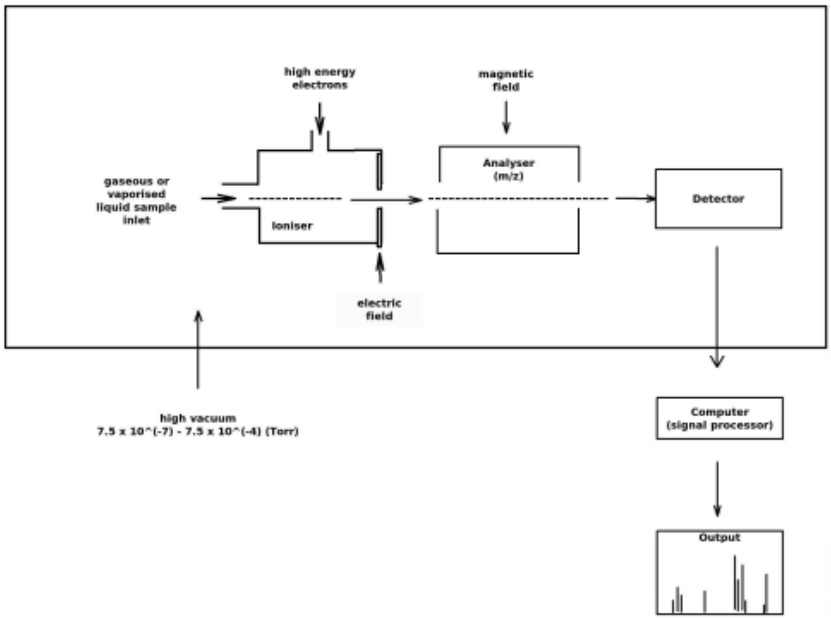


Figure 2.3: Schematic of the main components of a mass spectrometer (Prelorendjos, 2014).

The contents of the sample may be identifiable following ionisation, but often the sample needs to be directed through a chromatography instrument as it moves through the ionisation source. In that instance, the mass spectrometer is combined with a chromatography separation column, which makes the sample separate into its different components. Then the different components sequentially enter the instrument to be analysed individually. There are three forms of chromatography that are used most often:

1. Gas Chromatography (GC-MS) is mainly used to separate natural mixes that are volatile. The chromatograph is usually connected to covered, capillary columns, linked to a mass spectrometer (Kissinger, 2002; Williams and Fleming, 1995). The prime parts of a gas chromatograph are a flowing mobile phase, which is typically an inert gas, such as, nitrogen or helium, argon, an injection port, a separation column containing the stationary phase, a detector and an information recorder (Figure 2.4[3]).

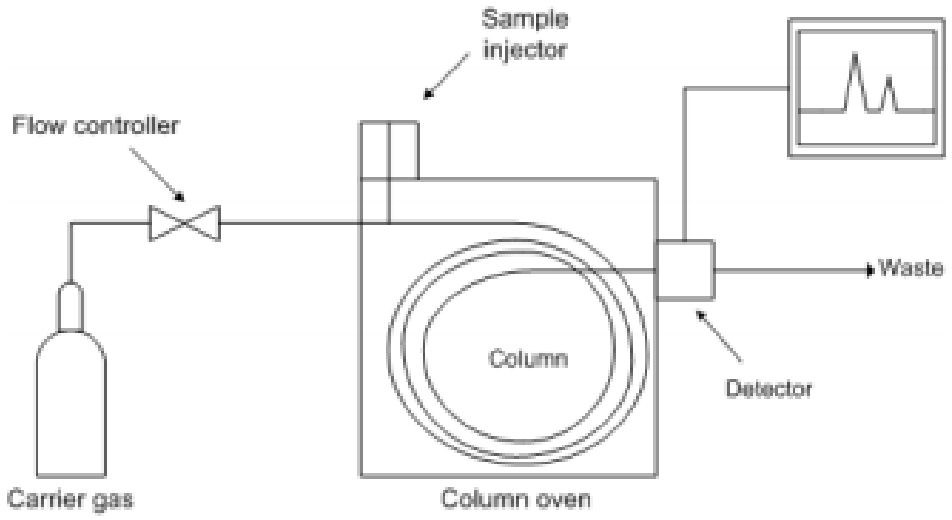


Figure 2.4: Diagram of a gas chromatography mass spectrometer (GC-MS)

2. Liquid Chromatography[4] (LC-MS) is generally used to isolate and purify compo-

[3]Source:http://upload.wikimedia.org/wikipedia/commons/8/87/Gas_chromatograph.png.

[4]Often high performance liquid chromatography (HPLC-MS) or ultra high pressure liquid chromatog-

nents within a mixture. Linked with mass spectrometry, LC can be used with all kinds of stationary phases, such as normal phase, reversed phase or ion exchange (Kissinger, 2002; Williams and Fleming, 1995). In this scenario, the analyte's chemical properties such as charge, hydrophilicity or hydrophobicity differentiate one from another. On account of more logical separations of solutions with the end goal of recognition or evaluation, more sophisticated instruments are required, such as High-Performance Liquid Chromatography (HPLC) or Ultra High-Performance Liquid Chromatography (UPLC) instruments. These technologies quickly provide high-resolution data for the samples under investigation. Figure 2.5[5] depicts the typical LC separation process.

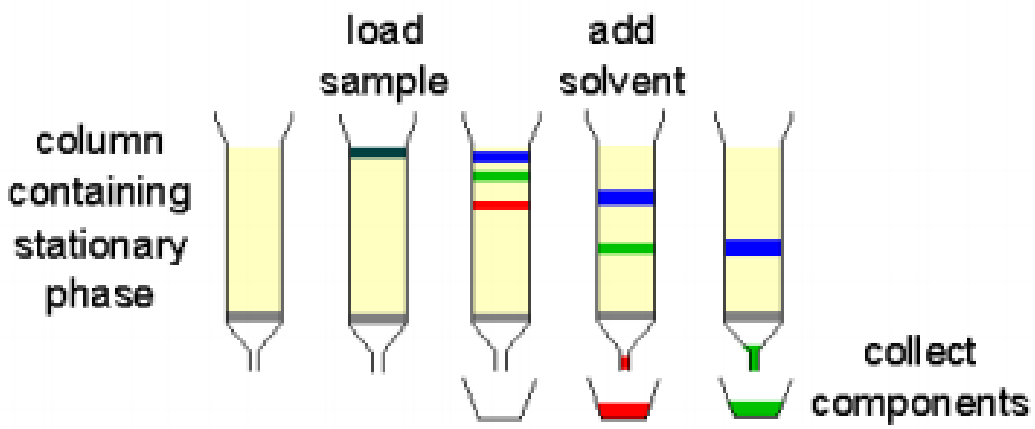


Figure 2.5: Liquid chromatography (LC) separation procedure

3. Medium pressure liquid chromatography (MPLC) is a preparative separation technique. It is used where large quantities of compounds need to be isolated from crude materials and purified before being subjected to additional techniques, such as HPLC. MPLC is suitable for this due to its low cost, high throughput capacity and high sample loading (Cheng et al., 2010). MPLC complements ash chromatog-

raphy (UPLC-MS) is needed.

[5]Source: http://www.chemistry.nmsu.edu/Instrumentation/lc-schem.gif.

raphy and the weight required is within the range of $5 - 20$ bars which is created through a cylinder pump with an adjustable rate. It can be differentiated from flash chromatography and other low weights techniques, as those techniques use low weights ranging from 1 to 5 bars (Weber et al., 2011). MPLC uses medium pressure that speeds up the rate at which samples elute through the column. Furthermore, the resolution is sufficiently effective that compounds that have a range of polarities can be separated from semi-purified samples. The size of samples that can be loaded in a single run can be up to 50 g. The fractions acquired can be further purified by re-chromatographing it; the reproducibility of the packing of columns and separation can all be attained (Claeson et al., 1993). A schematic representation of the key components of MPLC systems (Figure 2.6) and should be similar for any type of such equipment.

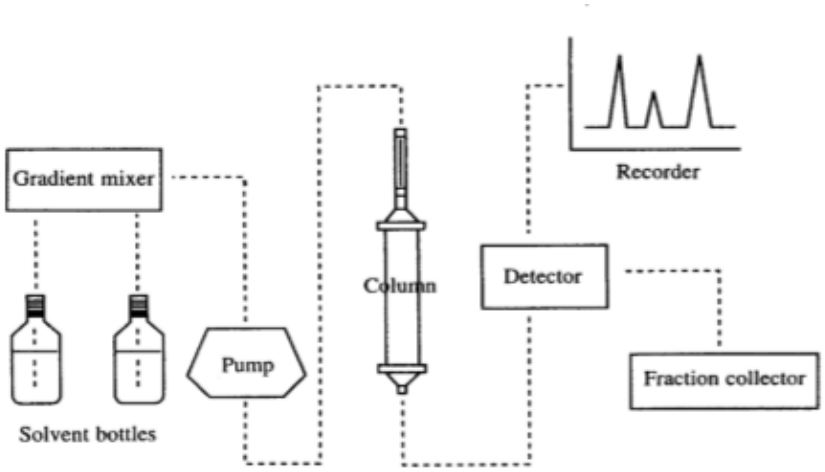


Figure 2.6: The main components of a medium pressure liquid chromatography system (Claeson et al., 1993).

2.4.4 High Performance Liquid Chromatography (HPLC)

Thin-layer liquid chromatography (TLC) is one of the most straightforward and easily applied preliminary tests that samples the polarity of a mixture, from which the solvent

system, for example, medium pressure liquid chromatography (MPLC), can be determined as being most suitable to separate the mix fully. Identifying and separating the components in regular compounds with chromatography can be achieved by several techniques including TLC and GC (gas chromatography). The HPLC system is the favoured analytical technique that produces superior outcomes for phytochemical studies. Compounds that lack chromophores in their structure are poor absorbers of UV; therefore, for such compounds, evaporative light scattering detection (ELSD) is typically used, making it a semi-universal detector for HPLC. Recently, reverse phase-HPLC with ELSD detection has become the option of choice to detect different classes of common compounds; this technology is particularly favoured for analysing food and drinks, as well as contributing to drug development (Dvovravckova et al., 2014). Because the technology is linear and relatively non-selective, even if the sample contains unknown compounds, the detected quantities of diverse constituents present in a complex mixture can be approximated (Cebolla et al., 1997). In phytochemical screening, the primary benefit of using ELSD rather than UV detectors is that it can discern diverse ranges of compounds such as glycosides, saponins and terpenoids that are poor absorbers of UV. The biggest limitations of the technique are that it is regarded to be destructive, and depending upon the composition of the solvent used in the mobile phase, the results can be inconsistent (Looney, 2012). As Figure 2.7 demonstrates, LC-ELSD is governed by three progressive techniques, which are nebulisation in the mobile phase nebulisation, and evaporation of eluent, followed by measuring the scattering of light by the retained analytes. The analytes are isolated in the column containing the mobile phase; together, these are turned into a fine spray of uniformly sized droplets by the nebuliser. The droplets are suspended in a transporter gas composed solely of nitrogen, forming an atomised spray. This is moved to a drift tube, which is heated, enabling the mobile phase to evaporate, whilst keeping the particles of interest in the evaporation tube. These form aggregates of dried particles, which can then be identified through their light scattering abilities (Young and Dolan, 2003). Light scattering is measured in the detected unit; a photodiode or photomultiplier captures the

data from the light scattered by the solid particles as the light source is directed at them (Looney, 2012). As the concentration of target components in a sample increases, the amount of scattered light increases accordingly (Campos et al., 2016). The association between the peak region of evaporative light scattering (A) and the quantity of analyte in the samples (m) is represented by the condition: $A = a \times m^b$, in which a and b are constants. The plot of the log of peak area against the log of analyte concentration will be linear (*i.e.*, $\log A = c + b \log m$, where $c = \log a$ in which a comes from $A = a \times m^b$) with slope b, and y-intercept a. The estimation of b lies in the interval [1,2], and seems to shift, from one analyser to another, depending on the specific design of the analyser, particularly the nebuliser component. Useful linearity is achieved when b is close to 1 (Young and Dolan, 2003).

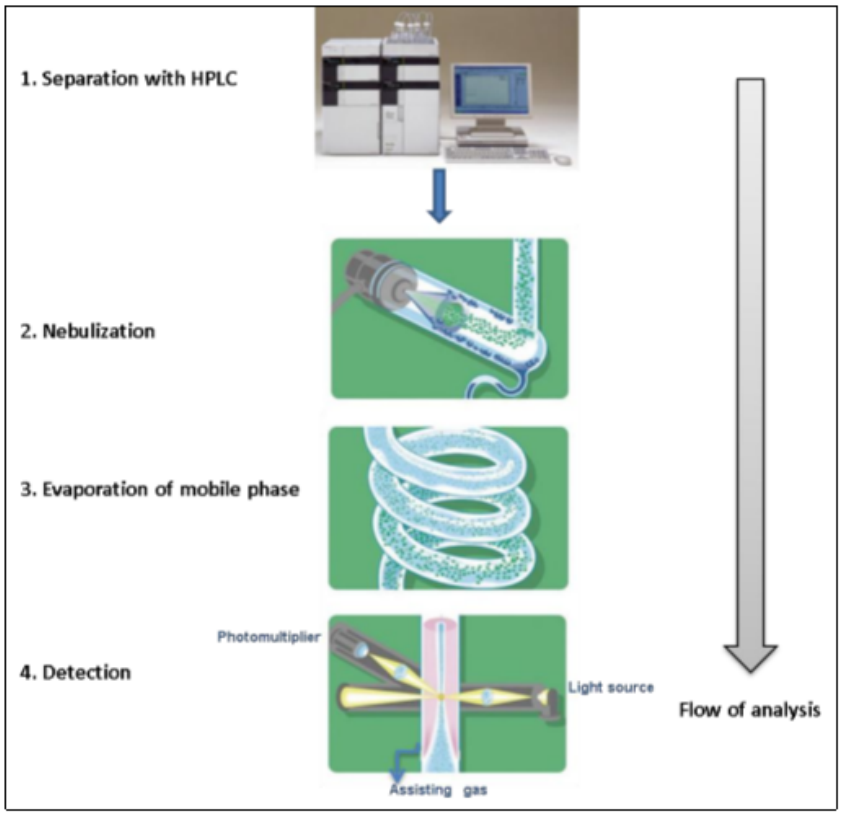


Figure 2.7: Principle of HPLC-ELSD detection of compounds during analysis (Campos et al., 2016).

2.4.5 High Resolution Mass Spectrometry

Whilst GC-MS is a valuable technique for determining the presence of unstable compounds in a sample, and HPLC is effective for measuring substances that absorb UV, neither of these technologies is considered particularly suited to evaluating propolis. This is due to its composition being complex. The more effective means that has been used recently is to couple HPLC to MS (LC-MS); this combination has been found effective to investigate and measure the diverse range of constituents of propolis (Midorikawa et al., 2001; Volpi and Bergonzini, 2006). This conjoined strategy is successfully applied to analysing natural products and pharmaceuticals, in addition to being an important advancement that has facilitated in-depth analyses of changes to metabolites in biological samples (metabolomics), genomics and proteomics. Furthermore, HPLC-MS and Tandem mass spectrometry (MS/MS) are also the preferred methods of analysing new medicinal compounds at every stage of their development (Korfmacher, 2005).

There are various machines that can yield ions within the mass spectrometer's ionisation chamber. The ionisation source gives an interface between the chromatographic system utilised for the separation of analytes and whatever remains of the mass spectrometer. In summary, ions are produced either by ionising a neutral substance through capturing or ejecting an electron, protonation or deprotonation, cationisation or by converting molecule charge into a charged form in the gas phase (Kang, 2012). Ionisation modes are distinguishable in their ability to fragment or broadly maintain the analytes during ion formation. Those modes that fragment are described as hard ionisation systems (e.g. electron impact ionisation), whilst soft ionisation strategies (e.g. electrospray ionisation) minimally fragment the analyte. The ionisation methods used most often in LC-MS today include:

1. Electrospray ionisation (ESI) is an Atmospheric Pressure Ionisation (API) strategy. ESI is most effective when polar molecules less than 100 Dalton (Da) in mass size

are being investigated; however, it can be used for molecules greater than 1,000 KDa. In ESI, a fine spray of charged droplets is created by applying a high voltage (usually about 1-4 KV) to a capillary containing a flowing liquid. The procedure is generally improved through the use of a coaxial nebuliser gas, such as nitrogen (Figure 2.8[6]).

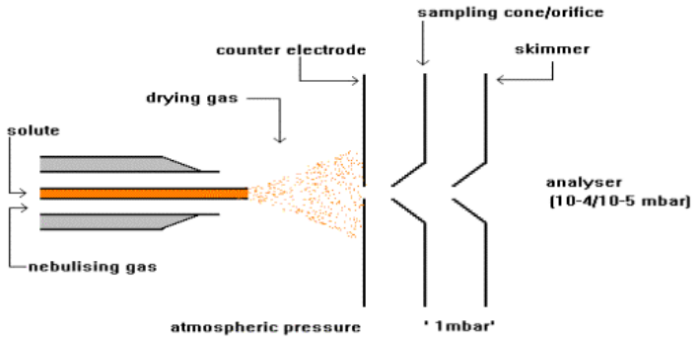


Figure 2.8: Basic diagram of an electrospray ionisation (ESI)

ESI is suitable for evaluating natural compounds, which have medium to high polarity. Since positive ionisation is dependent on protonation, molecules containing essential useful gatherings function admirably in this mode. In contrast, negative ionisation is the product of deprotonation; therefore, acidic functional groups such as carboxylate, imide phenol and are required for negative ESI, whereas amino, amide, ester and aldehyde groups are appropriate for positive ESI.

2. Matrix-assisted laser desorption/ ionisation (MALDI), is a laser desorption ionisation technique. The sample is combined in a saturated solution of a matrix, and a drop of the mix is deposited on the MALDI target. After the solvent dissipates and the matrix crystallises, the sample is put in the mass spectrometer source and is flashed with pulses of laser light. Energy is transferred between the excited matrix molecules and sample molecules as a result of desorbing both from the condensed

[6]Source: http://www.astbury.leeds.ac.uk/facil/MStut/mstutorial_files/image004.gif.

state. With molecules in the vapour phase, protons transfer between the matrix and the sample, leading to the formation of ions, to which high potential (usually 20 KV) is applied, accelerating the ions out of the source into a series of extraction electrodes and lenses (Figure 2.9[7]). This technique is particularly suited to examining organic, thermolabile, non-volatile compounds and those with high molecular masses. Consequently, MALDI is frequently used to investigate proteins, peptides, oligonucleotides and other biochemical compounds. However, MALDI has demonstrated its value in characterising polymers as well as large organic molecules and organometallic complexes (Go et al., 2007).

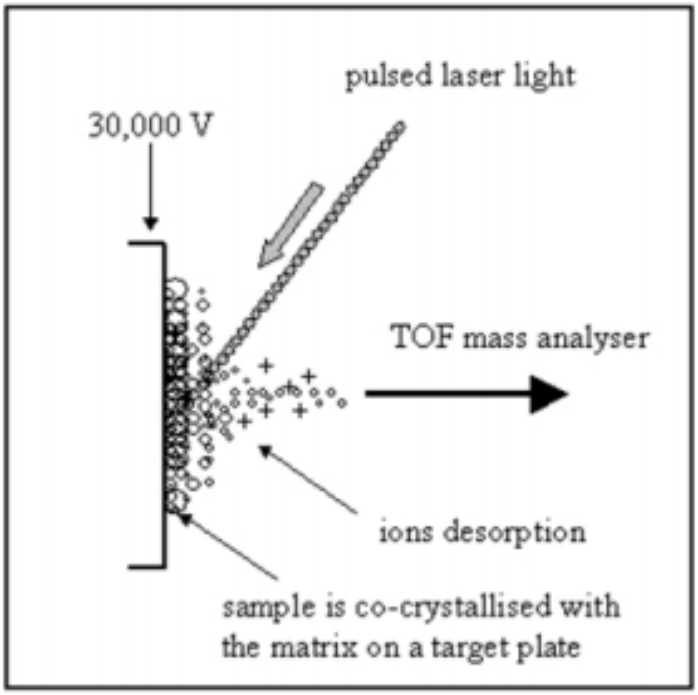


Figure 2.9: Basic diagram of a matrix assisted laser desorption ionisation (MALDI)

A mass analyser is used to separate and recognise the ions according to their respective mass/charge ratios. The capability of a mass spectrometer to distinguish between very similar masses determines its resolving power. Orbitraps, Fourier transform ion cyclotron resonance (FT-ICR) mass spectrometers and time of flight (TOF) instruments are

[7]Source: http://www.chem.pitt.edu/sites/default/files/users/Bhg5/figure%205.jpg.

amongst the mass analysers with the highest resolution currently available (Xian et al., 2012).

2.4.6 Mass Analyser

After the extraction of the ions by the ioniser, ions enter the ion analyser in order to be separated to their mass-to-charge $\frac{m}{z}$ ratios. The most commonly used mass analysers are:

1. The Orbitrap mass analyser is a recent invention, devised by Alexander Makarov at the end of the 20th century (Hu et al., 2005); it came on the market in 2005. It uses two specially designed electrodes; being barrel-like in shape, the outer electrode forms a C-trap that catches and stores ions temporarily, while the inner electrode is spindle-shaped (Figure 2.10). Applying a voltage causes the captured ions in the C-trap to move towards the inner -shaped electrode. Here, voltage is applied locally, causing the ions to circulate round the electrode, thereby getting caught in their motion (Hu et al., 2005). The ions adopt a spiral motion around the spindle-shaped electrode; this motion is sustained by balancing the outward centrifugal force caused by the initial tangential velocity upon ion injection, with the inward electrostatic attraction towards the central electrode. Measuring mass is contingent upon transients created by vibrating ions and their recurrence; the mass is independent of the ions' energy or their spatial distribution.

   The development of Orbitraps was inspired largely by the necessity of overcoming the issues of resolution and precision that afflict earlier technologies. It offers high mass accuracy (about 1-2 Parts per million (ppm)), high mass resolution of around 150,000, high $\frac{m}{z}$ range around 6000 and dynamic range around 104 (Hu et al., 2005). This dynamic range suggests that MS can distinguish different concentrations of analytes up to a factor of 104. This sensitivity is essential in the analysis of samples where one analyte is at a much lower concentration of another analyte that is present

in much higher concentrations, such as in studies of impurities. Hybrid systems that are highly accurate and have high mass resolving capabilities can be used to screen suspected components with or without the support of reference standards; even unknown compounds can be analysed in this manner. The technology provides copious data that not only addresses the mass of molecules or atoms and their molecular formula but through fragmentation designs created by MS/MS, it can also provide some detailed structural information (Krauss et al., 2010).

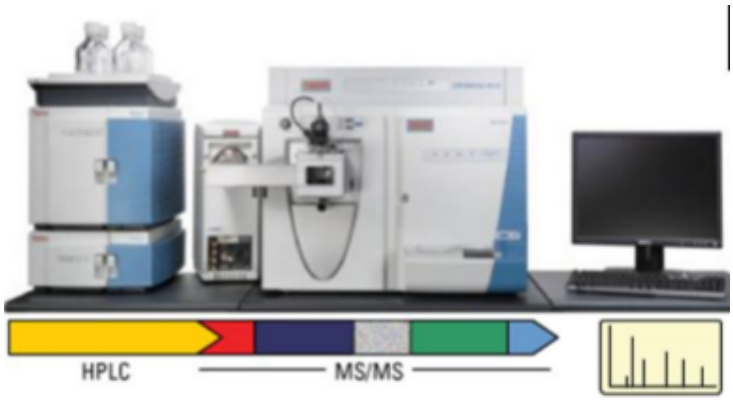


Figure 2.10: Orbitrap (Snider, 2014)

2. TOF analysers simultaneously accelerate ions, so all of the ions receive the same kinetic energy as each other. Therefore, the ions travelling a fixed distance through an evacuating flight tube separate based upon their mass-to-charge ratio and velocity. Figure 2.11[8] depicts the basic layout of a simple linear TOF analyser.

[8]Source: http://www.kore.co.uk/graphics/MS-200_tof.gif.

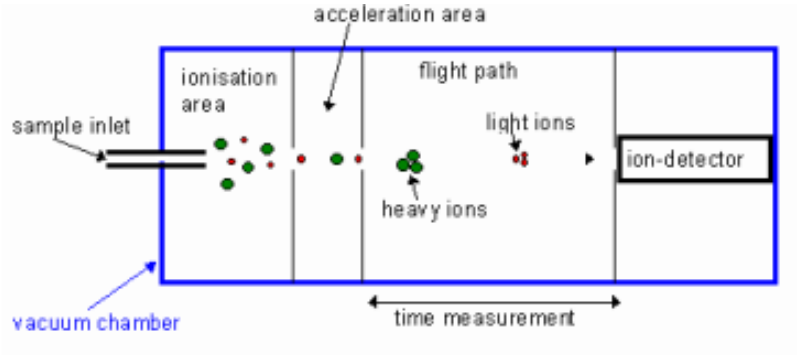


Figure 2.11: Time-of-Flight mass analyser

The prime advantages of TOF analysers are that they are capable of achieving resolutions between 5000 and 20000 Full width at Half Maximum (FWHM) and they are also relatively little in size, and cheap.

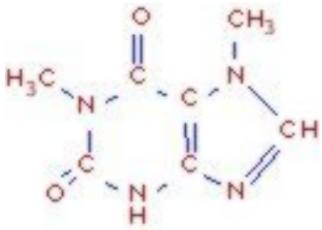
2.4.7 Detection of the Ions

The detector is the final component of the mass spectrometer. It monitors the ion current and increases it, then the signal is rapidly transmitted to the information system where the mass spectra are recorded. Detection of ions should be possible in a wide range of courses, contingent upon the kind of mass spectrometer in use. The identifiers that are used most often are the Charge (or Inductive) Detector, the Electron Multiplier, the Faraday Cup and the Photomultiplier Conversion Dynode (Go et al., 2007).

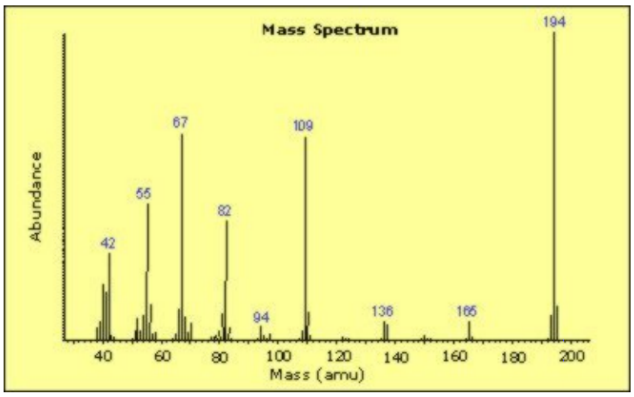
2.4.8 The Mass Spectrum (MS) and its Interpretation

In an MS plot, the $\frac{m}{z}$ values of the ions are plotted along with their abundance, which is shown as intensity. This data depicts the number of components in the sample, the molecular mass and relative abundance of each component in the sample. Figure 2.12 shows an example of an $\frac{m}{z}$ plot for a typical sample of an isolated compound. This is a

plot of relative intensity (abundance) versus the mass-to-charge ratio $\frac{m}{z}$. Several peaks are



(a) Structural formula of a compound



(b) M/Z plot for the isolated compound (mass spectrum)

Figure 2.12: Example of a Mass Spectrum for a compound (Prelorendjos, 2014).

apparent in the spectrum plot, with the most intense indicating the greatest abundance; this is referred to as the 'base peak'. All of the other peaks are characterised relative to the intensity of the base peak. Typically, the peak observed in the spectrum with the highest molecular mass represents the parent molecule, named as the molecular ion. From the plot in Figure 2.12, it is apparent that the most abundant peak is at mass 194; this peak represents the base peak, but because it is also the highest molecular mass, this peak also represents the molecular ion. However, this is not the norm and in most instances, the molecular peak or peaks in an MS plot differ from the most abundant one. The remaining peaks in this plot are ion fragments of various masses residual from the initial neutral molecule. The role of the spectrometer's mass analyser in presenting the spectra is very important. The accuracy, mass range, resolving power and scan speed of a mass spectral device are determined by the analyser (Barwick et al., 2006; Webb

et al., 2004). Accuracy is based on the instrument's stability and resolution; it reflects the detail with which a mass analyser can provide $\frac{m}{z}$ information. The resolving power is the mass spectrometer's ability to differentiate between ions of various $\frac{m}{z}$ ratios. With greater resolving power, there is a superior capacity to distinguish ions. It is defined as:

$$Resolution = \frac{M}{\Delta M} \tag{2.1}$$

where M is the mass-to-charge ratio $\frac{m}{z}$ and ΔM is the full width at half maximum (FWHM). The mass range of an analyser is effectively its $\frac{m}{z}$ range. The $\frac{m}{z}$ range varies in accordance with the type of analyser. If the resolution is sufficiently high, ions of different isotopes may be separated. The scan speed is the rate at which an analyser scans over a particle's mass range. Typically, it takes a few seconds, but again there is variation dependent upon the type of analyser used.

2.5 Conclusion

Within this chapter propolis and its properties have been introduced, and the main idea of metabolomics data was explained. The chemical processes for analysis of the composition of propolis samples were also described. In particular, mass spectrometry (MS) was investigated and discussed in detail; this part explains the creation of the data. MS data sets obtained from propolis samples are used in this thesis, to investigate the merits of different multivariate statistical methods for analysis of such data.

In the next chapter, metabolomics is discussed, where information about propolis as a molecular profile is considered as a sort of metabolomics (Bankova et al., 2016).

Chapter 3

Metabolomics and Analysis Techniques

After discussing extensively propolis and chemical analysis of propolis samples in chapter 2, this chapter now describes metabolomics data in general and introduces our data to be used in this thesis. This chapter is divided into four parts: the first part, in Section 3.1, provides an overview of metabolomics, and Section 3.2 provides an introduction to the main multivariate statistical analysis methods used in metabolomics. Section 3.3 provides and describes the data that are used in this thesis, and Section 3.4 gives a short conclusion.

3.1 Overview of Metabolomics

The propolis data sets to be studied in this thesis are of metabolomics type. Several established methods can be used for metabolomics data sets. The selection of the appropriate technique usually depends on the context of the investigation to be done (Griffin, 2004; Weckwerth and Morgenthal, 2005). Usually, the type of samples used in the analysis dictates the most suitable method to be used for the creation of the metabolomics data. The common methods include Nuclear Magnetic Resonance (NMR) spectroscopy,

Ultra-Violet spectroscopy (UV), Fourier Transform Infrared (FT-IR) spectroscopy and Mass spectrometry (MS).

Nuclear Magnetic Resonance spectroscopy is used if biofluids are involved in the analysis, and the method of Ultra-Violet spectroscopy is used to study the metabolic profiles of plants and plant materials (Bouchereau et al., 2000). Indeed, Fourier Transform Infrared spectroscopy is not used very often in metabolomics, where the disadvantage of this method is that it provides a very poor distinction between the various classes of metabolites (Griffin, 2004; Lindon et al., 2006).

In this project, MS is more appropriate for our samples. Before using the Mass spectrometry method, the Mass spectrometry requires a separation of the metabolic components. There are many available separation methods such as gas chromatography (GC), liquid chromatography (LC), high-performance liquid chromatography (HPLC), capillary electrophoresis (CE) and ultra performance liquid chromatography (UPLC). All these methods generate complex multivariate data sets, which need further analysis and interpretation with the appropriate chemometric tools (multivariate statistical methods).

The word "metabolomics" was first considered in 1998 when it was used in the context of describing the metabolic conduct of microbial systems. Following this period, the term has been used extensively within the scientific community, including areas outside the field of microbiology. The suffix 'ome' in the word "metabolomics" signifies the aim of the field to "direct attention to holistic abstractions" based on those observations that are possible, although as only a part of that whole (Oldiges et al., 2013). Truly, the principal aim of metabolomics is to identify, quantify, and classify all cellular metabolites. Metabolomics is, a rapidly emerging field that can be expressed as a comprehensive study of all metabolites–the end products of regulatory developments in a cell (Fiehn, 2001). The level of metabolites indicates the response of biological systems to environmental and genetic changes (Fiehn, 2002). To respond to these changes, biological systems combine a set of metabolites which constitute its metabolome. The analysis of metabolomes can help by providing some clarification to explain how metabolite levels vary in response to genetic

and environmental changes as a single mutation (Fiehn, 2001). In a more comprehensive view, the analysis of metabolomes is crucial for the understanding of cellular function; it aids in uncovering the dynamic nature of metabolism. This is done through the supply of informed knowledge about the various types and quantities of existing metabolites, and the environment which exists in cell systems and living organisms (Tomita and Nishioka, 2006). A clearer understanding of metabolism leads to a better understanding of the overall physiological state of organisms. In metaphorical terms, metabolomics has been described as a direct 'functional readout of the physiological state' of a living organism (Roessner and Bowne, 2009). Thus, metabolomics is a powerful tool which is capable of understanding the knowledge of underlying principles of the feature of living organisms. Metabolomics has been applied in different fields such as pharmaceutical analysis, plant science, toxicology and disease diagnosis, environmental and human nutrition research.

In the next section, a description of the main aspects of analysis of metabolomics data is given.

3.1.1 Analysis of Metabolomics Data

There are three main approaches which are used as tools for the analysis of metabolic networks and pathways, which are metabolite fingerprinting, metabolite profiling and metabolite target analysis (Fiehn, 2002; Nielsen and Oliver, 2005; Ryan and Robards, 2006).

More precisely, metabolite fingerprinting is considered as spectra generated by analytical approaches, such as NMR and MS, which provide a fingerprint of the metabolites produced by a cell. It is used to classify a large number of samples with the aid of multivariate statistics. This procedure has a disadvantage, which is that there is no information about or differentiation of individual metabolites.

Additionally, metabolite profiling is concerned with the identification and quantisation

of a predefined group of known or unknown metabolites, which is a type of metabolites belonging to a selected metabolic pathway. This is the oldest and most established metabolite analysis approach and was considered as the precursor for metabolomics.
The main metabolite target analysis is concerned with qualitative and quantitative analysis of a specific metabolite or metabolites which participate in a specific part of the living system's metabolism. Thus, only signals from the required metabolites are retained for analysis, while the other signals are treated as negligible.

3.1.2 Mass Spectrometry in Metabolomics

Mass spectrometry (see chapter 2) is an analytical technique that acquires spectral data in the form of a mass-to-charge ratio ($\frac{m}{z}$) and a relative intensity of the measured compounds. For the spectrometer to generate the peak signals for each metabolite, the biological sample first needs to be ionized. The resulting ionized compounds from each molecule will then generate different peak patterns that define the metabolite profiling of the original molecule. A wide range of instrumental and technical variants are currently available for MS spectrometry. These variants are mainly characterised by different ionization and mass selection methods (El-Aneed et al., 2009).
In metabolomics, MS is generally preceded by a separation step. This step reduces the high complexity of the biological sample and allows the MS analysis of different sets of molecules at different times. Liquid and gas chromatography columns (LC and GC, respectively) are the most commonly used separation techniques (Theodoridis et al., 2011). This chromatographic separation technique is based on the interaction of the different metabolites in the sample with the adsorbent materials inside the chromatographic column. This way, metabolites with different chemical properties will require different amounts of time to pass through the column. The time that each metabolite requires, called "retention time", is used together with the $\frac{m}{z}$ MS values to generate the two axes of the LC-MS and GC-MS spectral data. Figure 3.1 shows examples of LC-MS spectra.

Amongst the three different strategies used in metabolomic studies, stated above, there are two main research directions: metabolic profiling and metabolic fingerprinting. Metabolic profiling is focused on the analysis of a set of metabolites related to specific biochemical pathways or a group of compounds (Dettmer et al., 2007). In pharmacology, the objective of metabolic profiling is obtaining the catabolic outcome of administered drugs (Fiehn, 2002).

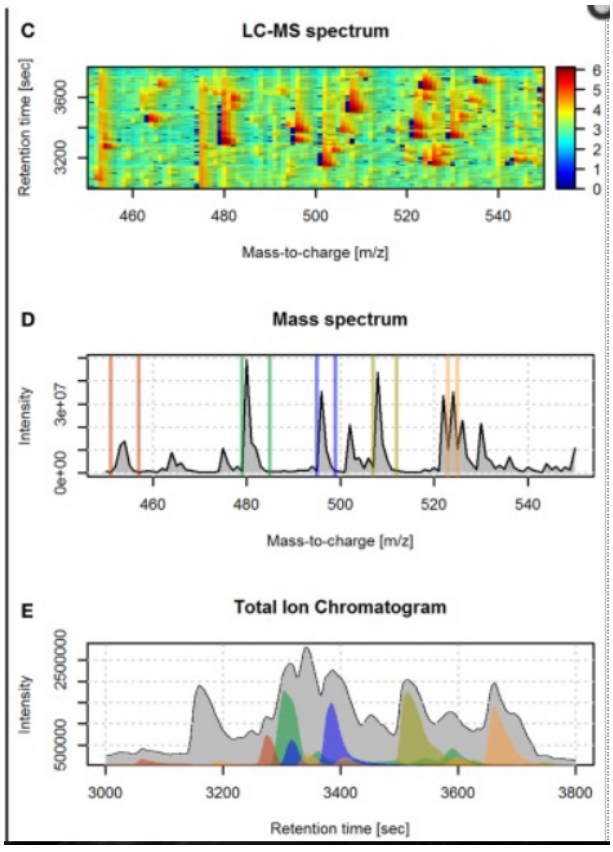


Figure 3.1: Examples of spectra obtained with LC-MS technologies. (C) An example of a LC-MS spectrum with colour-coded intensity and referred to the $\frac{m}{z}$ and retention time axes. (D) The sum of the LC-MS spectrum across the $\frac{m}{z}$ axis. (E) The total ion chromatogram (i.e., sum of the LC-MS spectrum across the retention time axis). The coloured regions in (E) correspond to the sum of the LC-MS spectrum limited to the $\frac{m}{z}$ ranges depicted with the same color in (D) (Alonso et al., 2015).

Metabolic profiling is usually hypothesis-driven rather than hypothesis-generating: metabolites are selected for investigation on the basis of the hypothesis. Therefore, the drawbacks of metabolic profiling are obvious: although this method can corroborate or refute the hypothesis, its capacity to reveal new aspects is limited. Another drawback of this method is that it is location-limited because it focuses on specific groups of metabolites (Dettmer et al., 2007).

Metabolic fingerprinting can be carried out on a broad range of biomaterial: urine, tissues, cells, amongst others. Contrary to metabolic profiling, metabolic fingerprinting is a main 'omics' approach. This is because it can be applied concurrently to a wide range of metabolites. Strictly speaking, as highlighted in various papers in the literatures, metabolic fingerprinting is a more global approach (Dettmer et al., 2007; Ellis et al., 2007). Metabolic fingerprinting focuses on comparing the changing patterns of metabolites. Metabolic fingerprinting is widely utilised as a diagnosis tool in medicine because it aids the identification and separation of diseased subjects, as well as to assess the dynamics of biotic, abiotic, and genetic perturbations (Ellis et al., 2007).

Metabolic fingerprinting and metabolic profiling depend on various analytical tools, one of which is nuclear magnetic resonance (NMR) technology, which aids the screening of samples for various patterns. The principal advantage of NMR is its non-destructive feature, meaning that the sample can be used for further analysis by other techniques. Furthermore, it is highly selective, with the capability to distinguish between several closely related chemical compounds. In addition, NMR requires minimal sample preparation requirements. However, the drawback of this method is that it only allows for the identification of medium and high-level metabolites, as a result of the NMR tool possessing limited sensitivity. In particular, Scalbert et al. (2009) estimated that no more than 60 different metabolites could be assessed in a biological sample.

This limitation of NMR, however, can be countered by performing mass spectrometry (MS) analysis instead. MS is a highly selective and sensitive instrument. The MS method details spectral information such as the precise mass of molecular ion and fragmentation

patterns. This, therefore, allows for the identification of the metabolites. As a consequence, a rapid increase of MS-based metabolomics studies has been evident. MS analysis is often used in conjunction with liquid chromatography (LC) analysis. The combination is usually referred to as LC-MS-based metabolomics. In LC-MS analysis, the prepared biological samples are admitted into a mass spectrometer through LC. In a simplified description, LC-MS analysis works as follows: comparative abundances of metabolites are estimated, data is processed, and analysis takes place (Chen et al., 2007). The creation of the LC-MS technology was mainly driven by the pharmaceutical industry. The drive for this was the industry's need for high sensitivity and precision, factors required for studying drugs and their metabolic effects (Lindon et al., 2011). The LC method alone is insufficient in providing a comprehensive analysis due to its limited sensitivity and selectivity. In metabolic profiling MS works as a separation method; it separates metabolites based on their mass-to-charge ratio. However, when LC is combined with MS, a higher degree of sensitivity and specificity is achievable (Fiehn, 2002).

In metabolic fingerprinting, the function of LC-MS analysis is not to separate the analyte but to provide precise data for further identification and analysis of biomarkers. This function also requires a high degree of sensitivity, which the LC-MS technique allows. The main advantage of the MS technique in metabolic fingerprinting is its capability to deliver high mass accuracy, which in turn produces a good anatomic structure of data. A further implication of this is that the number of potential identities for candidate markers is decreased (Theodoridis et al., 2013). Therefore, the overall process of identification and analysis is aided.

3.1.3 Advantages and Disadvantages of MS

Advantages, and disadvantages of MS are given below (Chao et al., 2010; Kealey and Haines, 2002) The main advantages of MS are:

1. Providing selective qualification and quantification of metabolites.
2. It can simultaneously identify and measure a variety of metabolites.
3. High sensitivity.
4. Offers rapid detection of metabolites.

However, in using the MS method, the researcher can meet the following problems:

1. The detection limits are lower if the substance to be analysed can be ionized.
2. Before applying MS it is necessary to apply a number of different separation techniques depending on the classes of the substances to be analysed.
3. MS methods require confirmation from standard compounds, which is often not available, especially for unknown compounds.
4. MS is a destructive analytical technique (unlike NMR). That means, after an MS analysis the samples cannot be reused for other analyses.

3.1.4 Metabolomics Applications

Metabolomics has several applications, where metabolite profiling is performed for medical and diagnostic purposes (Gomez-Casati et al., 2013). Furthermore, metabolomics aids the classification and description of plants and fungi (Gomez-Casati et al., 2013; Hong et al., 2016; Smedsgaard and Nielsen, 2005), such as detecting and quantifying mycotoxins which cover a path, to characterisation of fungi. The study of mycotoxins was also utilised to progress regulations related to food safety (Nielsen and Jewett, 2007). Additionally, metabolomics is an important tool in functional genomics (Bino et al., 2004). Specifically speaking, it aids the discovery of the functions of genes. A further example is a role metabolomics plays in providing a classification of molecular signatures, which accounts for a phenotype of unknown and silent mutations. In addition to these, metabolome stud-

ies have been utilised to characterise attributes which account for a silent plant phenotype. Metabolomics tools have also aided the developing of hypotheses about the impact of certain phenotypes on amino acid and carbohydrate metabolism (Nielsen and Jewett, 2007). There are other potential applications of metabolomics in evolution studies. For example, studies highlight that certain secondary metabolites are very species-specific and are considered to be potential markers for phylogenetics and taxonomy (Roessner and Bowne, 2009). As a result, they can aid in revealing the evolution of certain species. In the field of pharmacology, metabolomics studies are widely used for the purposes of drug discovery and development (Wishart, 2008). Specifically speaking, metabolomics is applied in lead compound discovery. Furthermore, metabolomics aids identifying biomarkers, which are essential in monitoring diseases as well as assessing drug efficiency, thus, it is also used in drug metabolism studies. Finally, metabolite research has been used in drug toxicity assessment, clinical trials and post-approval drug monitoring (Wishart, 2008). Hence, it is observed that metabolomics is associated with essentially all stages of drug development, from discovery to post-approval maintenance.

In general metabolomics data sets are highly multivariate, i.e. many more variables are recorded than the number of observations present.

3.2 Multivariate Analysis in Metabolomics

The data generated in a metabolomics experiment generally can be represented as a matrix of intensity values containing N observations (samples) of K variables (peaks). In general, analysis of metabolomics data involves the application of multivariate statistical methods and informatics used for chemically-based data ("chemometrics"). The main aim of metabolomics is to classify a spectrum, where the data are generated by a metabolomics analytical technique and contain the metabolic profile information from a biological sample. More precisely, the benefit of the spectrum is to identify its basic patterns of peaks,

and also metabolites corresponding to these peaks. This approach can require reducing the dimensionality of these complex data sets, for example by two or three-dimensional mapping procedures to enable easy visualisation of any clustering or similarity of samples. In addition, supervised chemometric methods can be used to model multi-parametric data sets, so the class of separate samples can be predicted based on a series of mathematical models derived from the original data (Green, 2014).

The aim of metabolomics is to supply a universal snapshot of biological fluids and all small-molecule metabolites in a sample, free of observational biases present in more focused studies of metabolism. However, the huge information content of such universal analyses generating very high dimensional data introduces another challenge, as efficiently drawing biologically useful conclusions from any one metabolomics data set requires specialised forms of data analysis (Chatfield, 2018).

One path to finding meaning in metabolomics data sets involves multivariate analysis (MVA) methods such as principal component analysis (PCA), hierarchical clustering analysis (HCA) and partial least squares projection to latent structures (PLS), in which spectral features contributing most to variation or separation are identified for further analysis. However, these methods are not a panacea; Worley and Powers (2013) discuss the use of multivariate analysis for metabolomics, as well as pitfalls and misconceptions. Metabolomics uses various statistical methods for data analysis. The choice of the method depends on the aims of the study. If the aim is to classify samples and if there is no prior information about the sample identity, then principal component analysis (PCA) and hierarchical clustering analysis (HCA) are used as exploratory methods to find out properties of biomarkers, while if the sample identity is known, then such supervised methods as partial least squares (PLS) can be used (Dettmer et al., 2007).

PCA is one of the most common statistical methods used, as metabolomics data are highly multivariate. This tool is used to reduce complexity or number of parameters, by projecting the data onto a lower-dimensional space. PCA allows observation of differences among samples and identification of variables which contribute to these differences. Also,

PCA is a powerful visualisation tool, which enables visual detection of sample patterns, through the projection of multidimensional data onto 2D and 3D plots (Figure 3.2).

HCA, on the other hand, is an unsupervised method which produces a dendrogram (tree-like diagram) to group data points. It, as well as other clustering methods, is used to evaluate in a multivariate way the similarity of a set of samples on the basis of the metabolite profiles of these samples. The use of HCA can allow classifying unknown samples according to their closeness to known ones. This method, however, is criticised as poorly reproducible and mathematically unjustified. It is also claimed that this method lacks adequate measurement for the quality of clusters (Goodacre et al., 2004).

Whilst the unsupervised nature of PCA gives a means to achieve dimensionality reduction, its only reveals group structure when within-group variation is sufficiently less than between-group variation. Therefore, supervised forms of discriminant analysis such as Partial Least Squares, that depend on the class membership of each observation, are also commonly used in metabolic fingerprinting experiments (Wold et al., 2001). However, this requires knowledge of pre-existing classes. PCA followed by HCA or other clustering method allows identification of unknown groups that may be present in the data, and is an approach often used in the literature (e.g. Miyagi et al., 2010; Worley and Powers, 2013).

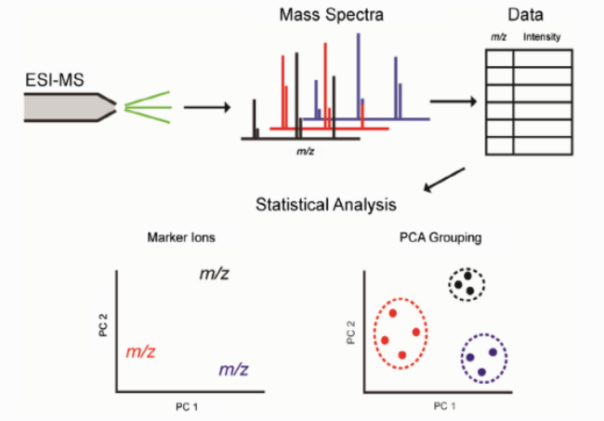


Figure 3.2: Statistical analysis of data via PCA to group samples and indicate the marker ions for each group (Bankova et al., 2016).

3.3 Data Description

We use three different data sets from propolis samples in this thesis. The first data set used to study methodology contains three sub-sets of data from Scotland and the second data set comes from propolis samples from different sites in Libya. These are used to compare methods of analysis, while a third data set is introduced in a case study in Chapter 10. Each one is described below:

- **Data from Scotland**

  Samples of propolis were collected during July and August 2014 by beekeepers from several of their honey bee colonies, in three different areas of Scotland, i.e. Aberdeenshire (north-east Scotland), Dunblane in central Scotland, and Fort William in the north-west (see Figure 3.3). These samples were profiled using liquid chromatography-high-resolution mass spectrometry (LC-MS) in Dr David Watson's lab in SIPBS, at the University of Strathclyde. The propolis samples contained several hundred compounds, many of which are still unknown structures. The Aberdeenshire data has 27 samples with 921 variables, there are 17 samples with 511 variables from Fort William and 9 samples from Dunblane with 498 variables, from 9 hives or colonies, 6 colonies and 3 colonies respectively. Every 3 samples come from the same hive, except for samples 10 and 11 in the Fort William data, which come from the same hive, but sample 12 comes from the same hive as samples 13 and 14.
  In all data sets, the rows are chromatographic peak areas (heights of the trace for that sample; see Figure 3.1) for putatively identified compounds. The column headings relate to a label for the sample (hive or colony), with 3 repeat analyses per hive (see Figure 3.4). I will refer to data set I, II and III for Aberdeenshire, Fort William and Dunblane respectively. The data were transposed for analysis, so that

metabolites relate to columns in the data and observed samples to rows. The proportion of zero values are 3%, 0.94% and 0.11% for Aberdeenshire, Fort William and Dunblane respectively.

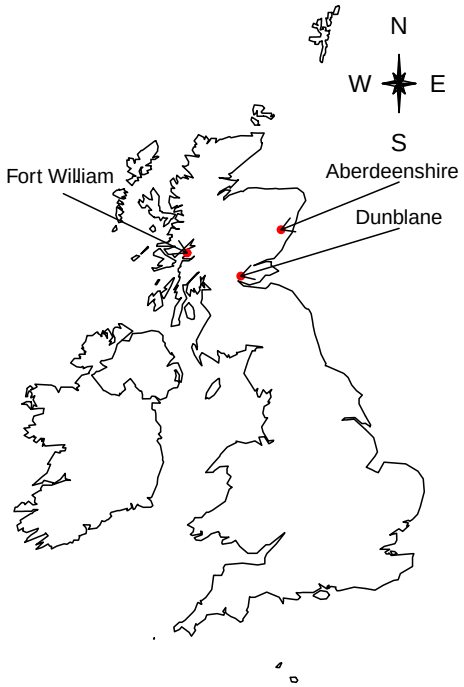


Figure 3.3: The UK map, including the locations of the colonies supplying the analysed Scottish propolis samples.

- **Data from Libya**

Twelve raw propolis samples were available from different geographical localities in Libya with 300 variables (see map in Figure 3.5); Tukra (Al Aquriyah), a small village located about 70km east of Benghazi city, Libya (P1); Qaminis (53km south of Benghazi) (P2); Bayda (east of Benghazi city) (P3); Quba (east of Benghazi city) (P4); Kufra A (south-east Libya) (P5); Kufra B (south-east Libya) (P6); Kufra C

	A	B	C	D	E	F	G
1	row ID	row m/z	row rete	Name	CABIN-	CABIN-	CABIN-
2	6504	121.0295	21.5	benzoic acid	1.29E+08	7.04E+07	6.22E+07
3	38	121.0295	11.7	benzoic acid	3.26E+08	7.86E+07	8.51E+07
4	4008	135.0452	8.2	phenylacetic acid	1917576	3085732	1517065
5	6946	135.0452	14.8	phenylacetic acid	3698759	7009556	1234282
6	1282	135.0452	12.3	phenylacetic acid	2.30E+07	2490245	6813808
7	6518	135.0452	16.0	phenylacetic acid	3060457	2932880	2128639
8	6581	135.0452	3.2	phenylacetic acid	440710.4	9163289	1816607
9	81	135.0452	9.3	phenylacetic acid	1.74E+08	1.87E+07	9.21E+07
10	3949	135.0452	7.4	phenylacetic acid	1.49E+07	2145642	2050572
11	1263	135.0452	2.3	phenylacetic acid	1486823	1589281	5794787
12	1266	135.0452	6.1	phenylacetic acid	2051629	3750220	1.32E+07
13	4996	135.0452	2.0	phenylacetic acid	1661488	2180287	1049138

Figure 3.4: An example of a data set for propolis, where column A shows the ID for each mass spectrum from the MassBank library (Horai et al., 2010), column B shows $\frac{m}{z}$ total ion chromatogram displayed for the detected peaks, column C shows retention time, column D shows name of components where this is available and columns E, F, G relate to a label for the hive (or colony).

(south-east Libya) (P7); Ghadames (south-west Libya) (P8); Tripoli (north-west Libya) (P9); Kasser Khiar (located 80 km east of Tripoli) (P10); Khumas (located 120km east of Tripoli) (P11); and Khumas (located 120km east of Tripoli) (P12). Samples P1-P12 were all used in this study. The proportion of zero values is 0.66% for Libya data.

Samples P1 and P2 were collected in December 2012, samples P3-P7 were collected in July 2013 and the other samples P8-P12 are from March 2014. In this data the rows are also chromatographic peak areas for putatively identified compounds. The column headings relate to a label for the samples (hives or colony; see Figure 3.4).

3.4 Conclusion

In this chapter, metabolomics has been described, as well as analysis of metabolomics data and metabolomics applications, and multivariate statistical analysis for metabolomics data, including PCA and cluster analysis using a hierarchical approach (HCA). The data to be used in this thesis were also introduced. These are metabolomics data sets resulting from MS analysis of propolis samples from Scotland and Libya. A further, European,

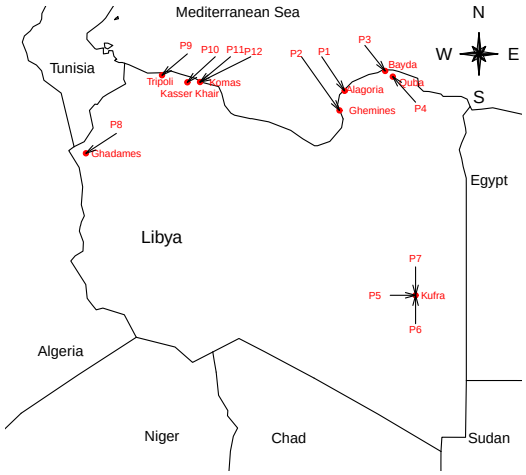


Figure 3.5: Map of Libya (after Siheri et al., 2016) including the locations of the colonies supplying the analysed Libyan propolis samples: P1 (Al Aquriyah), P2 (Qaminis), P3 (Bayda), P4 (Quba), P5 (Kufra (A)), P6 (Kufra (B)), P7 (Kufra (C)), P8 (Ghadames), P9 (Tripoli), P10 (Kasser Khiar), P11 (Khumas (A)), P12 (Khumas (B)).

data set is described and used in Chapter 10.

The next chapter of the thesis describes the most popular pre-processing and pre-treatment methods for the enhancement of the quality and accuracy of the metabolomics data, and the preparation of the data to make it suitable for further statistical analyses. A number of pre-treatment methods are also evaluated on the Scottish and Libyan data sets, as part of the applied work of this thesis.

Chapter 4

Pre-processing and Pre-treatment of the Data

After extensive discussion concerning propolis, chemical analysis of propolis samples and metabolomics, and introducing our data, this chapter now describes the most important and commonly used pre-processing and pre-treatment methods for metabolomics data. Pre-processing methods are described in Section 4.2 and pre-treatment methods can be found in Section 4.3. These methods are applied to the data sets already described, to compare their use.

4.1 Overview of Methods

Pre-processing and pre-treatment are an essential part of any chemometric data analysis. They concern the application of certain operations to data, either to remove unwanted variation or noise or to reduce it to an acceptable point. Goodacre et al. (2007) stated that pre-processing of a data set is the general term for those processes used to convert the raw instrumental data into clean data, in order to make it suitable for pre-treatment

and further applications. These pre-processing methods include Noise Filtering, Deconvolution, Peak Detection, Alignment and Baseline Corrections, among others. On the other hand, pre-treatment involves the transformation of the pre-processed data to prepare it for data analysis.

Metabolomics data are mostly presented in tabular form, with each row of such a table relating to a specific sample and each column to a single measurement (or variable). Pre-treatment includes scaling operations used on the rows and columns (most commonly Standardisation, Range scaling, Pareto-scaling, Vast scaling and Level scaling), and transformations of individual elements of a data set (usually logarithmic or power transformation). The effect of data pre-treatment will be illustrated by the application of eight data pre-treatment methods to metabolomics data sets. Pre-processing and pre-treatment of the data usually have either positive or negative effects on the outcome of the analysis. Pre-treatment methods are also applied to data sets to convert the clean data to a different scale (for instance, relative or logarithmic scale). Therefore, they aim to focus on the relevant (biological) information and to reduce the influence of disturbing factors such as measurement noise.

4.2 Pre-processing of Raw Data

After generating the signals, it is often necessary to apply specific techniques to clean the data. Pre-processing for MS data typically includes noise filtering, baseline correction, peak alignment, peak detection, peak quantification and spectral deconvolution. It should be noted that not all of the aforementioned processing steps are included in all methods, nor are they necessarily performed in the same order (Coombes et al., 2005). Additionally, analytical instruments do not provide clean and comparable lists of metabolites. The raw data must be processed to generate a practicable data matrix in a variety of ways (Castillo et al., 2011). The key step is to eliminate the variance and bias in

the data analysis, to reduce the complexity and enhance metabolically significant signals (Smith et al., 2006). Therefore, several algorithms have been developed, and multiple open-source programs have been applied to process raw MS data acquired through liquid chromatography-mass spectrometry (LC-MS) or gas chromatography-mass spectrometry (GC-MS). Among these, the following have attracted particular attention for their practicability and effectiveness: XCMS (https://xcmsonline. scripps.edu/) (Coombes et al., 2005); MZmine (http://sourceforge.net/projects /mzmine/) (Katajamaa et al., 2006); OpenMS (http://openms. sourceforge. net/) (Sturm et al., 2008); and MetAlign (http://www.metalign.nl) (De Vos et al., 2007). Most members of the research community in metabolomics work with these tools, and new programmes, such as MetSign, MSFACTs and Metabolite Detector (De Vos et al., 2007; Duran et al., 2003; Wei et al., 2011), have been steadily developed to increase the quality and efficiency of data pre-processing. Most of these tools are freely available. Furthermore, through these tools, the exchange of algorithms and data within the community is convenient.

In general, tools for raw data pre-processing include three basic modules, namely, noise filtering and baseline correction, peak detection and deconvolution and alignment. In the following sections, we will introduce different chemometric algorithms and strategies for these modules.

4.2.1 Noise filtering and baseline correction

Noise filtering is designed to separate component signals from the background originating from the chemical matrix or instrumental interference, and to remove measurement noise or baseline distortions (Katajamaa and Oresic, 2007). Regularly, during the baseline correction of one-way data, the two ends of a signal peak are manually identified by analysts and a piecewise linear approximation is then applied to fit a curve as the baseline (Zhang et al., 2010). However, this procedure is time-consuming, and its accuracy highly depends on the user's operating skills. Thus, numerous algorithms have been developed

for better estimation of the baseline. For MS-based data sets, the methods for removing random noise are typically implemented by traditional signal processing techniques in chemometrics. Noise filtering of LC-MS data is more complicated than that of GC-MS data because chemical and random noises are both included in the former (Hilario et al., 2006). This type of noise can lead to a baseline shift in the intermediate mass range of LC-MS spectra. In order to resolve this problem, several filtering methods have been proposed. Often in MS experiments, the generated spectra may appear to show baseline inconsistencies. Baseline offsets from spectrum to spectrum can affect the outcomes of the data analysis in many ways. They affect negatively the abundance of MS, hence causing problems in the accurate peak assignment and quantification (Xi and Rocke, 2008). For example, in a PCA model, baseline effects may cause the introduction of extra components in the model, and as a consequence, the results and interpretation of the analysis could be significantly altered from those taken from the actual model (Gemperline, 2006). There are different types of baseline effects, which vary from a simple offset to extremely complex shapes such as an upward or downward sloping line or even a broad curved shape. The ways to remedy these problems depend on the type of baseline error in the spectra. In simple offset cases, knowing that a specific region in the spectra has signal values equal to zero, it is usually sufficient to subtract the average value of the signal in this region ($\frac{m}{z}$) for each spectrum, from each metabolite in the respective regions. In more complex cases, it may be necessary to fit a polynomial function through all the valleys in the spectra. This polynomial line is then subtracted from the corresponding spectrum to correct the baseline differences (Gemperline, 2006). These methods are also called frequency domain correction methods (Xi and Rocke, 2008).

4.2.2 Peak detection and deconvolution

Peak detection and deconvolution are important to identify and quantify the signals corresponding to the molecules (e.g., the metabolites) in a sample (Castillo et al., 2011). A

peak detection method can identify the true signals correctly and avoid false positives. However, high response values do not always guarantee real peaks as some sources of noise can also produce high signals. Conversely, low peaks may correspond to real signals. Therefore, constraints on the peak shapes and criteria of minimal intensity, area or signal-to-noise are widely applied to distinguish real peaks from noise. Conventionally, peak detection algorithms follow two strategies: derivative techniques or matched filter response. A common problem in MS metabolomics studies is the appearance of overlapping peaks in the spectra. Deconvolution is a pre-processing technique that is used to overcome this difficulty (Goodacre et al., 2007). In fact, the fragments, adducts and molecule isotopes increase the difficulty of detecting peaks in the signals; therefore, it is necessary to improve the detection procedure. In MS, it is necessary to use the profile resolutions of both spectral and chromatography steps. This can be done by correlating the sample profiles with the retention time, in order to regroup ions coming from the same metabolite. In addition, deconvolution can be used to reduce the complexity of chromatograms obtained with soft ionisation methods by filtering multiple charged types, clusters and adducts (Jonsson et al., 2005). Additionally, to match the peaks extracted from files, all mass spectra and scans (chromatograms) have to be aligned across the total dataset and/or matching criteria should be set, commonly the mass and retention time windows (Dettmer et al., 2007). Recently, Tsugawa et al. (2015) proposed an open-source software pipeline, called MS-DIAL, instead of an R package, for data independent acquisition-based metabolite identification and quantification by mass spectral deconvolution.

4.2.3 Alignment

The purpose of alignment of detected features in different samples is to remove shifts in samples for a given signal, to guarantee downstream extraction of useful information. So far, several alignment techniques have been developed to minimise run-to-run shifts (Smith et al., 2015). To make them applicable to chromatographic systems coupled with

sophisticated detection instruments, e.g., LC-MS, which have yielded large amounts of two-dimensional data, the dimensionality should be reduced. The data reduction could be accomplished by generating integrated peak areas or total ion chromatograms. For one-dimensional data, some kinds of time alignment procedures could be employed as a useful method for tackling this problem of retention time shifts (Johnson et al., 2003). Other alignment methods attempt to integrate peak areas. Despite being time consuming, this approach is recommended as the process of data cleaning because the retention time shift, noise pollution and background shift are cleared simultaneously.

4.2.4 Conclusion

Pre-processing is concerned with the cleaning of the generated signals, to remove problems, for instance, overlapping peaks, baseline drifts, signal phasing and the existence of an extremely large number of metabolites in the data. A range of methods to overcome such problems was briefly described above, with emphasis on those methods most suitable for MS signals, as the spectra used in this project were generated by mass spectrometry. However, the available spectra had already been signal-processed by Dr David Watson (the data provider) (and colleagues); therefore there was no need to apply any of the aforementioned techniques.

4.3 Pre-treatment Methods

Once pre-processing of the data has been completed, it is quite often necessary to apply pre-treatment methods, in order to prepare the data for processing. It is common in metabolomics data analyses that not all the observed variation is desirable, but it is related to biological and technical variation (sampling, sample work-up and analytical measurement errors). Additionally, the data is more often than not heteroscedastic. Pre-

treatment methods are used to decrease the effects of these problems as much as possible. These methods depend on both the required biological information and the processing method to be used for the statistical analysis of the data. In all data sets in this project, the rows are a metabolite, and the columns are the samples (hives or colony) and these were transposed for analysis. Pre-treatment methods can be applied to the columns (column scaling), to the rows (row scaling) and to individual elements of a data set, called transformations (Brereton, 2009). The normal order of performing the following pre-treatment methods in a data set is usually to first transform individual elements of the data set, then to apply row scaling and finally to scale the columns (Brereton, 2009). Mean-centring may be used as part of scaling.

4.3.1 Transformations

In general, metabolomics data suffer from heteroscedasticity and are often skewed. In addition, interactions between the different metabolites are not necessarily additive but can be multiplicative (Boccard et al., 2010). The multivariate statistical methods used for the analysis of metabolomics data are more effective when the data is symmetric, and many statistical significance tests often assume that the distribution of the data is approximately normal. Therefore, it is useful to convert the data, so it approximates normality as closely as possible (Brereton, 2009). Thus, the transformations of the elements of metabolomics data sets are important in helping towards this aim. There are two common transformation approaches, which are the logarithmic and power transformation.

Logarithmic Transformation

A logarithmic transformation is important as it minimises the problem with heteroscedastic data, converts multiplicative models to additive and reduces the influence of large data values, for instance, outliers and occasional high peaks. This is achieved by replacing an

element x_{ij} by $\log(x_{ij})$. Here we use log to base 10 (van den Berg et al., 2006). Although this has advantages, it has some limitations such as handling zeroes or very close to zero values (especially when these values are very close to the limit of detection). If the values are below the limit of detection, then they are considered as zero, and therefore their logarithms are not defined (Brereton, 2009; van den Berg et al., 2006). Usually, a small value is added to x_{ij} in the case of x_{ij} being zero, before taking the log, even to all values. For instance, here we added 1 for all x_{ij} before taking the log.

Power Transformation

Power transformation has some strengths such as (Brereton, 2009):

1. It reduces the influence of large values such as outliers and occasional high peaks.

2. It can cope with zero values, eliminating the need to replace values below the limit of detection.

3. Any uncertainties in small values do not affect the data analyses as much as in the case of logarithmic transformation. The smaller a value is relative to other values, the smaller its influence on the n^{th} root transformed data will be. The drawbacks of this transformation can be summarised as:

1. All values should be positive.

2. If the distribution of the data is approximately log-normal, then power transformation cannot convert the distribution of values to a symmetric one.

3. There are many options for the value of the power. Trial and error are needed to identify the most appropriate choice. Especially in multivariate data such as in metabolomics, where each metabolite may have a different distribution, it can be quite difficult to decide on the power.

Power transformation is performed by replacing x_{ij} with x_{ij}^n. For $n = 1/2$, this is the square root transformation, and so on (Brereton, 2009). Additionally, the most popular of power transformation method is the square root (van den Berg et al., 2006).

Application of Transformation of Data Sets I, II and III

We now compare the effects of the different pre-treatment methods (mean-centring, log transformation and power transformation) on the Scottish data sets, in terms of the results of PCA (we consider the Libya data later) (Shlens, 2003). The effect of the various transformation methods on the PCA scores and loadings of data set I can be seen in Figures 4.1 and 4.2 respectively. The score plots in Figure 4.1 indicate that the scores of the twenty-seven samples of data set I (Aberdeenshire) are quite similar in shape, with the scores of the true data and power transformed data (for both n values) having the highest similarity. The log data show a slightly different pattern.

In this section we consider only the effect of transformation on the patterns seen in the plots of the PCA results, not interpretation of the principal components themselves. The interpretation is considered later in the thesis. Concerning the loadings plots in Figure 4.2, there is a similar pattern to that of the scores, as the loadings on both PCs in the true data and (both) power transformed data sets are similar in shape, whereas the log transformed plot is quite different. The loadings on PC1 and PC2 of the log transformed data (Figure 4.2) showed many large peaks, while after power transformation only a few large peaks were present. In the other plots, there are fewer peaks, but the values of them are higher in magnitude, hence identifying a few metabolites contributing to PC1 and PC2.

It is evident that different results will be obtained when we use different means of pre-treatment as the input for data analysis. In general, there is more variation in the loadings of PC1 and PC2 for the log data, and the shapes of the scores and loadings plots for the true data and both powers have the highest similarity among all plotted data sets.

Now we will discuss which transformation methods is the best. The general idea behind

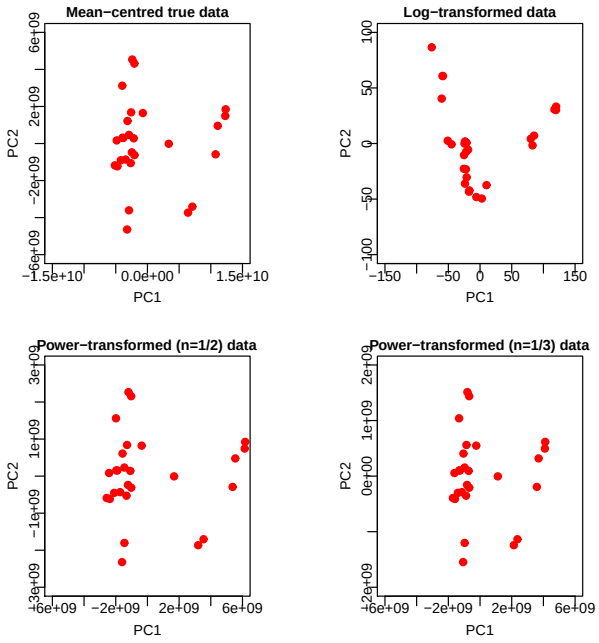


Figure 4.1: PC1 vs PC2 scores plots for the transformed Aberdeenshire data sets, using mean-centring, and log transformation or power transformation after mean-centring with n=1/2 and n=1/3.

the transformation is to make a variable more symmetric. Therefore, we can try various transformations and test for normality, as well as using visual displays, Q-Q plots, etc. Figure 4.3 shows histograms of the data set I values, with a normal curve superimposed (command plotNormalHistogram from R package "rcompanion"). Looking at the grey bars, this data is skewed strongly to the right (positively skew) and the log data looks much more normal. The grey bars deviate noticeably from the normal curve for the true and power transformed data (for both n values). The effect of the log and the power transformations on the data as a means to correct for heteroscedasticity is seen in Figure 4.3. Compared to the true data, power transformation was not able to remove the heteroscedasticity. The log transformation was able to remove heteroscedasticity, however only for the metabolites that are present in high concentrations. In contrast, the standard

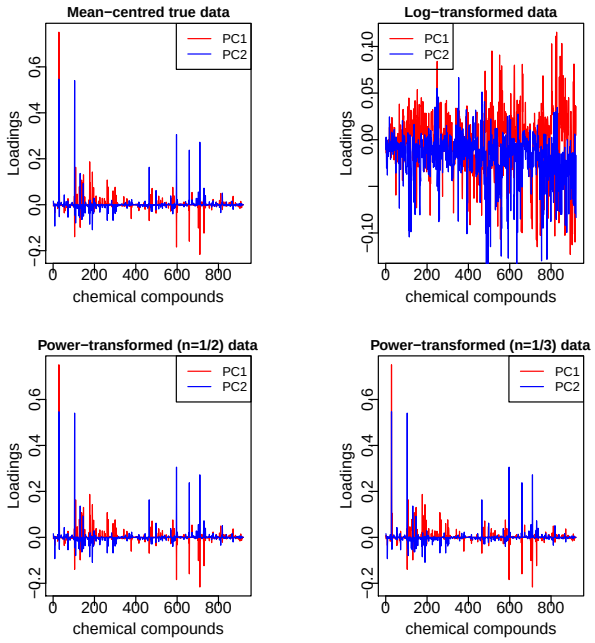


Figure 4.2: PC1 vs PC2 loadings plots for the transformed Aberdeenshire data sets, using mean-centring, and log transformation or power transformation after mean-centring with n=1/2 and n=1/3.

deviations of metabolites present in low concentrations were inflated after log transformation due to the large relative standard deviation of these less abundant metabolites. Therefore, log transforms tend to have a significant effect on distribution shape, and in visualisations can bring extreme outliers closer to the remainder of the data. Since data set I has many zero values, we added one to all values before log transforming to ensure that they are positive. This value can of course also influence the result (Brereton, 2009).

We now look at the Fort William data (data set II). The score plots in Figure 4.4 also indicate that the scores of the seventeen samples of data set II (Fort William) are quite similar in shape, with the scores of the true data and power transformed data (for both n values) having the highest similarity. Again the pattern is a bit different for the log data. Concerning the loadings plots in Figure 4.5, there is a similar pattern to that

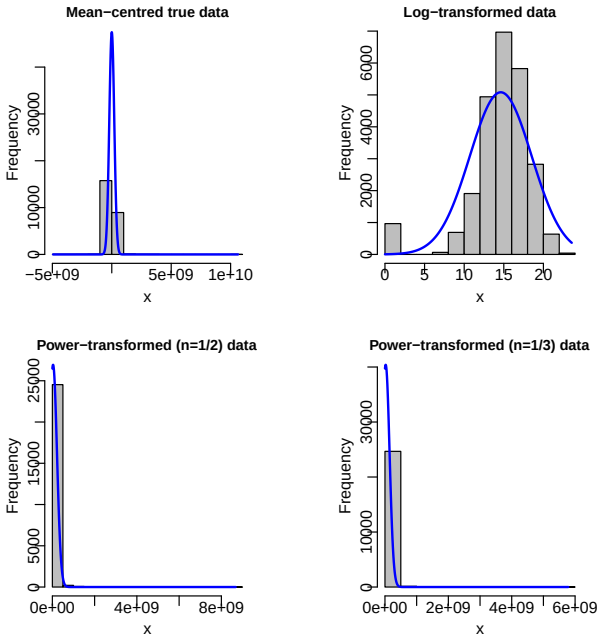


Figure 4.3: Histograms of the data set I values, with a normal curve superimposed.

of the scores, as the loadings on both PCs in the true and both power transformed data sets are similar in shape, whereas the log transformed plot is quite different. The loadings on PC1 and PC2 of log transformed data (Figure 4.5) showed many large peaks, while after power transformation, only a few large peaks were present. In the other plots there are fewer peaks but the values of them are again higher in magnitude, hence identifying a few metabolites contributing to PC1 and PC2. In general, there is more variation in the loadings of PC1 and PC2 for the log data, and the shapes of the scores and loadings for the true data and both powers have the highest similarity among all plotted data sets.

Figure 4.6 shows histograms of the data set II values, with a normal curve superimposed. Looking at the grey bars, this data is skewed strongly to the right (positively skew) and the log data plot looks most normal. The grey bars deviate noticeably from the normal curve of the true and power transformed data (for both n values). The effect of the log and the power transformation on the data as a means to correct for heteroscedasticity is shown in Figure 4.6. Again, compared to the true data, the power transformation was

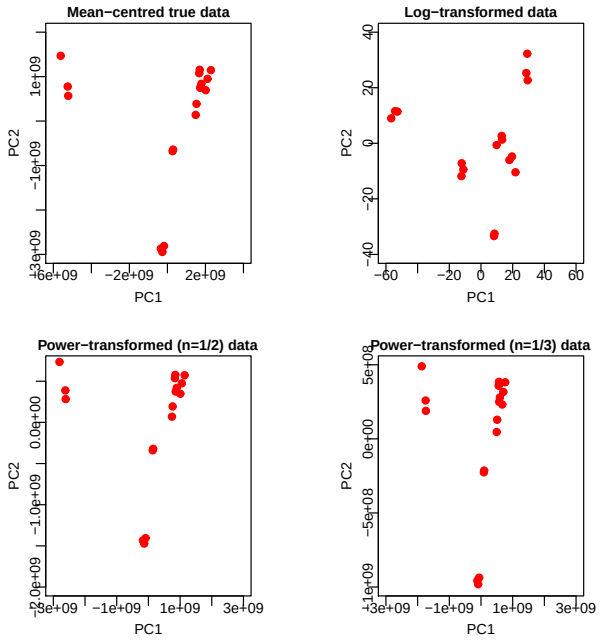


Figure 4.4: PC1 vs PC2 scores plots for the transformed Fort William data sets, using mean-centring, log transformation and power transformation with n=1/2 and n=1/3.

not able to remove heteroscedasticity. The log transformation was able to remove heteroscedasticity, however only for the metabolites that are present in high concentrations. As before, the standard deviations of metabolites present in low concentrations were inflated after log transformation due to the large relative standard deviation of these less abundant metabolites. Again, as data set II has many zero values, we added one to all values before taking the log, to ensure they are positive.

We now look at the Dunblane data (data set III). The score plots in Figure 4.7 indicate that the scores of the nine samples of data set III (Dunblane) are quite similar in shape, with the scores of the true data and power transformed data (for both n values) having the highest similarity, although as similar as with the other data sets. Concerning the loadings plots in Figure 4.8, there is a similar pattern to that of the scores, as the loadings on both PCs in the true and both power transformed data sets are similar in

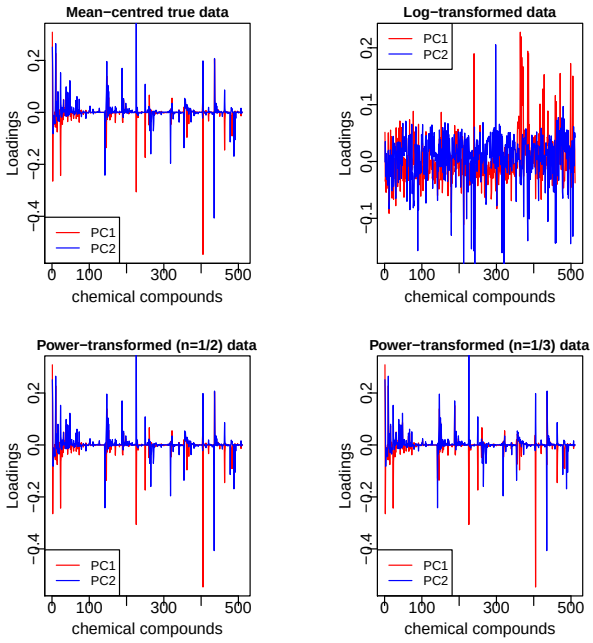


Figure 4.5: PC1 vs PC2 loadings plots for the transformed Fort William data sets, using mean-centring, log transformation and power transformation with n=1/2 and n=1/3.

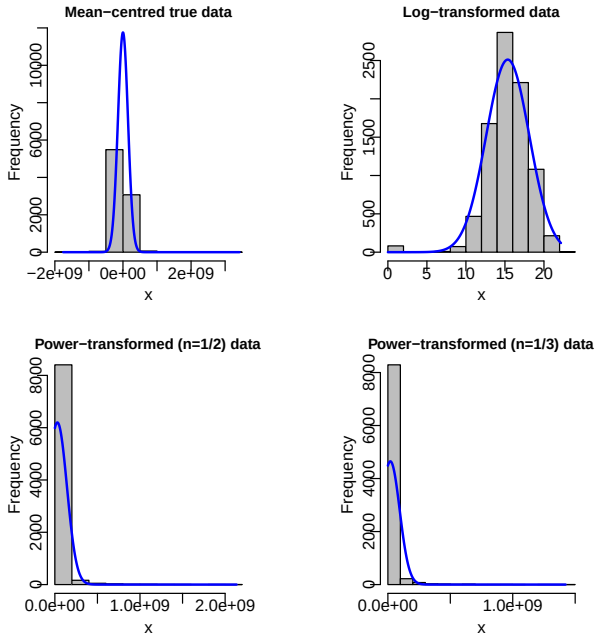


Figure 4.6: Histograms of the data set II values, with a normal curve superimposed.

shape, and again the pattern is a different for the log data. The loadings on PC1 tend to be negative for the log data. The loadings on PC1 and PC2 of the log transformed data showed many large peaks, while after power transformation, only a few large peaks were present.

In general, the shapes of the scores and loadings for the true data and both power transformed data sets have the highest similarity among all plotted data sets.

Figure 4.9 shows histograms of data set III, with a normal curve superimposed. This

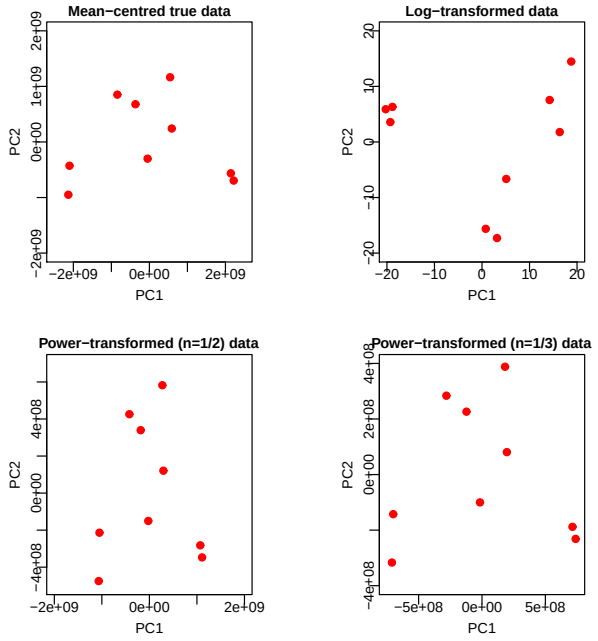


Figure 4.7: PC1 vs PC2 scores plots for the transformed Dunblane data sets, using mean-centring, log transformation and power transformation with n=1/2 and n=1/3.

data is skewed strongly to the right (positively skew) and the log data looks most normal. The grey bars deviate noticeably from the normal curve of the true and power transformed data (for both n values). The power transformation was not able to remove heteroscedasticity. The log transformation was able to remove heteroscedasticity, however, as before, only for the metabolites that are present in high concentrations. Again, as data set III

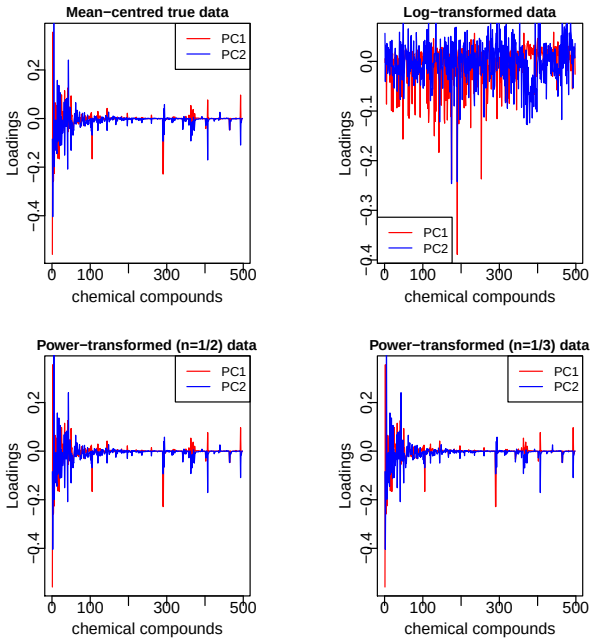


Figure 4.8: PC1 vs PC2 loadings plots for the transformed Dunblane data sets, using mean-centring, log transformation and power transformation with n=1/2 and n=1/3.

has many zero values, we added one to all values to ensure they are positive, before taking the log.

The normal order of performing pre-treatment methods in a data set is usually first to transform the individual elements of the data set, then to apply row scaling and finally to scale the columns (Brereton, 2009). Before considering scaling, we draw conclusions about transformations, from the comparisons made on these data sets.

Conclusions from Transformation of Data Sets I, II and III

Transformations are non-linear conversions of the data. Transformations are in general applied to correct for heteroscedasticity (Kvalheim et al., 1994), to make skewed distributions (more) symmetric, and to convert multiplicative relations into additive relations. In biology, relationships between variables are not necessarily additive but can also be

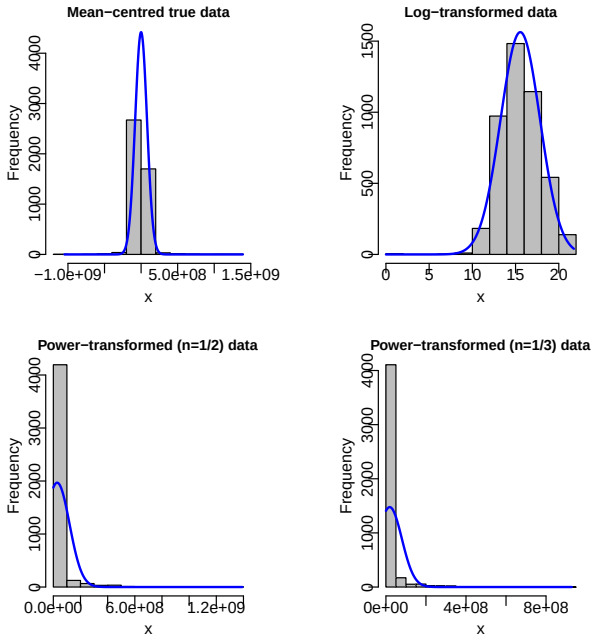


Figure 4.9: Histograms of the data set III values, with a normal curve superimposed.

multiplicative (Sokal and Rohlf, 1995). A transformation may be necessary to identify such a relationship.

Since the log and the power transformation minimise big values in the data set relatively more than small ones, the transformations have artificial scaling effect as differences between big and low values in the data are reduced. However, the artificial scaling effect is not determined by the multiplication with a scaling factor, as for a real scaling effect, but by the impact that these transformations have on the original values. This artificial scaling effect is therefore rarely sufficient to adequately adjust for the magnitude of variances. Consequently, it can be useful to apply a scaling method after transformation.

However, for data sets I, II and III it was decided not to use any transformation in further analysis. Although in every case the log transformation brought the data much closer to a normal distribution, the log transformed data led to much less interpretable plots from PCA. As PCA is a main part of our data analysis, it was decided the log transformation was unsuitable. The power transformation led to results that were very similar to those

from the raw data. Therefore, we decided to use the raw data for data sets I, II and III. We now consider scaling.

4.3.2 Scaling

Variables are often scaled in principal component analysis. This is especially recommended when variables are measured on various scales (e.g. different injection volumes in chromatography) which is very much the case with the columns of metabolomics data; otherwise, the PCA results obtained will be badly affected, and dominated by the more variable values or larger values (Jolliffe, 2011). The aim in scaling is to make the variables comparable, which is critical while performing principal component analysis (PCA). PCA tries to identity features with maximum variance, and the variance is high for high magnitude features. This skews the PCA towards high magnitude features. Therefore, the scaling method is a pre-treatment operation used to adjust the importance of the various elements in the data to the model-fitting procedure. The adjustment usually involves the weighting of the metabolites with a factor that can be estimated by using either a dispersion criterion or a size measure (Boccard et al., 2010). To clarify, we aim to make all metabolites on the same measurement scale to be comparable to each other. The two common scalings that are used are row-scaling and column-scaling, with different types of scaling possible in each case. We now consider these.

Row-scaling

- **Normalisation**

  To remove or minimise the variability from sample to sample, normalisation of the samples can be applied. This operation puts all the samples on the same scale, thus allowing for comparisons in the various samples. Normalisation involves dividing

each variable of a sample vector by a constant. There are a number of different constants that can be used, such as the 1-norm of the vector (Beebe et al., 1998; Brereton, 2009). For example, the 1-norm vector normalisation is given by:

$$\tilde{x}_{ij} = \frac{x_{ij}}{\sqrt{(\sum_{j=1}^{N_{metabolites}} x_{ij}^2)}}, \tag{4.1}$$

where x_{ij} is the element in the ith row and jth column. This 1-norm approach was used in the data analysis here (with the pre-processing of data done by Dr. Watson and his team). So the sum of squares of the elements of vector x_i after the normalisation is equal to one. The selection of the appropriate normalisation constant depends on the type of systematic variation in the samples. Normalisation belongs to the row-scaling methods (Brereton, 2009). It is an important step, as its purpose is to remove any systematic variation, retaining all the biological information in the data.

Column-scaling

1. Centring

Generally, centring pre-treatment methods allow the researcher to focus on the differences but not the similarities in the data. They focus on isolating and removing the systematic variation in the data. However, attention is needed when data are heteroscedastic, as the effects from centring methods might not be sufficient. Usually, centring methods are applied in combination with other pre-treatment methods. They belong to the column-scaling methods (Goodacre et al., 2007). The following methods are the most commonly used column-scaling methods in metabolomics.

- **Mean Centring**

This is a centring method in which each column of the data is expressed in deviations from its mean. The mean of the columns is subtracted, translating the centre of gravity of the data set to the origin. The formula for mean centring is

$$\tilde{x}_{ij} = x_{ij} - \bar{x}_j, \qquad \text{where} \quad \bar{x}_j = \frac{1}{N}\sum_{i=1}^{N} x_{ij}, \tag{4.2}$$

where $\tilde{x}_{ij}$ represents the data after mean centring, $\bar{x}_j$ is the overall mean of variable j and N is the number of samples in the data set.

- **Weighted Centring**

Weighted centring aims to convert all metabolite concentrations to fluctuations around zero instead of around their mean. Hence, it retains only the relevant variation (the variation between the samples) for the analysis. It is also known as reference subtraction. This method is particularly useful in PLS-DA (partial least squares discriminant analysis) classification where it takes into account several classes with different numbers of samples in each class (Brereton, 2009). The weighted mean for a data set with N_c classes can be estimated as:

$$\tilde{x} = \frac{\bar{x}_g + \frac{\sum_{h \neq g} \bar{x}_h}{N_c - 1}}{2} \tag{4.3}$$

where $\bar{x}_g$ and $\bar{x}_h$ are the mean vectors for groups g and h respectively. For two classes, $N_c = 2$, the above formula becomes

$$\tilde{x} = \frac{\bar{x}_1 + \bar{x}_2}{2}$$

where $\bar{x}_1$ and $\bar{x}_2$ are the mean vectors for groups 1 and 2 respectively, and $\tilde{x}$ is a global mean (but not the overall mean, which may be biased in favour of one

of the two classes, especially when the two classes have different sample sizes) (Brereton, 2009). Weighted centring can then be achieved by subtracting the weighted mean from each column of the data set, as long as there are N_c classes in the column.

2. Scaling Based on Data Dispersion

These scaling methods use a dispersion measure for scaling the data and more specifically the columns of a data set (Boccard et al., 2010; Goodacre et al., 2007; van den Berg et al., 2006). In all these methods, the mean and standard deviation are defined as:

$$\bar{x}_j = \frac{1}{N}\sum_{i=1}^{N} x_{ij}, \qquad \text{and} \quad s_j = \sqrt{\frac{\sum_{i=1}^{N}(x_{ij} - \bar{x}_j)^2}{N-1}}. \tag{4.4}$$

The following methods are the most commonly used column-scaling methods in metabolomics (Brereton, 2009):

• Standardisation

This is a form of scaling performed by mean-centring each metabolite value and using afterwards the standard deviation as the scaling factor. The formula is given by

$$\tilde{x}_{ij} = \frac{x_{ij} - \bar{x}_j}{s_j} \tag{4.5}$$

Standardisation is also called autoscaling or unit variance scaling, as after the standardisation procedure, all metabolites have a standard deviation equal to one, allowing the metabolites to be compared using correlations instead of covariances. The main advantage is that all metabolites become equally

important (van den Berg et al., 2006). After standardisation, the data becomes dimensionless.

- **Range Scaling**

The scaling factor in the range scaling method is the range within each metabolite. In this way, the formula is

$$\tilde{x}_{ij} = \frac{x_{ij} - \bar{x}_j}{x_{j_{max}} - x_{j_{min}}} \tag{4.6}$$

for metabolite j. Range scaling allows the comparison of metabolites with respect to their biological response range. In this approach, all metabolites are equally important, and their scaling is related to the biology of the data. However, an increase in measurement errors and sensitivity to outliers may be noticed when applying this scaling method. As in the case of standardisation, the data becomes dimensionless.

- **Pareto Scaling**

Here the square root of the standard deviation is used as the scaling factor. It aims to reduce the influence of large values without losing significant information concerning the structure of the data. The formula is:

$$\tilde{x}_{ij} = \frac{x_{ij} - \bar{x}_j}{\sqrt{s_j}}. \tag{4.7}$$

Pareto-scaled data is closer to the original than standardised data, but this depends very much on the large values in the data set.

- **Vast Scaling**

  This is an extension of standardisation. It aims to give more importance to those metabolites that appear to have small variances. To achieve that, the method uses the coefficient of variation statistic as a scaling factor. The formula is given by:

$$\tilde{x}_{ij} = \frac{(x_{ij} - \bar{x}_j)}{s_j} \cdot \frac{\bar{x}_j}{s_j} \tag{4.8}$$

  where $\frac{\bar{x}_j}{s_j}$ is the inverse of the coefficient of variation of x_j . This method is not useful when large induced variation exists, and there is no group structure in the data.

3. **Scaling Based on Average Value**

- **Level Scaling**

  Scaling based on average value methods uses a size measure instead of a spread measure. Level scaling is one such method. It converts the changes in metabolite concentrations into changes relative to the average concentration of the metabolite by using the mean concentration as the scaling factor. The resulting values are changes in percentages compared to the mean concentration. The formula for level scaling of metabolite j is given by:

$$\tilde{x}_{ij} = \frac{x_{ij} - \bar{x}_j}{\bar{x}_j} \tag{4.9}$$

  This method is suitable for the identification of biomarkers. It is however prone to increase measurement errors. Level scaling, like the scaling methods based on data dispersion, also belongs to the column-scaling methods.

We now apply scaling methods to our data.

Application of Scaling to Scottish Data Sets I, II and III

Information regarding the effect of applying column-scaling on the PCA scores and loadings of data set I can be seen in Figures 4.10 and 4.11 respectively (Shlens, 2003). The scores plots in Figure 4.10 indicate that there are some differences among the scores of the six scaling methods used on data set I. These six scaling methods are *standardisation*, *range*, *Pareto*, *vast* and *level*, and each one including mean-centring.

Concerning the loadings plots in Figure 4.11, the loadings on PC1 and PC2 for *standardisation*, *range*, *vast* and *level* scaling have a similar shape. In the other two scaling methods used on data set I, the mean-centred true and Pareto results, the loadings on both PCs have similar shapes. In general, the shapes of the loadings for the true and Pareto cases have the highest similarity among all plotted data sets.

The application of different pre-treatment methods on data set I had a large effect on the resulting data used as input for data analysis, as depicted in Figure 4.10 and 4.11. The different pre-treatment methods resulted in different effects. For instance *standardisation*, *range*, *vast* and *level* scaling showed many large peaks, while after Pareto-scaling, only a few large peaks were present. It is evident that different results will be obtained when differently pre-treated data sets are used as the input for data analysis.

We will now show an objective numerical comparison between the scaling methods in Table 4.1 for PCA, where PCA constructs orthogonal uncorrelated linear combinations of variables that explain as much common variation as possible. From Table 4.1, it can be observed that Pareto-scaling performed much better than the other pre-treatment methods in terms of PCA, because it explains more of the variation in the data set, and we will use the command *prcomp* in *R*, which uses singular value decomposition (SVD) (R Core Team, 2013; Shlens, 2003). The first two PCs explain 79% of the total variation of the data set I, considerably more them for any other method. Therefore it is best for data set I, to mean-centre and Pareto-scale prior to using PCA. We now consider data set II. An illustration of the effect of applying column-scaling on the PCA scores and loadings of

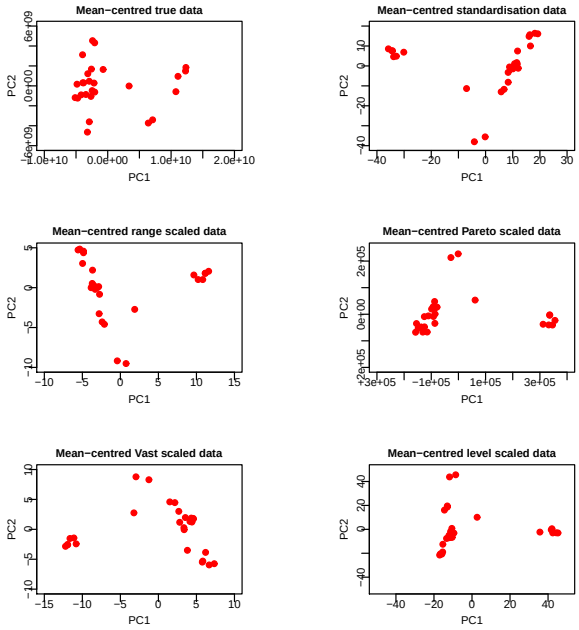


Figure 4.10: PC1 vs PC2 scores plots for the scaled Aberdeenshire data sets, using mean-centring, Standardisation, Range, Pareto, Vast and Level scaling.

pre-treatment methods	Standardisation	Range	Pareto	Vast	Level
% of variance of PC1	40.18	47.21	68.90	46.28	38.34
% of variance of PC2	19.48	17.00	10.10	15.66	19.93
Cumulative %	59.66	64.21	79.00	61.94	58.27

Table 4.1: Percentage of variance explained by the first two PCs of Aberdeenshire data, using different scaling approaches. The best method is shown in red.

data set II, can be seen in Figures 4.12 and 4.13 respectively. The scores plots in Figure 4.12 indicate that there are some differences among the scores of the six scaling methods used on this data set. Concerning the loadings plots in Figure 4.13, the loadings on PC1 and PC2 for *standardisation*, *range*, *vast* and *level* have a similar shape. On the other hand, the loadings of the true and Pareto-scaled data, for both PCs, have different similar shapes. In general, the shapes of the scores and loadings for the true and Pareto cases have the highest similarity among all plotted data sets.

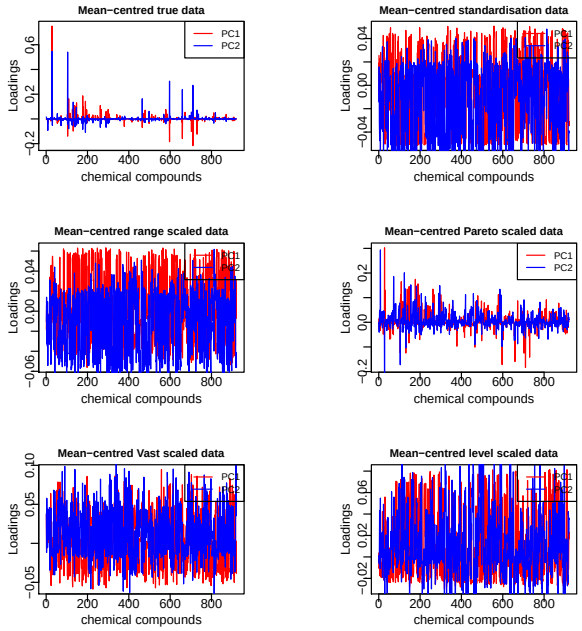


Figure 4.11: PC1 vs PC2 loadings plots for the scaled Aberdeenshire data sets, using mean-centring, Standardisation, Range, Pareto, Vast and Level scaling.

The application of different pre-treatment methods on data set II had a large effect on the resulting data used as input for data analysis, as seen in Figure 4.12 and 4.13. For instance *standardisation*, *range*, *vast* and *level* scaling showed many large peaks, while after Pareto-scaling only a few large peaks were present. It is evident that different results will be obtained when differently pre-treated data sets are used as the input for data analysis. From the Pareto scaling, or true data, there are many fewer peaks that are large in magnitude, which is likely to be more interpretable in identifying important metabolites.

From Table 4.2, it can be observed that Pareto-scaling performed better than the other pre-treatment methods in terms of PCA because the PCA explains more of the variation in the data set, and we will use the command *prcomp* in *R*, which uses singular value decomposition (SVD) (R Core Team, 2013; Shlens, 2003). The first two PCs explain 74.8% of the total variation of the data set II. Therefore, the data set II was mean-centred and

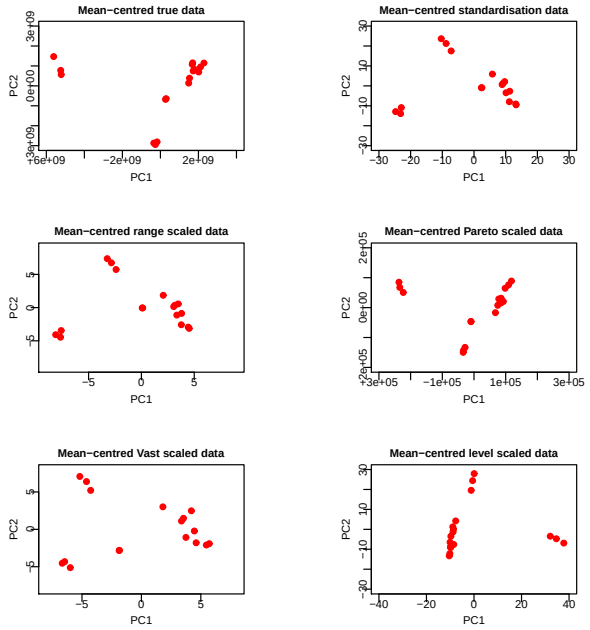


Figure 4.12: PC1 vs PC2 scores plots for the scaled Fort William data sets, using mean-centring, Standardisation, Range, Pareto, Vast and Level scaling.

Pareto-scaled prior to using PCA in further analysis.

We now consider data set III. For information regarding the effect of column-scaling on

pre-treatment methods	Standardisation	Range	Pareto	Vast	Level
% of variance of PC1	35.68	38.65	52.73	36.58	41.3
% of variance of PC2	25.64	25.41	22.05	23.57	21.99
Cumulative %	61.32	64.06	74.78	60.15	63.29

Table 4.2: Percentage of variance explained by the first two PCs of Fort William data, using different scaling approaches. The best method is shown in red.

the PCA scores and loadings of the data set III, see Figures 4.14 and 4.15 respectively. The scores plots in Figure 4.14 indicate some differences among the scores of the six scaled data sets. Additionally, the scores on PC1 and PC2 for *standardisation* and *range* have similar shapes. Concerning the loadings plots in Figure 4.15, the loadings on PC1 and PC2 for *standardisation*, *range*, *vast* and *level* scaling again have a similar shape. For

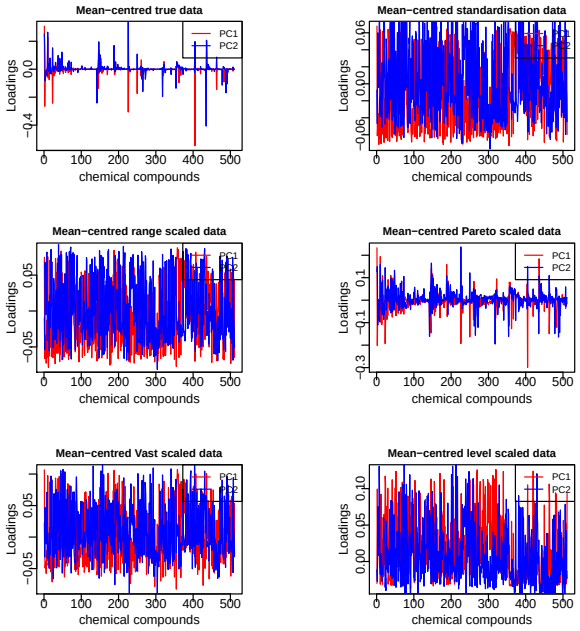


Figure 4.13: PC1 vs PC2 loadings plots for the scaled Fort William data sets, using mean-centring, Standardisation, Range, Pareto, Vast and Level scaling.

the other two scaling methods used on data set III, the true and Pareto-scaled data, the loadings on both PCs have different similar shapes. There are many fewer peaks that are relatively large in magnitude, which is likely to be more interpretable in practice for identifying important metabolites.

In general, the shapes of the scores and loadings plots for the true data and Pareto-scaled data have the highest similarity among all plotted data sets.

From Table 4.3, it can be observed that Pareto-scaling performed better than the other pre-treatment methods in terms of PCA (we will use the command *prcomp* in *R*, which uses singular value decomposition (SVD) (R Core Team, 2013; Shlens, 2003)), because the PCA explains more of the variation in the data set, where the first two PCs explain 76.9% of the total variation of data set III. Therefore data set III was mean-centred and Pareto-scaled prior to using PCA in further analysis.

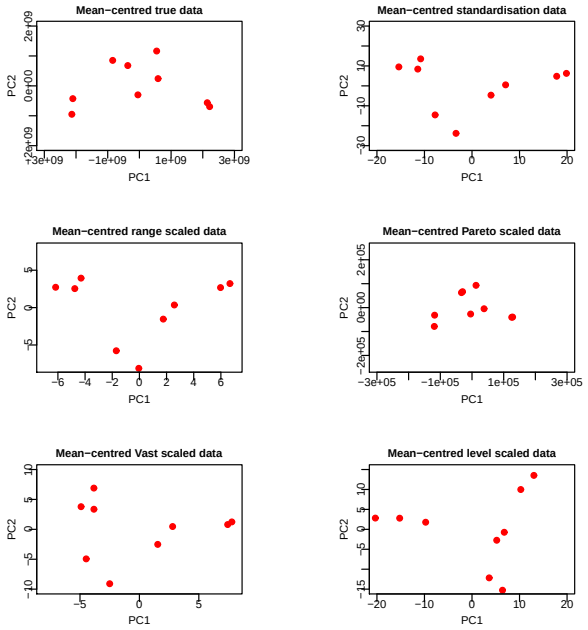


Figure 4.14: PC1 vs PC2 scores plots for the scaled Dunblane data sets, using mean-centring, Standardisation, Range, Pareto, Vast and Level scaling.

pre-treatment methods	Standardization	Range	Pareto	Vast	Level
% of variance of PC1	33.47	35.45	53.38	32.81	43.01
% of variance of PC2	30.33	30.67	23.53	30.16	26.03
Cumulative %	63.80	66.12	76.91	62.97	69.04

Table 4.3: Percentage of variance explained by the first two PCs of the Dunblane data, using different scaling approaches. The best method is shown in red.

4.3.3 Summary and Conclusions for Pre-treatment for Data Sets I, II and III

Before any chemometrics analysis takes place, it is necessary most of the time to process the generated metabonomics data to remove or reduce to acceptable levels the amount of systematic variation in the data, to make the data more suitable for statistical analyses. There are two stages in the preparation of the data, the pre-processing, then pre-

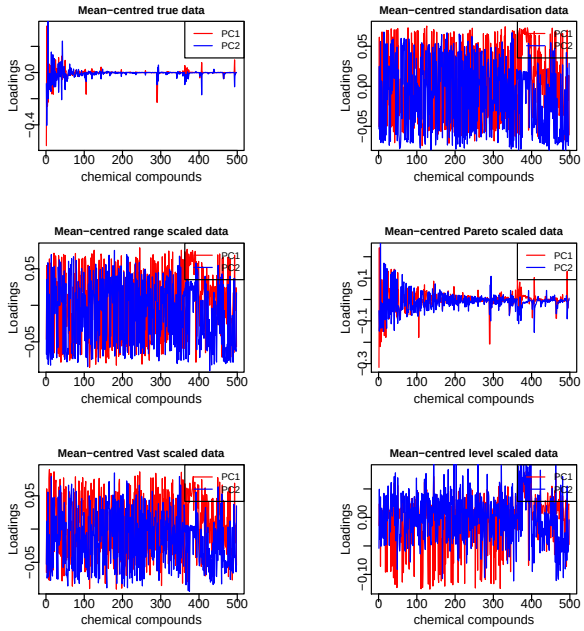


Figure 4.15: PC1 vs PC2 loadings plots for the scaled Dunblane data sets, using mean-centring, Standardisation, Range, Pareto, Vast and Level scaling.

treatment. There was no need to apply any of the previously mentioned techniques of pre-processing in this case, as it was already done.

Pre-treatment occurs in the second stage of data processing to remove or reduce any uninduced variation (due to sampling, sample work-up and analytical measurement errors) as much as possible and, if it exists, heteroscedasticity of the data. Description of the most well-known methods was given with respect to the three ways that the techniques can be applied to the data: i.e. transformations of the elements of the data matrix, row-scaling and column-scaling. Two different approaches for transforming the elements of a data matrix were discussed, i.e. the log and power transformations, and were used on the data. The results indicated that there was no clear improvement to be made to the results for data sets I, II and III by using either of these transformation techniques. Scaling methods (both row and column) were classified as *centring*, scaling based on *data dispersion* and scaling based on *average values*. The advantages and disadvantages of applying the pre-

treatment methods were discussed, as well as graphical representations (PC scores and loadings) of the effect on the first two PCs of applying these to the metabolomics data sets I, II and III.

Using these results, and considering the type of data to be analysed here, the pre-treatment methods chosen for data sets I, II and III used for exploratory analyses and clustering in Chapter 5-8, were *mean centring* with *Pareto* scaling of the columns, with no need found to scale to a constant total the rows of the data matrix. More specifically, the elements of each column in the data matrix were transformed by subtracting the column mean from each element, then dividing by the square root of the standard deviation, effectively making the columns more comparable to each other in the various analyses in Chapters 5-8. No element transformation was chosen. From Tables 4.1, 4.2 and 4.3, it can be observed that Pareto-scaling performed better than the other pre-treatment methods in terms of PCA, because the results explain more of the variation in the data. Therefore the data sets I, II and III were mean-centred and Pareto-scaled prior to using PCA.

4.4 Application of Transformation and Scaling on all Three Data Sets Combined (IV)

Here we will merge the three data sets I, II and III together (as Data set IV) and look at any relationships between location and results for the three data sets mentioned previously. These data sets are from three different sites in Scotland: the Aberdeenshire data contains 27 samples with 921 variables, Fort William has 14 samples, after removal of 3 samples which were outliers. These were detected in analysis reported in Section 5.4.4. The Fort William data has 511 variables, and the Dunblane data has 9 samples with 498 variables. Therefore, data set IV contains 50 samples with 1930 variables (chemical compounds). The data set I, II and III are organised in a one block diagonal matrix, since they were collected at different times and so the variables recorded are not necessarily the same.

That means the columns and rows are different in each matrix.

An illustration of the effect of the various transformation methods on the PCA scores and loadings of data set IV can be seen in Figures 4.16 and 4.17 respectively. The score plots in Figure 4.16 indicate that the scores of the 50 samples of data set IV are quite similar in shape for all except the log-transformed data. The log data shown a slightly different pattern. Concerning the loadings plots in Figure 4.17, there is a similar pattern to that of the scores, as the loadings on both PCs for the true and both power transformed data sets are similar in shape, whereas the log transformed plot is quite different. The loadings on PC1 and PC2 of the log transformed data (Figure 4.17) show many large peaks, while after power transformation, only a few large peaks were present. In all plots except the log transformed one there are fewer peaks, but the values of them are higher in magnitude, hence identifying a few metabolites contributing to PC1 and PC2. In general, the scores and loadings for the true case and both power cases have the highest similarity among all plotted data sets.

Figure 4.18 shows histograms of data set IV, with a normal curve superimposed. Looking at the grey bars, this data is skewed strongly to the right (positively skew). The grey bars deviate noticeably from the normal curve in every case, since there is a large bar corresponding to very small values in the log plot (which is a result of the combination of the data sets leading to the presence of many zeroes to which a value of one was then added). Compared to the true data, the power transformation was not able to remove heteroscedasticity. The log transformation was able to remove heteroscedasticity to some extent, however, the spike at 0 means that a normal curve does not fit. As there is no clear advantage to using any of these transformations, we decided to use the untransformed data in later analysis.

The effect of column scaling on the PCA scores and loadings of data set IV can be seen in Figures 4.19 and 4.20 respectively. The scores plots in Figure 4.19 indicate that there are some differences among the scores of the six scalings of data set IV. Additionally, the scores on PC1 and PC2 for *standardisation* and *range* scaling have a similar shape.

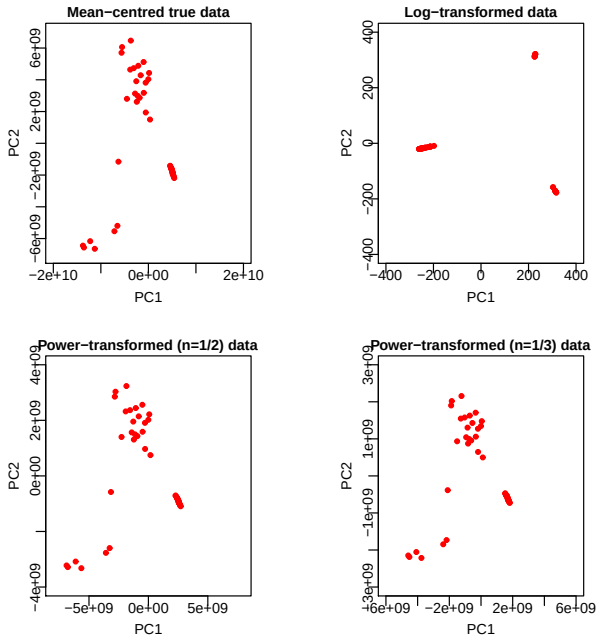


Figure 4.16: PC1 vs PC2 scores plots for the transformed data set IV, using mean-centring, log transformation and power transformation with n=1/2 and n=1/3.

Concerning the loadings plots in Figure 4.20 the majority of loadings on PC1 for *level* and *standardisation* are negative peaks. The loadings of the true and Pareto-scaled cases on both PCs have fairly similar shapes. The different pre-treatment methods resulted in different effects. For instance, *standardisation*, *range*, *vast* and *level* scaling, showed many large peaks, while after Pareto-scaling, only a few large peaks were present. It is evident that the Pareto-scaling result is likely to be more interpretable in practice for identifying important metabolites.

The purpose of applying the previous methods is to remove or reduce any uninduced variation. To verify this result, we will look at the PCA results, and we will use the command *prcomp* in *R*, which uses singular value decomposition (SVD) (R Core Team, 2013; Shlens, 2003). From Table 4.4, it is observed that Pareto-scaling performed much better than the other pre-treatment methods, as the first two PCs explain 66% of the total variation of data set IV, which is much higher than for any of the other methods of

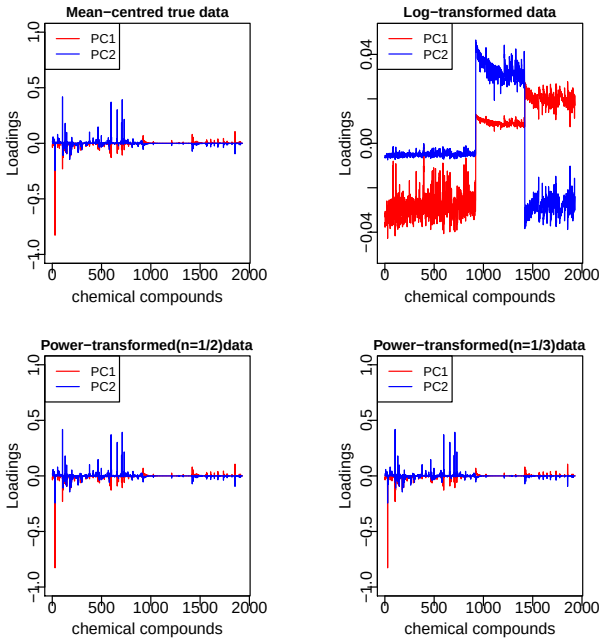


Figure 4.17: PC1 vs PC2 loadings plots for the transformed data set IV, using mean-centring, log transformation and power transformation with n=1/2 and n=1/3.

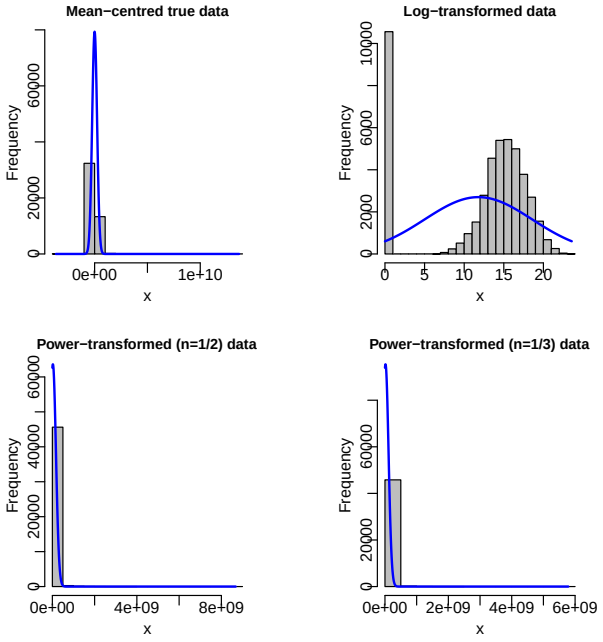


Figure 4.18: Histograms of the data set IV values, with a normal curve superimposed.

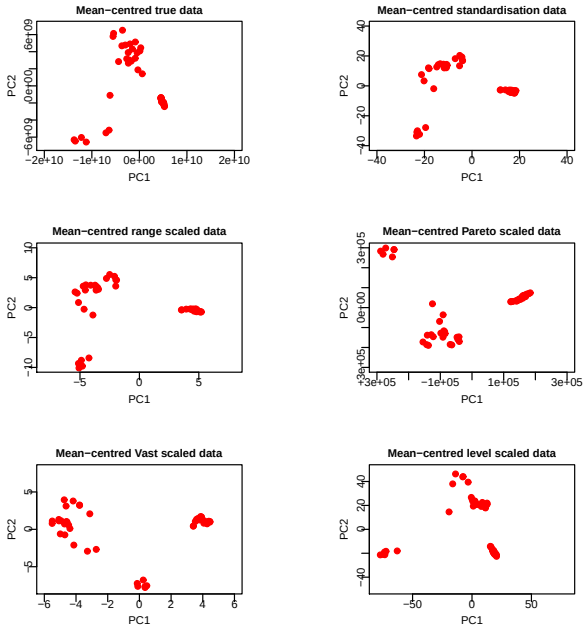


Figure 4.19: PC1 vs PC2 scores plots for the scaled data set IV, using mean-centring, Standardisation, Range, Pareto, Vast and Level scaling.

scaling. Therefore, this data set was mean-centred and Pareto-scaled prior to using PCA in later analyses. The results of the above analyses are quite consistent.

pre-treatment methods	Standardisation	Range	Pareto	Vast	Level
% of variance of PC1	28.04	30.78	36.79	35.64	29.37
% of variance of PC2	22.73	26.07	29.21	20.6	19.29
Cumulative %	50.77	56.85	66.00	56.24	48.66

Table 4.4: Percentage of variance explained by the first two PCs of data set IV, using different scaling approaches. The best method is shown in red.

We now try analysis of data from a different country, to see whether the same conclusions are valid for that.

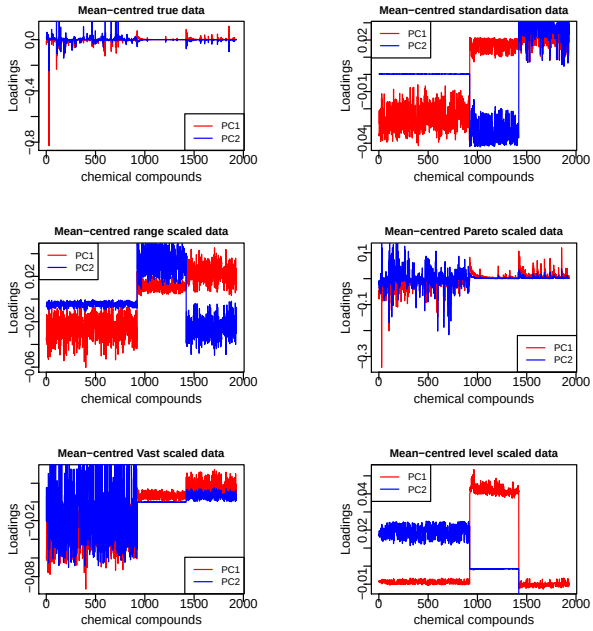


Figure 4.20: PC1 vs PC2 loadings plots for the scaled data set IV, using mean-centring, Standardisation, Range, Pareto, Vast and Level scaling.

4.5 Application of Transformation and Scaling for Libya Data

An illustration of the effect of the various transformation methods on the PCA scores and loadings of the Libya data set can be seen in Figures 4.21 and 4.22 respectively. The score plots in Figure 4.21 indicate that the scores of the twelve samples of the Libya data are similar in shape for all except the log transformed data. The log data shows a different pattern. Concerning the loadings plots in Figure 4.22, there is a similar pattern to that of the scores, as the loadings on both PCs in the true and power transformed data sets are similar in shape, whereas the log transformed plot is quite different. The loadings on PC1 and PC2 of all transformed data showed many large peaks, but many more using log transformation and in that case it will be difficult to identify a few metabolites contributing to PC1 and PC2.

Now we will discuss which transformation methods is the best. Figure 4.23 shows

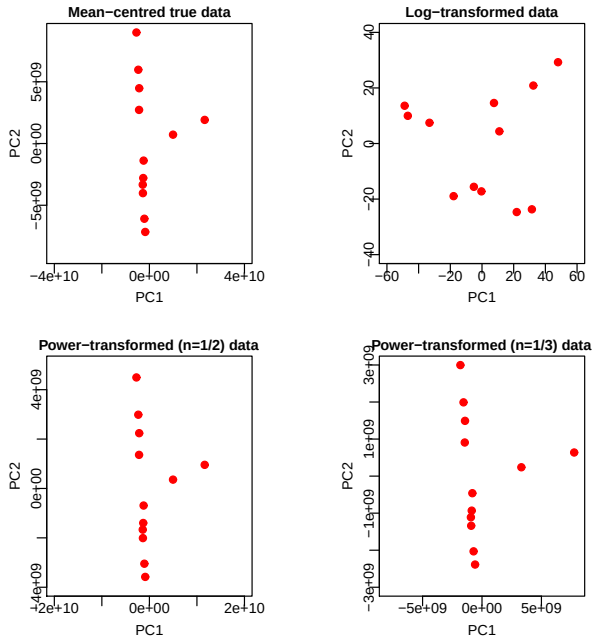


Figure 4.21: PC1 vs PC2 scores plots for the transformed Libya data set, using mean-centring, log transformation and power transformation with n=1/2 and n=1/3.

a histogram of the Libya data, with a normal curve superimposed. This data is skewed strongly to the right (positively skew) and the log data looks more normal than in the other cases for the true and power transformed data (for both n values). The effect of the log and the power transformation on the data as a means to correct for heteroscedasticity is shown in Figure 4.23. Compared to the true data, the power transformation was not able to remove heteroscedasticity. The log transformation was able to remove heteroscedasticity, however only for the metabolites that are present in high concentrations. Since the Libya data set has many zero values, we also added one for all values to ensure they are positive when taking the log. The purpose of applying the previous methods is to remove or reduce any uninduced variation. From the results above from the Libya data set, no element transformation was chosen since the results indicated no significant improvement in PCA results by using any of the two transformation techniques. Hence we deal with

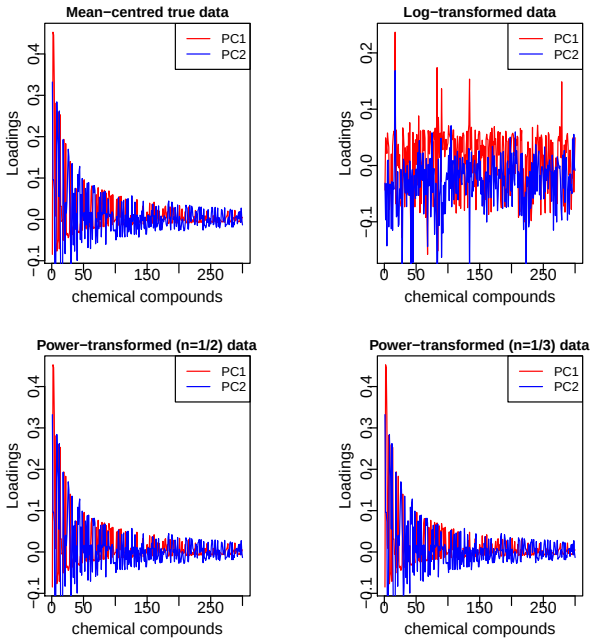


Figure 4.22: PC1 vs PC2 loadings plots for the transformed Libya data set, using mean-centring, log transformation and power transformation with n=1/2 and n=1/3.

the raw Libya data for further analysis.

The effect of column-scaling on the PCA scores and loadings of the Libya data set can be seen in Figures 4.24 and 4.25 respectively. The scores plots in Figure 4.24 indicate some differences among the scores of the six scalings of the Libya data set. Additionally, the scores on PC1 and PC2 for *standardisation* and *range* have a similar shape. Concerning the loadings plots in Figure 4.25, the majority of loadings on PC2 for *vast* scaling are negative peaks. The loadings of the true and Pareto-scaled data on both PCs have similar shapes. The different pre-treatment methods resulted in different effects. For instance, *standardisation*, *range*, *vast* and *level* scaling showed many large peaks, while after Pareto-scaling fewer large peaks were present. It is evident that the Pareto-scaling result is likely to be more interpretable in practice for identifying important metabolites. To verify this result, we will look at PCA results, and we will use the command *prcomp* in *R*, which uses singular value decomposition (SVD) (R Core Team, 2013; Shlens, 2003).

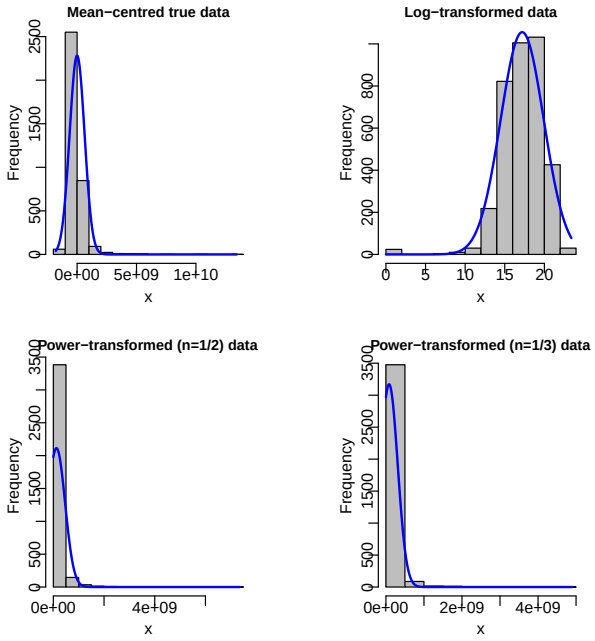


Figure 4.23: Histograms of Libya data set values, with a normal curve superimposed.

From Table 4.5, it is observed that Pareto-scaling performed much better than the other pre-treatment methods, as the first two PCs explain 71.78% of the total variation of the Libya data set, which is much more than for any of the other methods. Therefore, this data set was mean-centred and Pareto-scaled prior to using PCA.

pre-treatment methods	Standardisation	Range	Pareto	Vast	Level
% of variance of PC1	37.87	40.24	54.87	47.16	33.46
% of variance of PC2	19.55	19.58	16.91	15.80	21.80
Cumulative %	57.42	59.82	71.78	62.68	55.25

Table 4.5: Percentage of variance explained by the first two PCs of the Libya data, using different scaling approaches. The best method is shown in red.

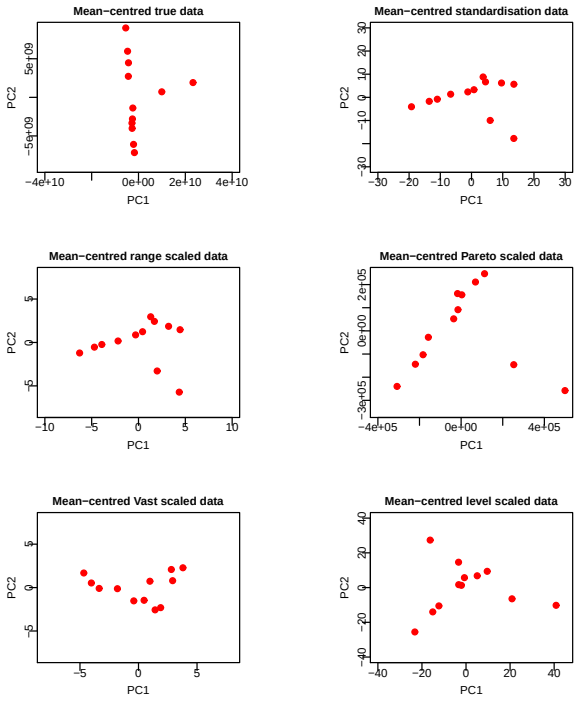


Figure 4.24: PC1 vs PC2 scores plots for the scaled Libya data set, using mean-centring, Standardisation, Range, Pareto, Vast and Level scaling.

4.6 Conclusions

In the introductory part of this thesis, the main aspects of generating metabolomics data and information about propolis have been discussed. The analysis of metabolomics data is achieved with multivariate statistical techniques. Therefore the application of such techniques to metabolomics data was also briefly mentioned in this part. The problem to be researched is to investigate statistical techniques that can be used in the analysis of metabolomics data. Also, all propolis data sets to be used in the statistical analyses of the problem were given. To generate a metabolomics data set from the propolis samples taken from different areas and different colonies, an analytical technique must be used, which is almost exclusively used to generate metabolic profiles, and this was described in detail, Mass Spectrometry (MS). For comparative purposes, the main advantages and

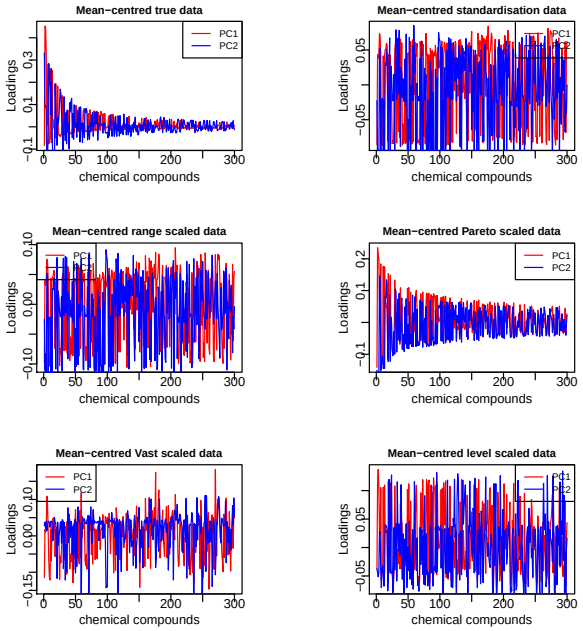


Figure 4.25: PC1 vs PC2 loadings plots for the scaled Libya data set, using mean-centring, Standardisation, Range, Pareto, Vast and Level scaling.

disadvantages of using different such analytical techniques are given.

Before any chemometrics analysis takes place, it is most of the time necessary to process the generated metabolomics data to remove or reduce to acceptable levels the amount of systematic variation in the data, that is, to make the data more suitable for the statistical analyses to follow. There are two stages in the preparation of the data, which are the pre-processing and the pre-treatment.

Pre-processing is concerned with the cleaning of the generated signals, from problems such as overlapping peaks, baseline drifts, signal phasing and existence of an extremely large number of metabolites in the data. A range of methods to overcome such problems was briefly described, as the spectra used in this project were generated by MS. However, the available spectra had been signal-processed by Dr. David Watson and his team. Therefore there was no need to apply any pre-processing techniques.

Pre-treatment is used in the second stage of data processing to remove or reduce any unin-

duced variation (due to sampling, sample work-up and analytical measurement errors) as much as possible and if it exists, heteroscedasticity of the data. Description of the most well-known methods was given with respect to the three ways that the techniques can be applied to the data: transformations of the elements of the data matrix, row-scaling and column-scaling. Two different approaches for transforming the elements of a data matrix were discussed, i.e. the log and power transformations and these were also used on the data. The results indicated that there was no conclusive (if any) improvement to be made to the data sets I, II, III, IV and the Libya data by using either of the two transformation techniques.

Scaling methods (both row and column scaling) were classified to *centring*, scaling based on *data dispersion* and scaling based on *average values*. The advantages and disadvantages of applying the pre-treatment methods were discussed, as well as graphical representations (PC scores and loadings) of the effect on the first two PCs of applying these to metabolomics data sets I, II, III, IV and the Libya data.

Using these results, and considering the type of data to be analysed, the pre-treatment methods chosen for data sets I, II, III, IV and the Libya data used for exploratory analyses and clustering in chapter 5-8, were *mean centring* with *Pareto* scaling of the columns, with no need in this case to scale to a constant total the rows of the data matrix. More specifically, the elements of each column in the data matrix were transformed by subtracting the column mean from each element, then dividing by the square root of the standard deviation, effectively making the columns more comparable to each other in the various analyses in Chapters 5-8. Conclusions about the data in terms of the PCA are given in Chapter 5.

In the next few chapters (5-9), the research will focus on the application of the most commonly used unsupervised multivariate techniques to the metabolomics data described in chapters 3, with the processing (centring methods and scaling) mentioned above. These include both linear and nonlinear dimension reduction and visualisation methods including

PCA and Multidimensional Scaling (MDS)/ Sammon's Non-linear Mapping (NLM). In addition, unsupervised clustering techniques, i.e. Hierarchical Clustering Analysis (HCA), k-means, and Self-Organising Maps (SOM), will be reviewed and applied to the selected metabolomics data sets.

Chapter 5

Unsupervised Techniques

After investigating pre-processing and pre-treatment methods for metabolomics data, we now consider commonly used multivariate technique for analysis of metabolomics data. The main problem is that we have data with high dimensions and we need a technique to reduce the dimensionality. This chapter describes the most important and commonly used unsupervised technique for the reduction of the dimensionality of the data, which is the Principal Component Analysis (PCA) method. Section 5.1 gives an overview of the need for PCA in this context. PCA is described in Section 5.2, and the theoretical background of the PCA technique is provided in Subsection 5.2.2. The application of PCA in metabolomics is presented in Sections 5.3, 5.4 and 5.5. Section 5.6 gives the conclusion.

5.1 Overview

The enormous amounts of data created by high-resolution MS spectra are as a result of the information contained in biological metabolomics data. With regards to metabolomics, a raw MS profile consists of as many as 500 metabolites, which are referred to as variables

here. As in other research, in our study it is necessary to establish potential relationships and/or correlations between the various variables. The level of difficulty and complexity associated with obtaining the required results is correlated with the amount of information available for analysis. To exemplify, the first data set described in Chapter 3, obtained from the Aberdeenshire propolis, contains 921 variables (compounds) for 27 samples, where 3 samples come from each hive (or colony). As such, it is observed that at higher resolutions where many metabolites are being introduced to the problem, it is difficult to properly examine and analyse the data. Thus, it is important to utilise statistical techniques to increase the possibility of determining potential similarities or differences between the various samples in the data. Therefore, it is necessary to reduce the dimensionality of the input space of the data to a smaller number of dimensions: commonly 2 or 3 dimensions are used. By doing so, the results of pattern recognition analyses of the data can be graphically illustrated.

PCA is the most commonly used unsupervised technique for the reduction of the dimensionality of the data, in metabolomics data applications and in Chemometrics and other studies in general. It finds orthogonal linear combination of the input variables which explain as much variation in the input data space as possible. The direction of these principal components does not use any response variable (Y). Therefore, it is an unsupervised approach. This technique is reviewed in Section 5.2. The application of PCA in metabolomics is presented in Section 5.4, 5.5 and 5.6, while the theoretical background of the PCA technique is provided in Subsection 5.2.2.

5.2 Data reduction

5.2.1 Overview

Principal components analysis (PCA) is the main tool used for data reduction. The PCA method creates a new set of variables as orthogonal linear combinations of the original variables in a data set (Horgan, 2000). PCA is a statistical method that aims to reduce the dimensionality, p, of a data space (Diamantaras and Kung, 1996; Olive, 2017). There is a possibility to describe the data and analyse the underlying structure of the data variance, by using a smaller number, m, of independent variables. However, the correlations between the original (observed) variables determine the intrinsic dimensionality, m, of the data. The higher the correlations, the smaller the number of independent variables needed. Following this, without the variation of the data being lost, the p observed variables can be represented as m functions of the observed independent variables (components), where $m < p$.

In common practice, PCA is performed on a data matrix that has rows for samples and columns for variables (i.e. R-mode PCA). On the other hand, Q-mode PCA is performed on the transposed data matrix to study a correlation between samples instead of descriptors (i.e. variables) (Legendre and Legendre, 2012). Conventionally, in most of the literature, R-mode PCA is applied to classify or discriminate between samples by uncovering relationships among variables, in which the loadings plot aids identification of important variables from a list of manifest variables, whereas the scores plot is used to identify sample clusters. On the other hand, loadings and scores matrices from Q-mode PCA are used for identification of important samples and to inspect clustering of variables, respectively (Legendre and Legendre, 2012). As well as its use for dimensionality reduction, PCA can be used before regression analysis.

PCA is a very common unsupervised technique in metabolomics and has been widely utilised for the extraction of important data descriptors. In addition, it has been widely used for reducing the dimensionality of the input space. Several studies of "omics" profiles of samples such as Janzekovic and Novak (2012) have given a detailed description of the application of PCA in bionomics studies; with emphasis placed on the metabolomics and metabolic profiles of samples.

However, a major shortcoming of PCA based on covariance matrices is the sensitivity of the principal components to the units of measurement used for each x variable. To clarify, if large differences exist between the variances of the variables, the variables with the largest variances will tend to dominate the first few PCs. As such, it might become particularly problematic to use PCs on a covariance matrix if the variables are recorded according to different measurement scales, the exception to this being if there is a strong conviction that the units of measurements chosen are the only ones that make sense. Nonetheless, even with this condition satisfied, using the covariance matrix will not produce very revealing PCs as long as the variables possess very differing variances. This is because the PCs are likely to recreate the variables with the largest variance. In addition, the PC scores might be difficult to interpret due to the differently scaled variables.

The standardised version of the covariance matrix is the correlation matrix. Since most analysis uses variables with different measurement scales, the analysis of the correlation matrix allows for these differences in measurement scales to be taken into account. Additionally, there is the possibility that variables measured using the same scale have different variances, which could cause problems when applying PCA. The use of the correlation matrix avoids this problem. From a statistical point of view, the analysis of the correlation matrix is favoured due to the correlation coefficients being insensitive to variations in the dispersion of data and so, producing better-defined factor structures (Tinsley and Tinsley, 1987).

Generally speaking, the results obtained using the covariance matrix differ from those obtained from the correlation matrix. Nonetheless, an important area of metabolic pro-

filing is toxicology and drug development. Keun (2006) shows how PCA of metabolic profiles could aid the detection of drug toxicity specific biomarkers. In addition, the use of PCA as a projection method is emphasised, and the analysis of the covariance matrix, provided that the variables are commensurable, is illustrated. In this thesis the results from chapter 4 indicated that the best pre-treatment for the propolis data sets (I, II, III, IV and Libya) was mean- centering and Pareto-scaling prior to using PCA. A further area of interest is the use of PCA in various fields of ecological research, such as in a determination of enterotypes of the human gut microbiome on the basis of specialisation of their trophic niches (Arumugam et al., 2011). In aquatic habitat studies, PCA has been applied for evaluation of aquatic habitat suitability, its regionalisation, analysis of fish abundance, their seasonal and spatial variation, and lake ecosystem organisation change (Ahmadi-Nedushan et al., 2006; Blanck et al., 2007; Catalan et al., 2009). However, it has also often been applied in analysing farming system changes (Amanor and Pabi, 2007).

5.2.2 Theoretical framework

There are several methods used to carry out PCA, and we will use the command *prcomp* in R, which uses singular value decomposition (SVD) (R Core Team, 2013). We have a data set that can be represented as an $(n \times p)$ matrix X. Typically, obtaining the sample principal components is done through the following steps (Shlens, 2003):

1. First, subtract the variable mean from each of the data dimensions; this gives an $(n \times p)$ matrix, P, with $(i,j)^{th}$ element $(x_{ij} - \bar{X}_j)$.

2. Divide the square root of the variable standard deviation from each of the data dimensions; this gives an $(n \times p)$ matrix, Q, with $(i,j)^{th}$ element $(x_{ij} - \bar{X}_j/\sqrt{S_j})$.

3. Calculate the correlation matrix of Q, $C_x = \frac{1}{n}Q^TQ$.

4. Calculate the SVD or the eigenvectors of the correlation matrix C_x (let A denote

the $(p \times p)$ matrix of eigenvectors):

$$C_x = USA^T$$

where U and A are orthonormal, and S is a diagonal matrix. The column vectors of U are taken from the orthonormal eigenvectors of QQ^T, and for A are taken from the orthonormal eigenvectors of Q^TQ, and ordered right to left from the largest corresponding eigenvalue to the smallest. Also, the principal components of Q are the eigenvectors of C_x, or the rows of A.

5. Order the eigenvectors by eigenvalues from highest to lowest and choose a set of significant eigenvectors (the first few) which explain a large part of the total variation. By doing this, the dimension of the matrix of eigenvectors is reduced to $(p \times m)$ instead of $(p \times p)$, where m is the number of eigenvectors chosen. Denote the matrix containing the chosen set of eigenvectors by A^*.

6. Finally, the required principal components scores, Z, are computed as:

$$Z = UA^*.$$

The eigenvalues of correlation matrix C_x, $\lambda = \lambda_1 \geq ..... \geq \lambda_p$, are the roots of

$$|C_x - \lambda I|$$

where $\sum_{i=1}^{p} \lambda_i = tr(C_x), \quad \prod_{i=1}^{p} \lambda_i = det(C_x),$

and the eigenvectors of C_x, $a = a_1 \geq ..... \geq a_p$, are the normalised eigenvectors satisfying

$$C_x a_i = \lambda_i a_i.$$

$$a_i^T a_j = \begin{cases} 1 & : \quad i = j \\ 0 & : \quad i \neq j. \end{cases}$$

Principal components (principal component axes) are computed in a way such that the sum of squared orthogonal distances between the data and their projections on the axes is a minimum. In other words, the principal component axes minimise the sum of squared errors in all the variables (Shlens, 2003).

The functional form of representation is a linear transformation (combination). The general transformation necessary for a p-dimensional space can be represented as:

$$\begin{aligned} y_1 &= a_{11}x_1 + a_{12}x_2 + \cdots + a_{1p}x_p \\ y_2 &= a_{21}x_1 + a_{22}x_2 + \cdots + a_{2p}x_p \\ \vdots &= \quad \vdots \quad + \quad \vdots \quad + \cdots \; + \quad \vdots \\ y_p &= a_{p1}x_1 + a_{p2}x_2 + \cdots + a_{p}x_p \end{aligned}$$

or in a matrix form as $Y = AX$, where Y is the p-dimensional component column vector $(y_1, y_2, ..., y_p)^T$, X is the p-dimensional column vector $(x_1, x_2, ..., x_p)^T$ and A is the $(p \times p)$ matrix of coefficients a_{ij},

$$\begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1p} \\ a_{21} & a_{22} & \cdots & a_{2p} \\ \vdots & \vdots & \cdots & \vdots \\ a_{p1} & a_{p2} & \cdots & a_{pp} \end{pmatrix}.$$

Geometrically, the reduction of the dimensionality of the data space represents the projection of the vector X onto an m-dimensional space. Commonly, the space is a line, a plane or a 3-dimensional space which allows for the data to be represented graphically, as well as to express the correlation between the variables.

Based on the transformation equations, it is observed that the elements of matrix A must be calculated in order for the components to be evaluated. Two factors of variation in the transformation procedure exist: the variation due to the reduction of the dimensionality

of the data space (projection error) and the variation of each component. Since individual components are illustrated graphically by a line in a specific direction, the projection error is the variation around the line. On the other hand, the variation of the component is represented by the spread of the data along its line. We aim to minimise the projection error while simultaneously amplifying the component variation. Through the use of the covariance matrix or the correlation matrix of the variables, we can evaluate the eigenvalues and eigenvectors of these matrices (Diamantaras and Kung, 1996). The estimated eigenvectors are the columns of matrix A, and hence their eigenvalues are the loadings of the components on the observed variables. Assuming matrix A is expressed as:

$$A = (a_1, a_2, ..., a_p) \tag{5.1}$$

where a_i is the column vector with elements $(a_{i1}, a_{i2}, ...a_{ip})^T$, $i = 1, ..., p$, we can then estimate each component y_i by the column vector a_i. The correlation matrix of the components, C_y, can be written in terms of the correlation matrix of the observed variables, C_x, as

$$C_y = AC_xA^T. \tag{5.2}$$

Since these components are independent; the derived correlation matrix is diagonal, with elements given by the computed eigenvalues, λ_i. From equation (5.2), for each vector a_i of matrix A, the following equivalent expression can be derived:

$$C_xa_i = \lambda_i a_i \tag{5.3}$$

where λ_i is the eigenvalue for component y_i (Diamantaras and Kung, 1996). Since it is crucial to identify the components in decreasing order of variation, the first component will have the maximum variance, $Var(y_1)$. This is calculated from equation (5.3) as the maximum eigenvalue, λ_1. The corresponding eigenvector to this eigenvalue, a_1, gives the direction of the first component axis, on which the data is projected. The spread of the projected data on this first component axis is given by λ_1. Solving equation (5.3) for the

second largest component variation, $Var(y_2)$, a_2 and λ_2 are obtained. The eigenvectors are orthogonal to each other: to exemplify, $a_i^T a_j = 0$ for every i and j. This process is repeated until the last component y_p is found. It is also worth noting that the total component variation is equal to the variation of the observed variables, that is

$$tr(C_x) = \sum_{i=1}^{p} \lambda_i \quad . \tag{5.4}$$

In addition, the portion of the total variation explained by a component, y_i is expressed as:

$$\frac{\lambda_i}{\sum_{i=1}^{p} \lambda_i} \quad . \tag{5.5}$$

As the number of PCs is p, dimension reduction is achieved by choosing a smaller number m of PCs corresponding to the largest eigenvalues, to represent the data but still explain a large part of the total variation.

5.2.3 Data Suitability for PCA

It is very important to evaluate the amount of information contained within a data set before accepting any results obtained by PCA. If the amount of information contained within the data set is very large, meaning that the descriptors in the data set are not correlated, it would be unnecessary to apply PCA. This is because there will not be a significant data reduction by using PCA.

Two statistical approaches which can be used to confirm the suitability or lack of it of a data set for PCA, are the Gleason - Staelin statistic (Jackson, 2003) and the normalised entropy, $\tilde{S}$, of a data set (Cangelosi and Goriely, 2007). The Gleason - Staelin statistic is expressed by:

$$\alpha = \sqrt{\frac{\|T\|^2 - p}{p(p-1)}} \tag{5.6}$$

for the correlation matrix T, where p is the dimensionality of the data set. i.e. the number of original variables in the data, and

$$\|T\|^2 = \sum_{i=1}^{p}\sum_{j=1}^{p} t_{ij}^2 \quad . \tag{5.7}$$

The statistic becomes the following when the covariance matrix B is used:

$$\alpha = \sqrt{\frac{\|B\|^2 - \sum_{i=1}^{p}(b_i^2)^2}{\sum_{i=1}^{p}\sum_{j\neq i}^{p}(b_i b_j)^2}} \quad . \tag{5.8}$$

Finally, α takes values in the range of [0,1]. If the variables within the data set are more correlated, this gives higher α values. For value 0, the variables are totally uncorrelated, and as such, it would be pointless to apply PCA to the data. However, for value 1, there is perfect correlation among the variables and the dimensionality of the data space is 1. So we look for a value between 0 and 1, in order to apply PCA.

The normalised entropy, $\tilde{S}$, of a data set is expressed as:

$$\tilde{S} = -\frac{1}{log_2 N}\sum_{i=1}^{N} r_i log_2 r_i \tag{5.9}$$

where r_i is the proportion of total the variation shown by component i, and N is the number of components that PCA calculated (the rank of the X data matrix). The statistic $\tilde{S}$ also takes values in the range [0,1]. Again, the higher the value, the more information is contained within the data set and the less necessary it would be to use PCA. For value 1, all variables in the data set are completely uncorrelated, and as such, the data space dimensionality is proportional to the number of variables in the data set. This is in contrast to when the value is 0, where all variables are completely correlated, and the dimensionality is 1; therefore only one component is essential to express all the information. So we look for a value between 0 and 1, in order to apply PCA.

The information dimension related to the normalised entropy can be defined as:

$$I_0 = \prod_{i=1}^{N} r_i^{-r_i} \tag{5.10}$$

where; r_i is the proportion of the total variation shown by component i, N is the number of components that PCA calculated, and it can be used to assess how many components to preserve.

5.2.4 Choosing the Number of Components to Retain

Establishing the maximum number of PCs required is a crucial part of the PCA process. There are debates in the literature regarding the most appropriate technique for estimating the number PCs to use. Yet, none of the proposed approaches to this problem is suitable for every possible situation. There are large numbers of stopping rules, which can be categorised into groups. To exemplify, the two most commonly used stopping rules are rules based on confidence intervals, such as parallel analysis and re-sampling methods (Besse and De Falguerolles, 1993; Horn, 1965), as well as those based on average test statistic values, such as the broken stick and Velicer′s MAP (Ferre, 1995; Peres et al., 2005; Velicer, 1976). The broken stick and parallel analysis approaches are expanded upon below, as well as the simple scree plot:

- **Broken stick**

  This method is established on the idea that by randomly dividing the total variance of a multivariate data set, the distribution of the eigenvalues follows a broken stick distribution. The concept is that if a line segment is randomly split into n pieces, the anticipated value of the length of the k^{th} piece can be expressed as:

$$E_k = \frac{1}{n} \sum_{u=k}^{n} \frac{1}{u}. \tag{5.11}$$

The component k is retained if the eigenvalue of the component k is larger than the respective anticipated value, E_k, of the broken stick distribution (Cangelosi and Goriely, 2007; Legendre and Legendre, 1998; Peres et al., 2005). However, same caution is needed when utilising this stopping rule. According to Cangelosi and Goriely (2007), there is a tendency to underestimate the appropriate number of principal components when using this method. As such, it would be advantageous to compare the results obtained from this method to those obtained from other stopping rules; to avoid retaining too few principal components.

- **Parallel Analysis (PA)**

PA is another method of determining the number of components in principal components (Zwick and Velicer, 1986). The PA stopping rule was introduced by Horn (Horn, 1965) and is based on the production of data sets with random uncorrelated normally distributed variables of a similar size to the original data. The method applies PCA to the generated random data set and retains the eigenvalues for individual principal components. This process is repeated many times, .e.g. 1000 times. The percentile intervals of eigenvalues for individual components are then estimated at confidence levels such as 95%. In the case that the obtained observed values exceed those of the calculated intervals at the chosen confidence level, the component is retained. The argument is that a component should be retained if its eigenvalue is greater (at the 95^{th} percentile) than that obtained at random. It is worth highlighting that as a result of this analysis relying on the normality of the produced data, it may not be the most suitable method for cases where the observed data is not normally distributed. In such instances, non-parametric re-sampling techniques such as bootstrap methods may produce more robust observations (Besse, 1992; Daniel, 1992).

- **Scree Plot**

  Another approach is based on Cattell's Scree test (Cattell, 1966), which involves visual graphical representation of the size of the eigenvalues. In this technique, the eigenvalues are presented in descending order and linked with a line. Then, the graph is examined to determine the point at which the last significant drop or break takes place, in other words, where the line levels off. The logic behind this technique is that this point divides the essential or significant factors from the minor or trivial factors. This technique has been criticised for its subjectivity, since there is not an objective definition of the cut-off point between the essential and trivial factors. Some cases may present different drops and possible cut-off points, such that the graph may be ambiguous and difficult to interpret. Zwick and Velicer (1986) mention that when analysing how examiners understand the Scree test, the outcomes can be very varied, depending on the nature of the solution and the training received by the examiners. Jackson (2003) suggest using one more component after the break in the line. Anther approach, which can be used with the scree plot is to choose the number of components according to how many eigenvalues are above 1. This is the Kaiser criterion.

The three techniques above (broken stick, parallel analysis and the scree plot) will be used in the analyses within this study, and results from all the propolis data sets (I, II, III, the combined data set IV and the Libya data) will be compared using these techniques.

5.3 Effect of Outliers on PCA

Two types of outliers can affect PCA, i.e. leverage points and orthogonal outliers. The former is related to their score distance, which is their projection's distance from the

centre of the PCA space, and the latter to their orthogonal distance to the space defined by the PCs. For example, Figure 5.1 shows that point 1 has a large orthogonal distance to the PCA space and this kind of outlier destabilises the estimation of PCA. Point 2 has a large orthogonal distance and a large score distance, which means the projection is far away from the centre of PCA space, and this kind of outlier is called a high (bad) leverage point because they can affect the estimation of PCA space. On the other hand, point 3 is called a good leverage point because it has a large score distance but a small orthogonal distance, and this kind of outlier stabilises the estimation of PCA (Varmuza and Filzmoser, 2016).

The score distance, SD, of a sample i is given by equation (5.12):

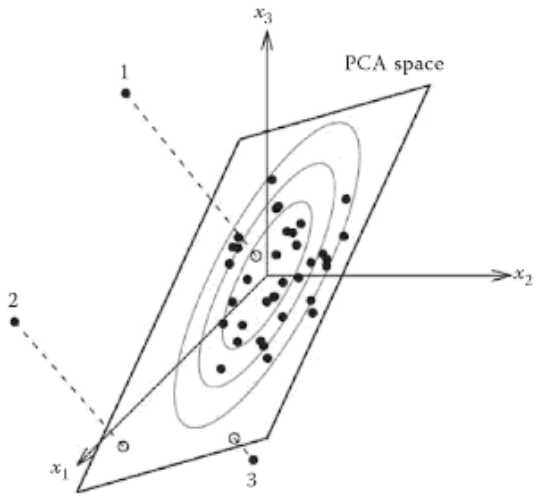


Figure 5.1: Visualisation of the different kinds of outliers that can affect PCA (Varmuza and Filzmoser, 2016).

$$SD_i = \left[\sum_{j=1}^{N_{pc}} \frac{f_{ij}^2}{\upsilon_j} \right]^{\frac{1}{2}} \tag{5.12}$$

where N_{pc} is the number of PCs forming the PCA space, f_{ij} are the elements of the score matrix and v_j is the variance of the j^{th} PC (Varmuza and Filzmoser, 2016). Assuming that the data is multivariate normally distributed, the squared score distances can be approximated by a chi-square distribution, $\chi^2_{N_{pc}}$, with N_{pc} degrees of freedom. A cut-off

value for the actual score distance can be taken as the 97.5% quantile, $\sqrt{\chi^2_{N_{pc},0.975}}$. If the score distance of a sample is larger than this cut-off point, then the sample is considered to be a leverage point. The orthogonal distance, OD, of a sample i is defined as:

$$OD_i = \|x_i - Lf_i^T\| \tag{5.13}$$

where x_i is the i^{th} sample of the centred data matrix, L is the loadings matrix using N_{pc} PCs and ${f_i}^T$ is the transposed score vector of sample i for N_{pc} PCs (Varmuza and Filzmoser, 2016). A cut-off value for the orthogonal distance is computed by Hubert et al. (2005), using the Wilson-Hilferty approximation for a chi-square distribution. That is, the distribution of $OD^{\frac{2}{3}}$ is approximately normal, with the centre (mean) and spread (variance) of the values being robustly estimated, e.g. using the median and the median absolute deviation (MAD) respectively. The cut-off value is then computed as $(median(OD^{\frac{2}{3}}) + MAD(OD^{\frac{2}{3}}).z_{0.975})^{\frac{2}{3}}$, where $z_{0.975}$ is the 97.5% quantile of the standard normal distribution. If the orthogonal distance of a sample is higher than the cut-off value, then the sample is considered as an orthogonal outlier.

To summarise, if an orthogonal outlier with large orthogonal distance also has a large score distance (so it is a leverage point), then the sample is a bad leverage point, as it can affect negatively the correct estimation of the PCA space. A leverage point that also has a small orthogonal distance but still is an orthogonal outlier, with a large score distance, is a good leverage point, as it can stabilise the estimation of the PCA space.

5.4 Application of PCA on Data Sets I, II and III

5.4.1 Overview

As described earlier, the production and pre-processing processes of the mass spectrometry (MS) profiling of the propolis samples were done by Dr. David Watson or his team, from

the Institute of Pharmaceutical and Biomedical Sciences at the University of Strathclyde (SIPBS). The original samples originate from Aberdeenshire, Fort William and Dunblane. The data sets I, II and III are made up of 27 samples with 921 variables, 17 samples with 511 variables and 9 samples with 498 variables for data sets I, II and III (Aberdeenshire, Fort William and Dunblane) respectively. The existence or lack of variation in the data's composition reflects local and regional variation in the composition of the propolis samples, caused by differences in the forage sources available to the honey bees. Each data set I, II and III was column-scaled by mean-centring and Pareto scaling, to enable more comparability in the samples. The column-scaling was done by dividing each element by the square root of the standard deviation of the variable (metabolite), thus transforming the variables for data I, II and III to be in the same unit of measurement.

The prospect of decreasing the number of variables to a smaller number of components, without suffering loss of important information from the original data, will be evaluated. Using PCA may also reveal any relationships between the samples and the variables (components). Examining loadings plots may also help to determine important metabolites in the PCs. It may also be possible to determine potential clusters from the resulting scores plots. Upon establishing the required PCs for the variables in the data sets and using appropriate statistical criteria, it would be possible to identify how many components would be needed to represent most of the variations in the data sets. The important information within the original data would be retained to high accuracy, by these PCs. This could also aid the identification of any determined clusters of samples and/or variables (chemical components). Finally, the characteristics of the variables (chemical compounds) will be evaluated to clarify potential relationship(s) between the variables (chemical compounds) and the sample. In other words, we will examine the chances of PCA identifying any natural clusters of samples, with characteristics of the important compounds in these samples.

5.4.2 Data Suitability of PCA for data sets I, II and III

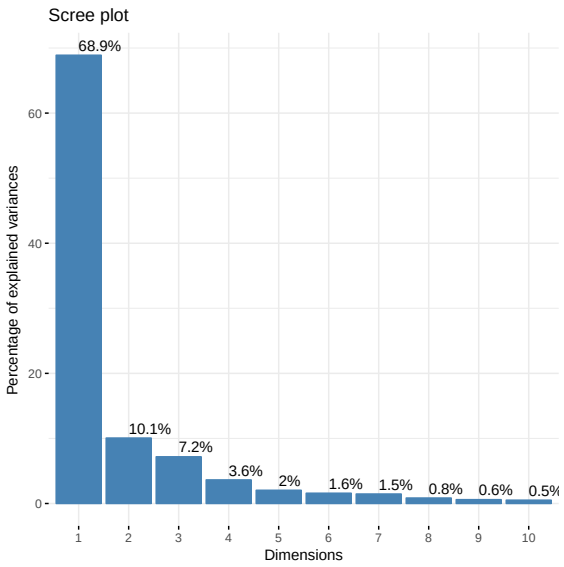
Before doing any PCA analysis, it is necessary to test the suitability of the data sets I, II and III for PCA. The Gleason-Staelin statistic and the normalised entropy are calculated using equations (5.6) and (5.9) respectively. The value of the Gleason-Staelin statistic using the correlation matrix is 0.465 for Aberdeenshire, 0.481 for Fort William and 0.480 for Dunblane, which indicate that the metabolites (variables) are sufficiently correlated to justify data reduction using techniques such as PCA. In addition, the normalised entropies for the Aberdeenshire, Fort William and Dunblane datasets are 0.393, 0.459 and 0.625 respectively, which also means that the metabolites are sufficiently correlated to apply PCA, with the dimensionality of the data being close to 4, 4 and 4 respectively (the value of the information dimension is $3.66 \approx 4$, $3.67 \approx 4$ and $3.95 \approx 4$ respectively (see equation 5.10)). Both statistics confirm that these data sets I, II and III (Aberdeenshire, Fort William and Dunblane) are suitable for PCA analyses.

5.4.3 Choosing the Number of Components to Retain

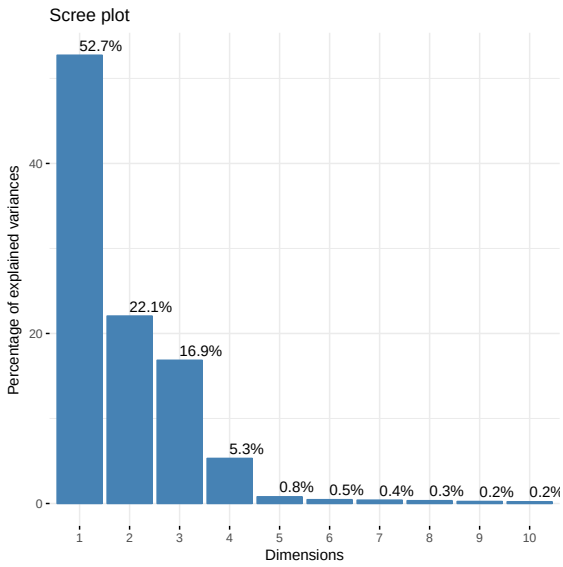
After confirming the suitability of data sets I, II and III for PCA, the next step in PCA is to identify the number of principal components to retain. This is done for the mean-centred and Pareto-scaled data sets I, II and III.

The percentages of the total variation in the pre-treated data sets I, II and III explained by the first ten principal components can be seen in Figure 5.2. Figure 5.2 (i) for data set I shows that about 79%, 90% and 95% of the total variation is explained by 2, 4 and 7 PCs respectively, Figure 5.2 (ii) for data set II shows that about 75%, 92% and 97% of the total variation is explained by 2, 3 and 4 PCs respectively, and Figure 5.2 (iii) shows that about 77%, 85% and 91% of total variation is explained by 2, 3 and 4 PCs respectively for data set III.

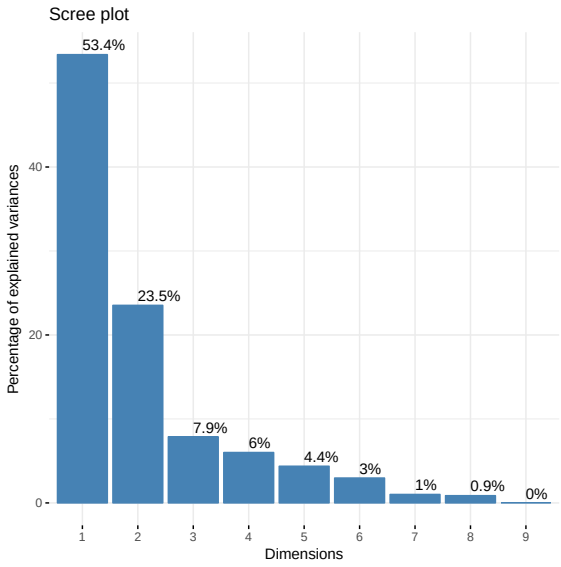
Table 5.1 contains the standard deviation, the percentages of total variance explained



(i) Aberdeenshire



(ii) Fort William



(iii) Dunblane

Figure 5.2: Percentages of the total variation in data sets explained by the first ten PCs.

and the cumulative percentages of variance for the first ten PCs of the data sets. These detailed results for the variance of the PCs indicate that 2 components could be retained for further analyses, as they explain a large part of the variation in the data sets I, II and III, approximately 79%, 75% and 77% respectively, while the variation of the remaining

components is likely to be due to measurement and instrumentation errors.

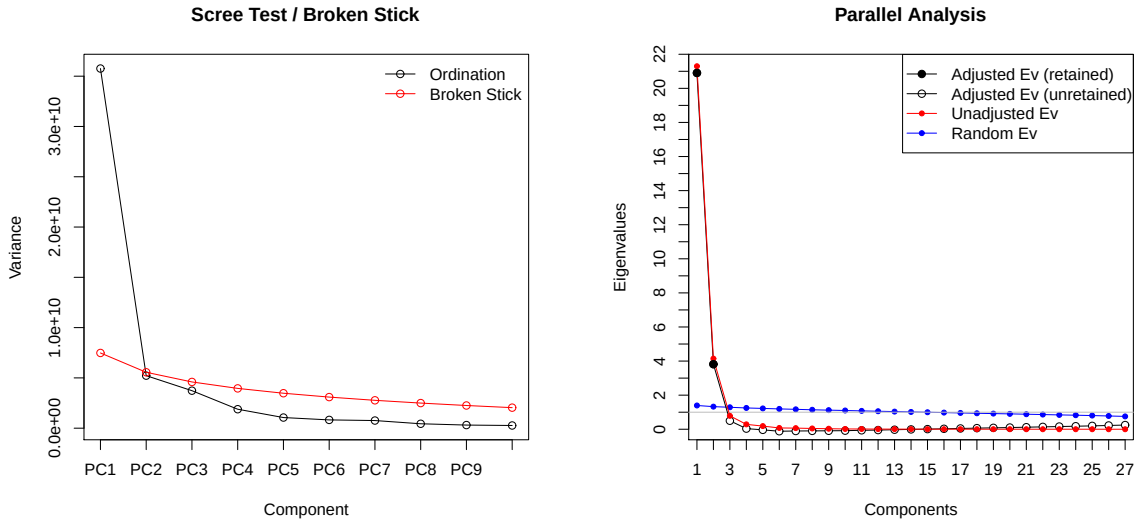
To confirm these findings, as described in Subsection 5.2.4, the broken stick, parallel

Aberdeenshire			
PCs	Standard Deviation	% of Variance	Cumulative %
PC1	1.891e+05	68.90	68.90
PC2	7.227e+04	10.10	79.00
PC3	6.106e+04	7.18	86.18
PC4	4.342e+04	3.63	89.81
PC5	3.252e+04	2.03	91.84
PC6	2.869e+04	1.59	93.43
PC7	2.755e+04	1.46	94.89
PC8	2.095e+04	0.85	95.74
PC9	1.759e+04	0.60	96.34
PC10	1.637e+04	0.52	96.86
Fort William			
PCs	Standard Deviation	% of Variance	Cumulative %
PC1	1.222e+05	52.73	52.73
PC2	7.904e+04	22.05	74.79
PC3	6.909e+04	16.85	91.64
PC4	3.877e+04	5.30	96.94
PC5	1.497e+04	0.79	97.73
PC6	1.129e+04	0.45	98.18
PC7	1.059e+04	0.39	98.58
PC8	9.713e+03	0.33	98.91
PC9	8243.6510	0.24	99.15
PC10	7.656e+03	0.20	99.36
Dunblane			
PCs	Standard Deviation	% of Variance	Cumulative %
PC1	8.934e+04	53.38	53.38
PC2	5.931e+04	23.53	76.91
PC3	3.430e+04	7.86	84.78
PC4	2.997e+04	6.00	90.79
PC5	2.556e+04	4.37	95.16
PC6	2.106e+04	2.96	98.12
PC7	1.229e+04	1.01	99.13
PC8	1.137e+04	0.87	100
PC9	2.693e-11	0.00	100

Table 5.1: Standard deviation, percentage of total variance explained, and cumulative percentages of variance for the first PCs of data sets I, II and III. (There are 9 samples in data set III).

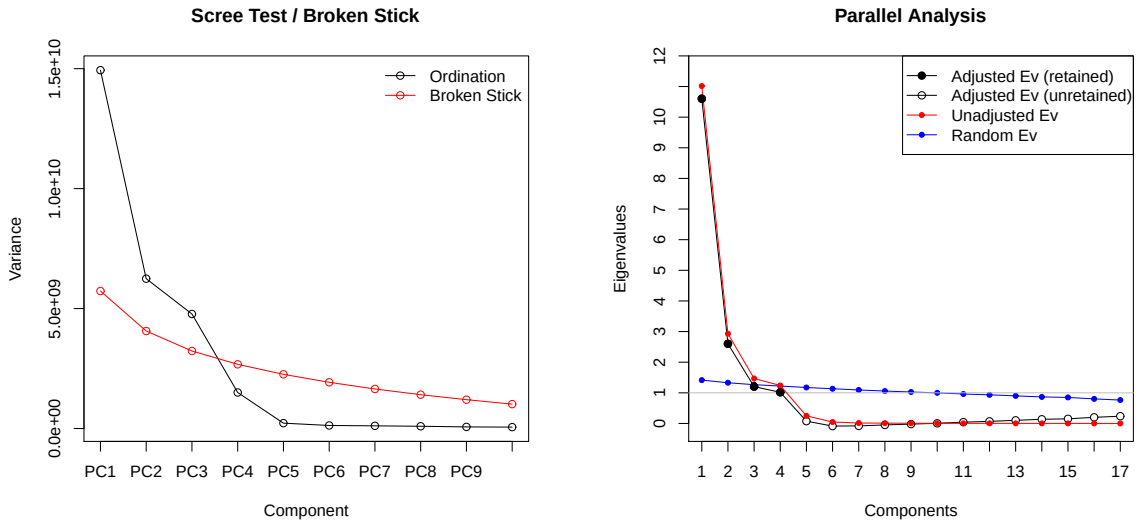
analysis and scree plot stopping rules will be used to identify the appropriate number of principal components. An illustration of the broken stick model can be seen in Figures 5.3, 5.4 and 5.5 (left Figures) for data sets I, II and III respectively. Figure 5.3 (left) for data set I shows that only one component should be retained, as only one eigenvalue is larger than the expected value of the broken stick distribution (red line) (though the second one is close). Cattell's scree test is also depicted in Figure 5.3 (black line in the left Figure), confirming that at most three components should be retained (using one more component after change in direction of the line ((Jackson, 2003)). Figure 5.4 (left) for data set II, shows that three components should be retained, as only three eigenvalues are larger than the expected values of the broken stick distribution (red line). Cattell's scree test is shown in Figure 5.4 (black line in the left Figure), confirming that at most three components should be retained. Also, the broken stick model seen in Figure 5.5 (left) for data set III, shows that only two components should be retained, as only two eigenvalues are larger than the expected values of the broken stick distribution (red line). Cattell's scree test shown in Figure 5.5 (black line in the left Figure), confirms that at most four components should be retained.

Parallel analysis was performed using the mean and the 99^{th} centile estimates for the calculation of the confidence intervals, and different numbers of random sets of up to 200 per variable. All runs retained 2, 4 and 1 components of data sets I, II and III respectively, independently of the confidence intervals and number of iterations used. The parallel analysis plot in Figure 5.3, 5.4 and 5.5 (right figures) for data sets I, II and III illustrates the adjusted and unadjusted eigenvalues and suggests that 2, 4 and 1 components should be retained respectively for the three data sets. The unadjusted eigenvalues are the eigenvalues of the observed data from an unrotated PCA. The random eigenvalues are the estimated eigenvalues (using either the mean or centile approaches) from 27630, 15330 and 14940 iterations of data sets I, II and III respectively, which is the default number of iterations, given by $30*$ *number of variables*, as used by the *R* function *paran*() to perform parallel analysis. The adjusted eigenvalues are given by the



(i) Scree and Broken Stick plots (ii) Parallel Analysis plot

Figure 5.3: Stopping rules for the number of PCs for Aberdeenshire; EV denotes eigenvalue.

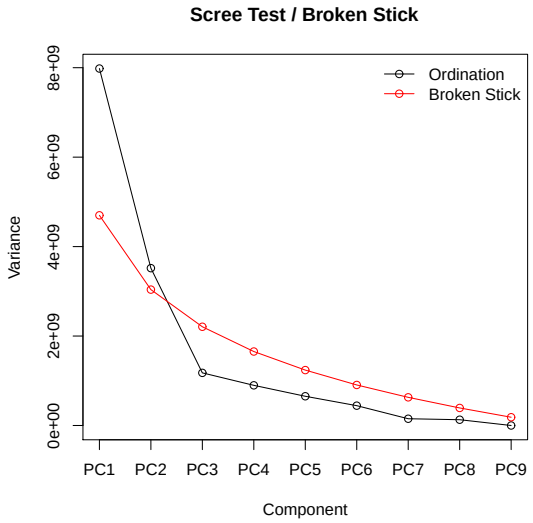


(i) Scree and Broken Stick plots (ii) Parallel Analysis plot

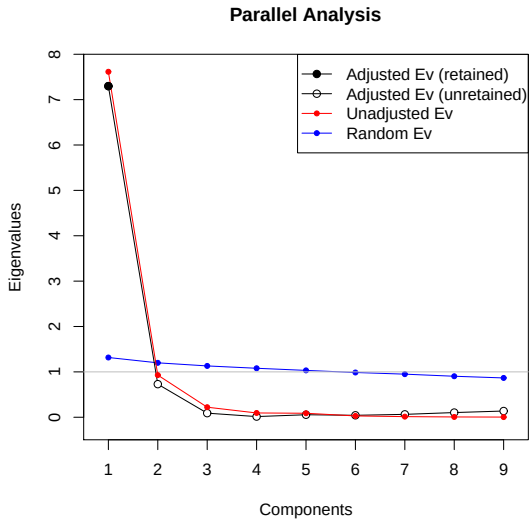
Figure 5.4: Stopping rules for the number of PCs for Fort William; EV denotes eigenvalue.

adjustment in equation (5.14) (Franklin et al., 1995)

$$AdjustedEig = UnadjustedEig - (SimulatedEig - 1), \tag{5.14}$$



(i) Scree and Broken Stick plots



(ii) Parallel Analysis plot

Figure 5.5: Stopping rules for the number of components of Dunblane; EV denotes eigenvalue.

and retained if their values are greater than 1.

Finally, in Table 5.2 a comparison of the results for a number of stopping rules can be seen. Table 5.2 shows that about 90%, 95% and 99% of the total variation is explained by 4, 7 and 18 PCs respectively for data set I. Of these, 18 is rather a large number of PCs to keep, while 2 to 4 PCs is probably a good number. The result of retaining 1 PC, from broken stick for the Aberdeenshire data set, although being the smallest number, if chosen will not be particularly interesting, and probably important information contained in the second PC will not be considered. Therefore, despite the first component explaining approximately 68.9% of the total variation, one PC is most probably not the appropriate number of PCs to retain. Retaining the first two or three principal components allows for proper graphical representation of the data set I and easier identification of any natural clustering in the structure of the input space. From the results here, four PCs could also be justified.

For data set II, Table 5.2 shows that about 90%, 95% and 99% of total variation is

explained by 3, 4 and 7 PCs respectively. Table 5.2 shows that 3 or 4 PCs should be retained for data set II. As the first two components explain approximately 75% of the total variation, two PCs is most probably an appropriate number of PCs to retain, but three can also be justified in this case.

Also, Table 5.2 shows that about 90%, 95% and 99% of the total variation is explained by 4, 5 and 7 PCs respectively for data set III. The results from parallel analysis and broken stick show that 1 or 2 PCs respectively should be retained. Again, the result of retaining 1 PC will not be particularly interesting, and probably important information contained in the second PC will not be considered. So, two PCs is most probably the appropriate number of PCs to retain, however three or four could also be justified in this case.

Number of Components retained			
Stopping rule	Aberdeenshire	Fort William	Dunblane
Parallel Analysis	2	4	1
Brocken Stick	1	3	2
Cattel's Scree Test	3	3	4
90% of Variance	4	3	4
95% of Variance	7	4	5
99% of Variance	18	7	7
Information Dimension	4	4	4

Table 5.2: Comparison of various stopping rules for data sets I, II and III.

5.4.4 Diagnostic Plots of PCA For Data I, II and III

Having identified that the first two or three PCs generally should be retained for further analyses, a graphical representation of data sets I, II and III (Aberdeenshire, Fort William and Dunblane) structure is the next step in the PCA analysis.

The PC scores (concerning the samples) and loadings (concerning the variables) can be plotted in many ways to give a visual summary of the data. These can be in 1, 2 or 3 dimensions. Plotting the PC scores is usually the first step in describing the data

graphically. The 1-dimensional scores plot is essentially a bar chart, where, for a selected PC, each score is plotted against sample number. It is often useful to re-order the sample in ways that can facilitate better the interpretation of the scores. Selecting a suitable order of the samples should indicate clearly in the bar chart if a specific PC is influenced by a specific grouping of the samples. One way of indicating a particular grouping of the samples (in this case hive samples) is by using colour. In the cases of propolis data sets I, II and III, the groupings of the samples will be defined by their chemical compounds of interest (peak areas) which may be expected to reflect location of the hives and the local environment for foraging. In a 2-dimensional scores plot, the scores on one PC are plotted against those of another PC for each sample. This is usually done for the first 2-3 PCs, which more often than not are sufficient to explain most of the variation in the data. In this case, the samples are plotted using the values of the scores as coordinates. This type of plot may indicate which of the PCs appears to be the best discriminator for a specific grouping of the samples. The groupings are usually represented by a different symbol and/or colour. In the cases of data sets I, II and III, whenever 2-dimensional score plots are used here, different colours will represent the groupings of samples according to their chemical characteristics, and also every three samples are from the same hive (or colony); for example samples 1, 2 and 3 relate to one hive and so on. Finally, if the results of the 1 and 2-dimensional plots are not conclusive, 3-dimensional scores plots can be used, such that each axis of the plot represents one PC. Colouring of the samples can be applied in an analogous way to that of the 1 and 2-dimensional plots.

In propolis data sets I, II and III, 3-dimensional plots will be used only if the results of the lower dimensional plots justify it. A general visual summary of data sets I, II and III can be seen in Figures 5.6, 5.8 and 5.11 respectively. The former scores plot describes most of the information in the data sets, as approximately 78.7%, 74.8% and 76.9% of the total variation are explained by the first two PCs of the Aberdeenshire, Fort William and Dunblane data, as shown in Table 5.1, 5.2 and 5.3 respectively, therefore investigating these two PCs of Aberdeenshire, Fort William and Dunblane should be sufficient.

Now we will look at the results of the PCA in more detail below:

- **Data Set I (Aberdeenshire)**

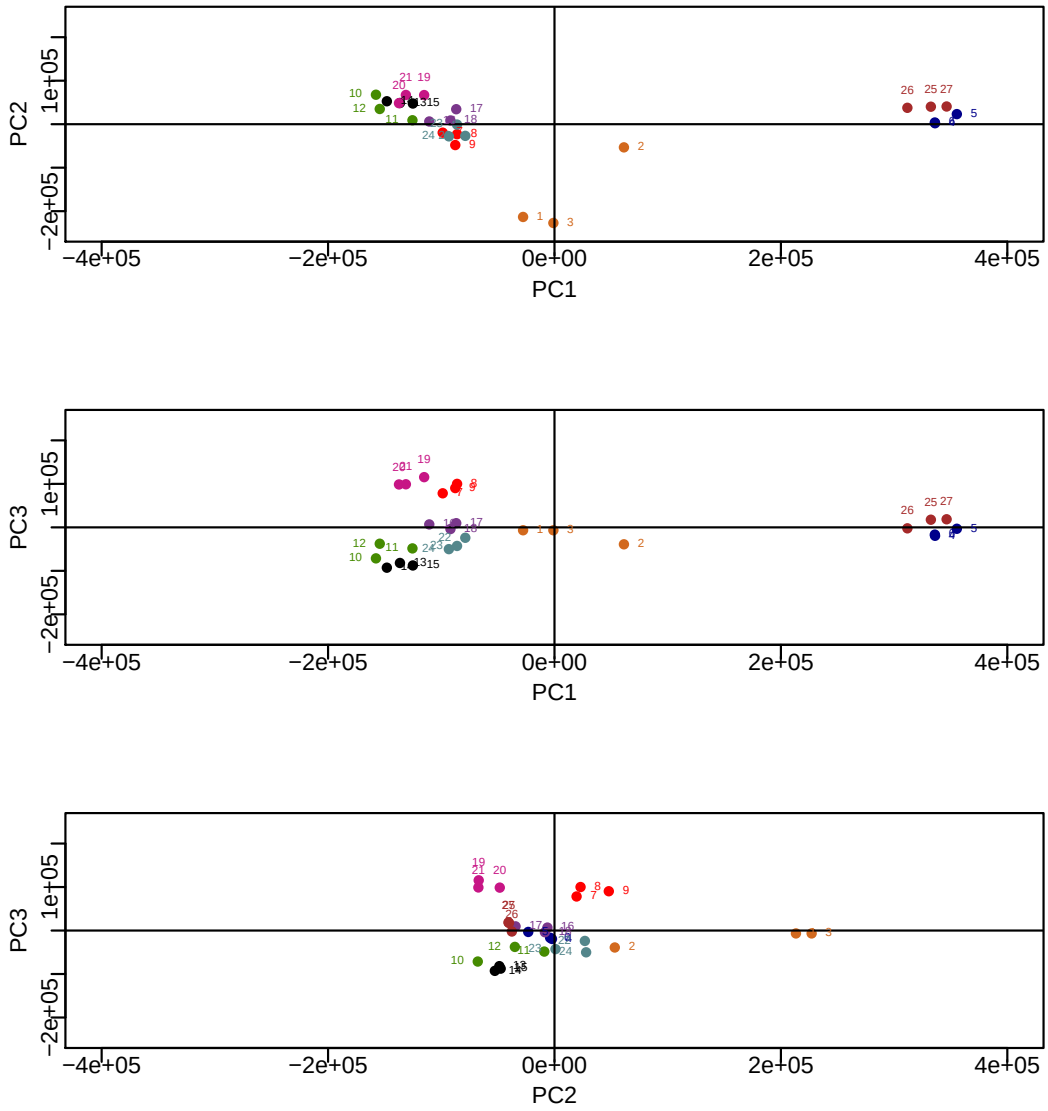


Figure 5.6: Scores plots of the Aberdeenshire data for the first three PCs, superimposed with the sample numbers (hives) and each colour indicates the same hive.

A general visual summary of data set I can be seen in Figure 5.6. The most inter-

esting plots of the three are the scores plot for the first two components and that of the pair PC1 and PC3. In general components PC1 and PC3, can discriminate the samples, where every three samples indicate the same hive (or colony) because the chemical analysis was repeated for three samples from each hive (or colony). As can be seen, samples 1, 2 and 3 are close, samples 4, 5 and 6 are close and so on. The first two PCs score plot describes most of the information in data set I, approximately 79% of the total variation, therefore it is necessary to investigate these two PCs.

The most interesting plot for data set I is the scores plot for the first two PCs, PC1 and PC2, which indicates that there are two samples, number 1 and 3, having an extreme negative score in PC2, which seem to be outliers in PC2. Sample 2 is a bit different from samples 1 and 3 from the same colony. Also, no samples have a high positive score in PC2. Samples with numbers 4, 5, 6, 25, 26 and 27 have a very high positive score in PC1. As PCA is affected by outliers, it is important to confirm whether these samples are outliers or not. Biochemically, these samples look different from the rest.

Diagnostic plots using the score distance and the orthogonal distance for the Aberdeenshire samples can be seen in Figure 5.7. The cut-off values for the score and the orthogonal distance are equal to 2.72 and 149563.6 respectively. It can be seen that there are samples with score distance higher than the cut-off, namely samples 1, 2, 3, 4, 5, 6, 25, 26 and 27, all mentioned above. On the other hand, for the orthogonal distances, there are no points with orthogonal distance higher than the cut-off. Hence samples 1, 2, 3, 4, 5, 6, 25, 26 and 27 are good leverage outliers and there were no bad leverage outliers (high score distance and high orthogonal distance). However, removing these samples from the data set and re-running the analyses showed that there was no effect from the inclusion of these samples in the PCA, as the results were similar. Therefore, the original data set of the selected 27 samples can be used for further analyses. The sample coordinates concerning the

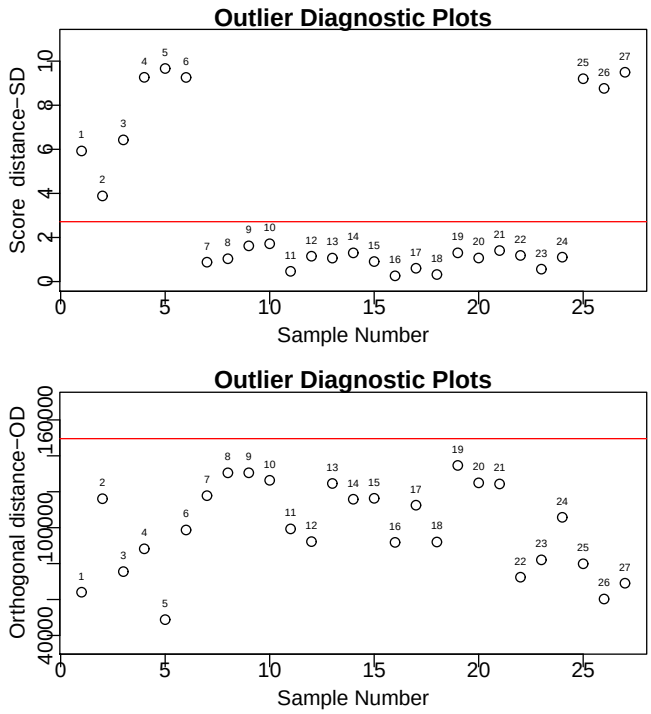


Figure 5.7: Outlier diagnostic plots using the score distance (SD) and the orthogonal distance (OD) for Aberdeenshire. The numbers in the plots are the numbers of the 27 samples. The horizontal lines in the two plots represent the cut-off values, such that any point above these lines is a leverage point (top plot) or an orthogonal outlier (bottom plot).

first two PCs, explain approximately 79% of the total variation in the Aberdeenshire data. Therefore, the first two PCs of the Aberdeenshire data will be used in further investigation.

In general, it might be proposed that the samples in the middle of the PCA plot are tending to use several different sources of propolis or more common sources, whereas the samples towards the periphery of the plot may focus on more restricted or unusual sources (Saleh et al., 2015). Therefore samples 1, 3, 4, 5, 6, 25, 26 and 27, from 3 different colonies, may relate to more restricted sources and the remaining samples to different sources of propolis.

- **Data Set II (Fort William)**

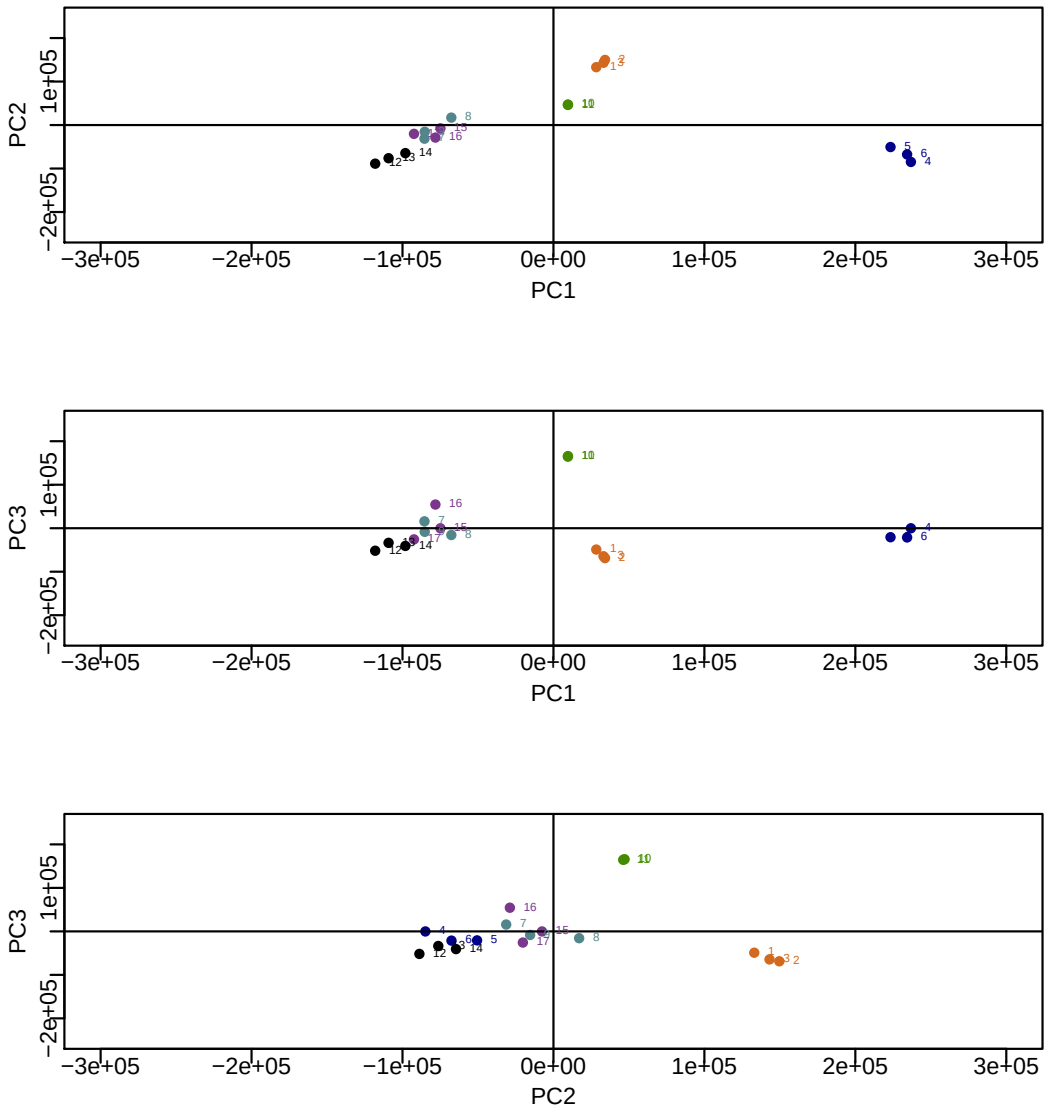


Figure 5.8: Scores plots of the Fort William data for the first three PCs, superimposed with the sample numbers (hives) and each colour indicates the same hive.

From Figure 5.8 for Fort William, in general PC1 and PC2 can discriminate the samples from different colonies. The first two PCs explain approximately 75% of the total variation of the data. The first two scores plots indicate that there are

three samples, 4, 5 and 6, having a high positive score in PC1, which seem to be outliers, influencing mainly PC1. Samples 1, 2 and 3 have a very high positive score in PC2 and could impact on PC2. No samples have a very negative score in PC2. Diagnostic plots using the score and the orthogonal distance of the Fort William

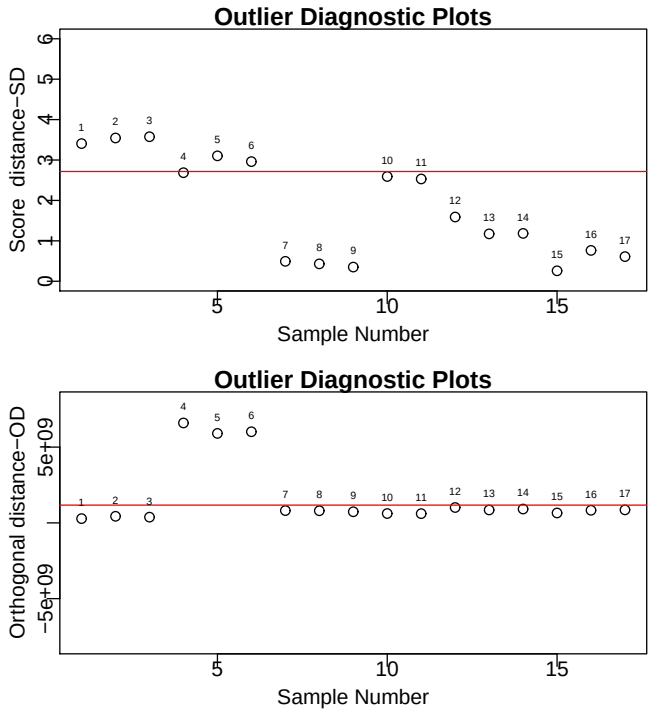


Figure 5.9: Outlier diagnostic plots using the score (SD) and the orthogonal distance (OD) for Fort William data. The numbers in the plots are numbers of the 17 samples. The horizontal lines in the two plots represent the cut-off values, such that any point above these lines is a leverage point (top plot) or an orthogonal outlier (bottom plot).

samples in the data can be seen in Figure 5.9. The cut-off values for the score and the orthogonal distance are 2.72 and 1170822521 respectively. There are some samples with score distance higher than the cut-off, namely samples 1, 2, 3, 5 and 6, and two samples 10 and 11 close to the cut-off value, and one sample with score distance approximately equal to the cut-off value (sample 4). In the case of the orthogonal distances, there are three samples with orthogonal distances higher than

the cut-off, namely samples 4, 5 and 6 (and one sample with orthogonal distance approximately equal to the cut-off value (sample 12)).

Since samples 4, 5 and 6 with a sizeable orthogonal distance also have a large score distance, they can affect the correct estimation of the PCA space negatively. Removing these 3 samples from the data set and re-running the analyses showed that there was an effect from the inclusion of these samples in the PCA, as the results were dissimilar, with an increase of the variance explained by the first two PCs (by approximately 9% compared to the first two PCs of the 17 samples (Table 5.3)). Therefore, the data set of the selected 14 samples rather than all 17 samples can be used for further analyses. Note that we omit samples 4, 5 and 6 from this data, so that now samples 7 and 8 in this reduced data set come from the same hive, but sample 9 comes from a different hive, the same hive as samples 10 and 11. Apart from samples 7 and 8, the other sets of 3 samples, i.e. samples 1-3, 4-6, 9-11 and 12-14, each come from a single hive or colony, so in total 5 colonies are now represented in the Fort William data.

The sample coordinates with respect to the first two PCs, which explain approx-

PCs	Variance for 17 samples	Variance for 14 samples
PC1	52.73	52.79
PC2	74.79	83.83

Table 5.3: Cumulative proportion of variance explained by the first two PCs for the Fort William data before and after excluding outliers.

imately 84% of the total variation in the reduced Fort William data, can be seen in Figure 5.10. Again, it might be proposed that the samples in the middle of the PCA plot are tending to relate to several different sources of propolis, whereas the samples towards the periphery of the plot (samples 1, 2, 3, 7 and 8) may focus on more restricted sources. Figure 5.10 shows the scores plot of the data after omitting the 3 outliers. Again there are some differences in the points.

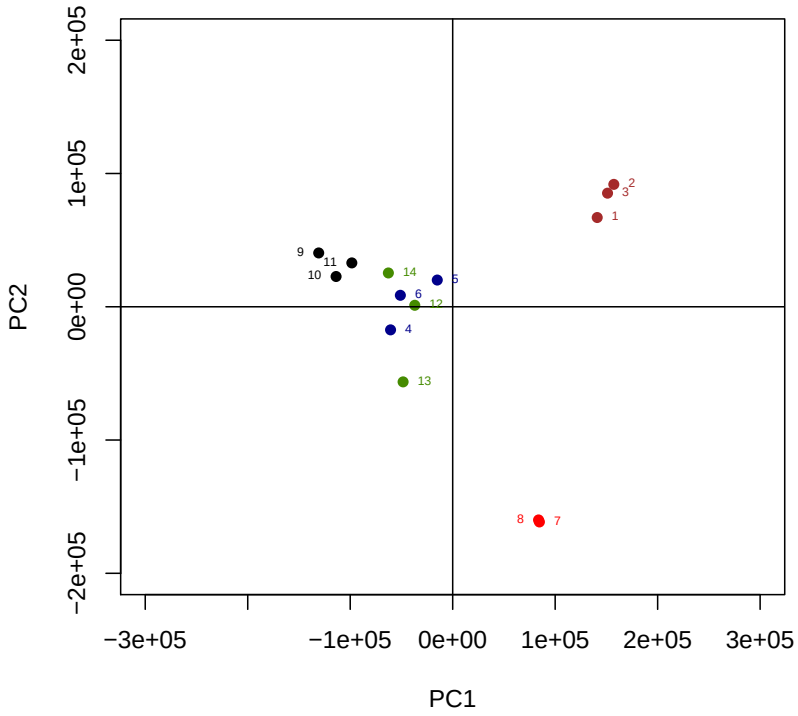


Figure 5.10: Scores plot for the Fort William data for the first two PCs after excluding outliers from the data.

- **Data Set III (Dunblane)**

  In the last data set, from Figure 5.11, no components can discriminate all samples well depending on every three samples from the same hive (or colony). In general there is variation between the samples from any one hive (or colony). The first two PC, explain approximately 77% of the total variation of the data. The first score plot indicates that there are points having slightly high positive scores (samples 8 and 9) and somewhat high negative scores (samples 4 and 5) on PC1, which seem to be outliers, influencing mainly PC1. In the PC2 dimension, points 1, 2, 3 are separated from the rest. Diagnostic plots for the Dunblane samples can be seen in Figure 5.12. The cut-off values for the score and the orthogonal distance are equal

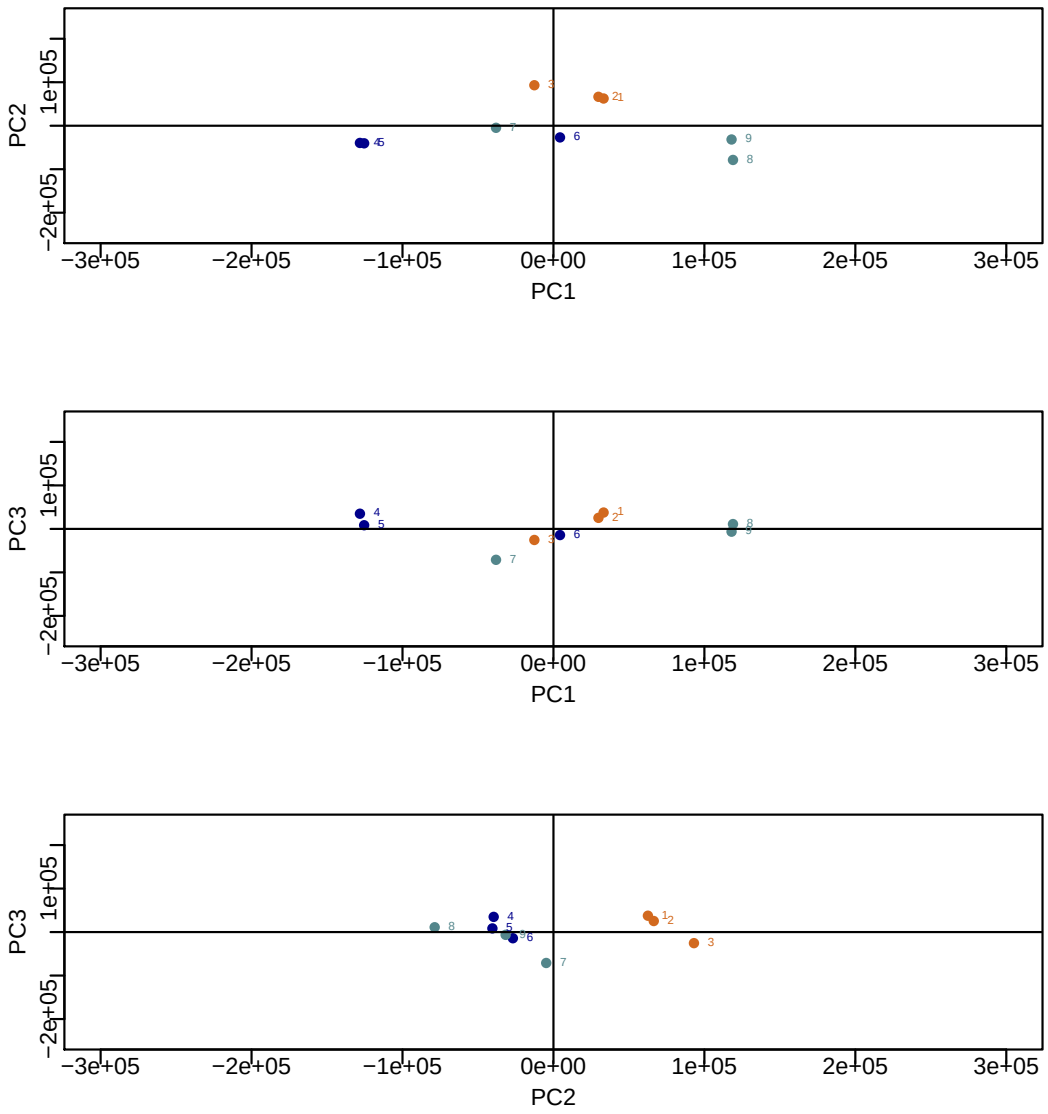


Figure 5.11: Scores plots of the Dunblane data for the first three PCs, superimposed with the numbers of the samples (hives) and each colour indicates the same hive.

to 2.72 and 99660.76 respectively. It can be seen that there are no samples with score distance higher than the cut-off. In the case of the orthogonal distances, there is also no point with orthogonal distance higher than the cut-off. Therefore, the original data set of nine samples can be used for further analyses. If the samples in the middle of the PCA plot are from several different sources of propolis, whereas

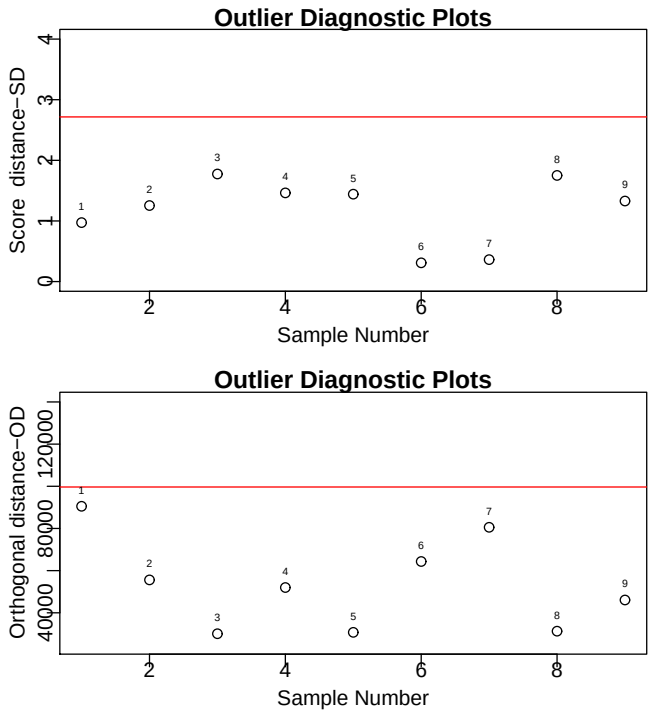


Figure 5.12: Outlier diagnostic plots using the score distance (SD) and the orthogonal distance (OD) for the Dunblane data. The numbers in the plots are the numbers of the nine samples. The horizontal lines in the two plots represent the cut-off values, such that any point above these lines is a leverage point (top plot) or an orthogonal outlier (bottom plot).

the samples towards the periphery of the plot focus on more restricted sources, there is variation in these samples. Samples 1, 2, 3, and 8, 9 and 4, 5 seem different from each other and from the other samples.

5.4.5 Contributions of Variables to PCs for Data Sets I, II and III

Loadings plots can help to provide a general idea of relationships between variables, as well as between samples and variables but are not easy to use with this kind of data

because variables here are represented as names of compounds, many names of chemical compounds in propolis are unknown until now, and there are 921, 511 and 498 compounds for the Aberdeenshire, Fort William and Dunblane data respectively.

In PCA, the scores plot is mainly used to discover groups, while the loadings plot is mainly used to find variables that are responsible for separating the groups. In the loadings plot, we mainly check the points that are further from the origin than most other points in the plot. In general, samples that cluster most closely together in the scores plot are usually well correlated. The loading plots show the contribution of variables to the PCs. The contributions of variables in accounting for the variability in a given principal component are expressed as percentages, and variables that are highly correlated with PC1 (i.e., dimension.1) and PC2 (i.e., dimension.2) are the most important in explaining the variability in the data set. Variables that do not correlate with any PC or are correlated with the last few PCs have a low contribution, and might be removed to simplify the overall analysis.

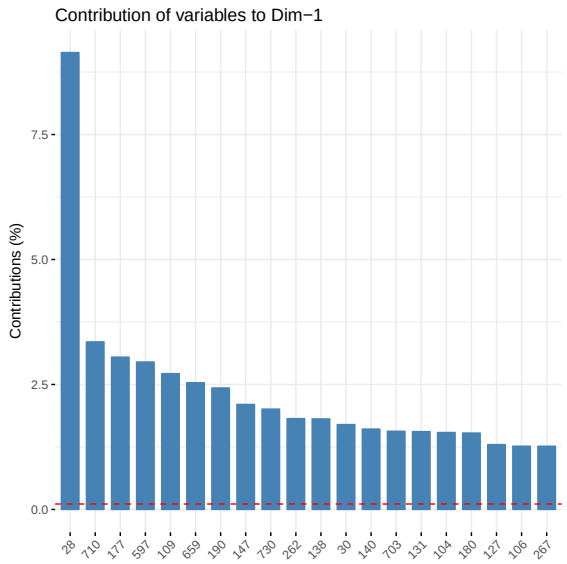
The function $fviz\_contrib()$ from the *factoextra* R package can be used to draw a bar plot of variable contributions. As data sets I, II and III contain many variables, we can decide to show only the top contributing variables. Figures 5.13, 5.14 and 5.15 show the top 20 variables contributing to the principal components for Aberdeenshire, Fort William and Dunblane respectively. The red dashed line on the graph indicates the expected average contribution. If the contribution of the variables were uniform, the expected value would be $1/length(variables)$ for data sets I, II and III. Also, the eigenvalues measure the amount of variation retained by each PC. Table 5.4 shows a comparison of contributions of the variables to the first two PCs of the data sets I, II and III.

As mentioned previously, propolis is a complex mixture made by bee-released and plant-derived compounds. In general, more than 300 constituents were identified in different samples of the data sets I, II and III and new ones are still being recognised during the

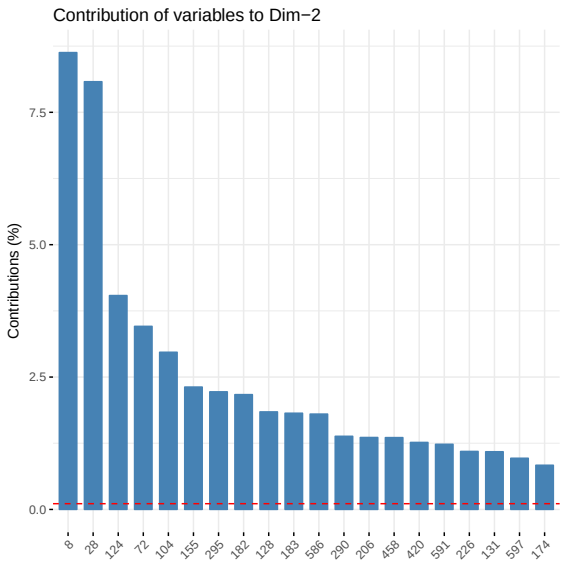
chemical characterisation of new types of propolis. Many constituents are contributing to the first two PCs in the data sets I, II and III, such as cinnamoyl caffeoyl acetyl glycerol and benzoyl hydroxy phenylacetic acid. Also, some constituents are contributing to the first two PCs in the data sets II and III (green colour in Table 5.4), such as coumaric acid, methyl pinobanksin, cinnamoyl caffeoyl acetyl glycerol, benzoyl hydroxy phenyl acetic acid, benzoyl dihydroxyphenylpropionic acid, coumaric acid cinnamyl ether, C16H11O5, caffeic acid hextrieneoate and phenylacetic acid. Some constituents are contributing to the first two PCs in data sets I and II (blue colour in Table 5.4), such as dicaffeoyl acetyl glycerol, cinnamoyl caffeoyl acetyl glycerol, benzoyl hydroxy phenyl acetic acid, coumaroyl feruoyl acetyl glycerol, Hydroxy phenyl acetyl dihydroxyphenylacetic acid and pinocembrin methyl ether. Some constituents are contributing to the first two PCs in the data sets I and III (red colour in Table 5.4), such as benzoyl hydroxy phenyl acetic acid, cinnamoyl caffeoyl acetyl glycerol, prenylated flavonoid and dimethyl pinocembrin benzoate.

From the results in Table 5.4, the composition of the Aberdeenshire propolis samples appears to be fairly different from the Dunblane samples. They differ from each other, but overall the compounds in Table 5.4 are in many cases not the same as the most important variables in the Dunblane samples. Moreover, the compounds from Fort William are different again but closer in character to the Dunblane samples than the Aberdeenshire samples. This result reflects only 20 compounds contributing most to the first two PCs in data sets I, II and III.

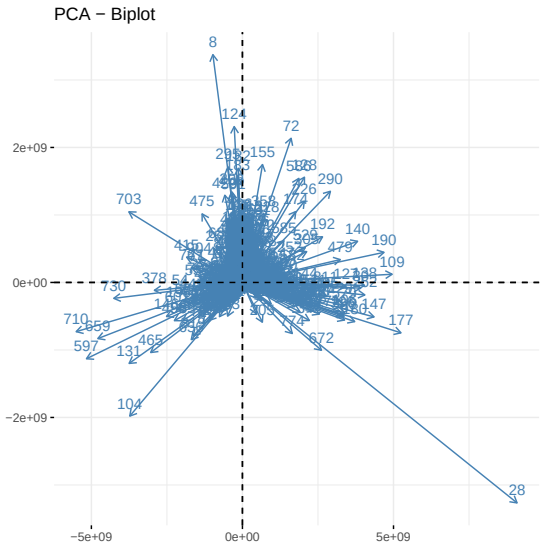
Regarding Figure 5.13 for the Aberdeenshire samples, variables 28, 177, 109, 190 and 147 have the highest positive loadings in PC1 (Figure 5.13, plot (iii), Appendix Table B), thus samples number 4, 5, 6, 25, 26 and 27 will tend to have larger values on these variables in PC1 (Figure 5.13, plot (iv)), as well as variables 8, 124, 72 and 155 in PC2, being observed to have the highest positive loadings in PC2 (Figure 5.13, plot (iii)), thus, samples 10, 19 and 21 will tend to have larger values on these variables (Figure 5.13,



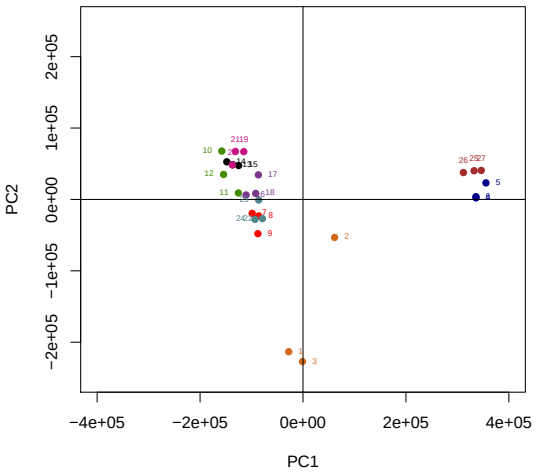
(i) contribution of variables to PC1



(ii) contribution of variables to PC2



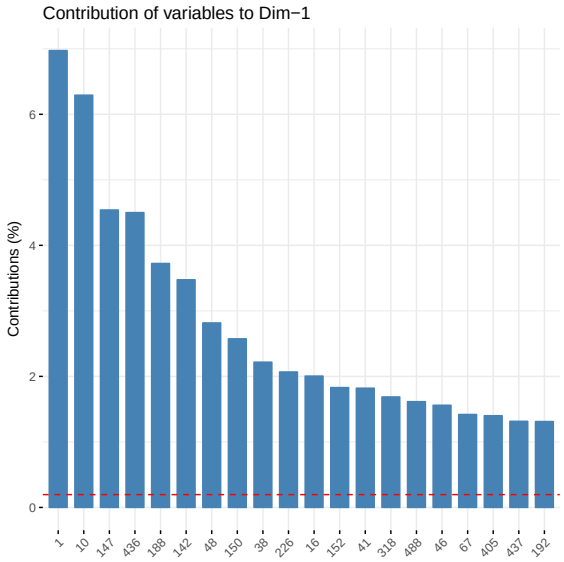
(iii) biplot for first two PCs



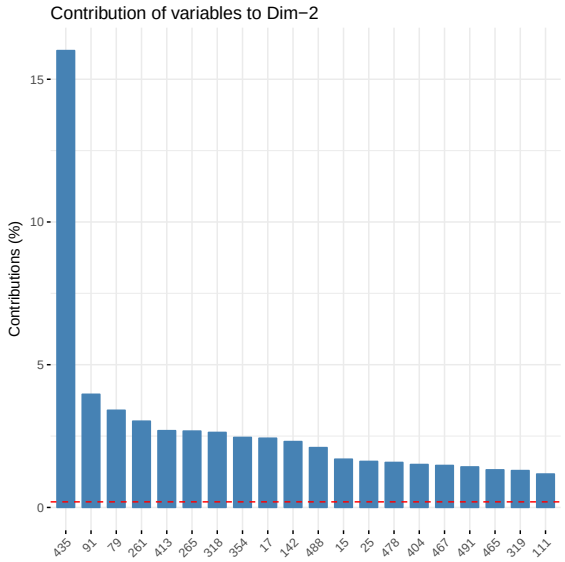
(iv) Score plot for first two PCs

Figure 5.13: Variables contributing to PC1 and PC2 of data set I (Aberdeenshire).

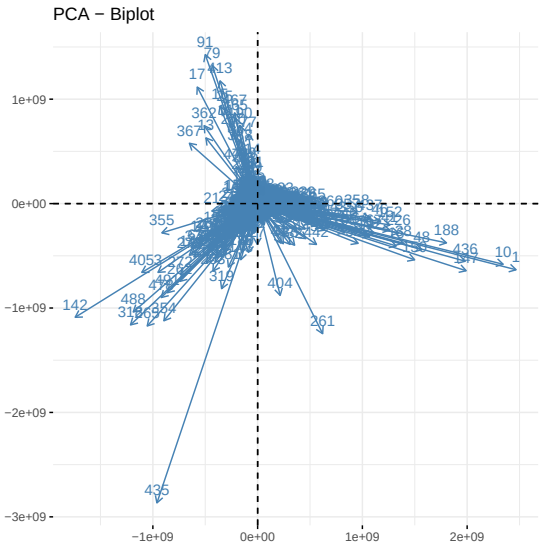
plot (iv), Appendix Table B). On the other hand, variables 28, 104, 131 and 597 have the highest negative loadings in PC2 (Figure 5.13, plot (iii)); thus, samples 1 and 3 will tend to have higher values on these variables in PC2 Figure 5.13, plot (iii), Appendix Table B). A PCA biplot shows both PC scores of samples and loadings of variables. The



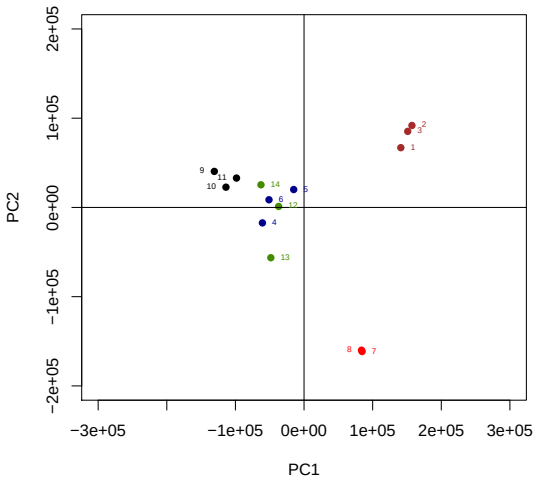
(i) contribution of variables to PC1



(ii) contribution of variables to PC2



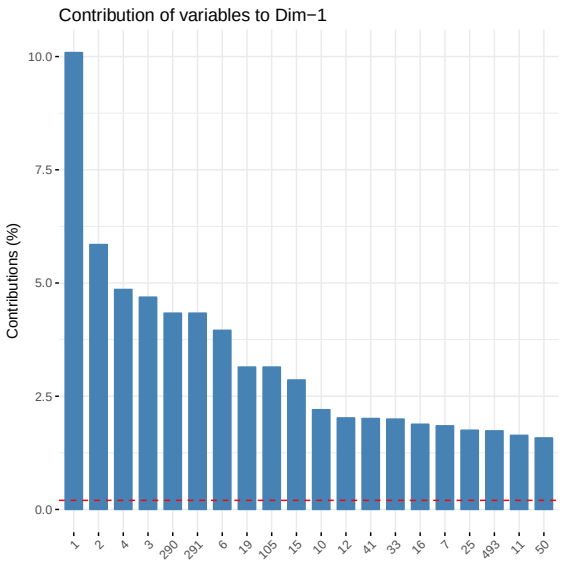
(iii) biplot for first two PCs



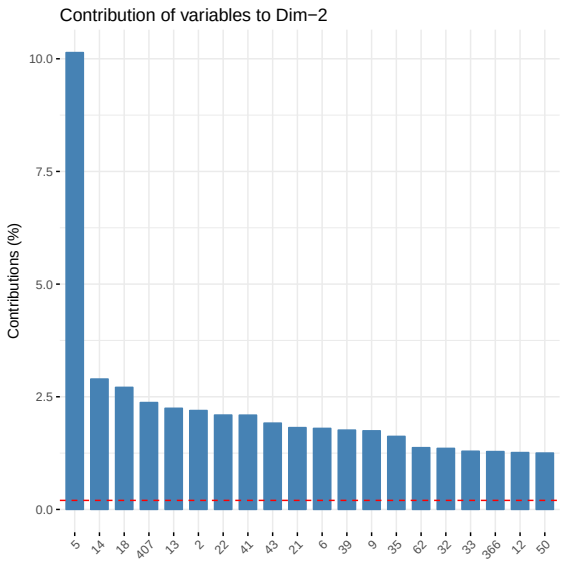
(iv) Score plot for first two PCs

Figure 5.14: Variables contributing to PC1 and PC2 of data set II (Fort William).

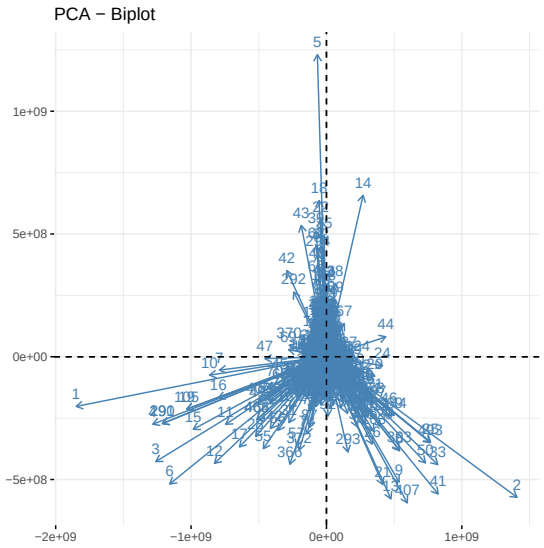
further away these loadings vectors are from a PC origin, the more influence they have on that PC (Kassambara, 2017). Loading plots also hint at how variables correlate with one another: a small angle implies positive correlation such as variables 28 and 672, a large one suggests negative correlation such as variables 28 and 703, and a 90° angle indicates



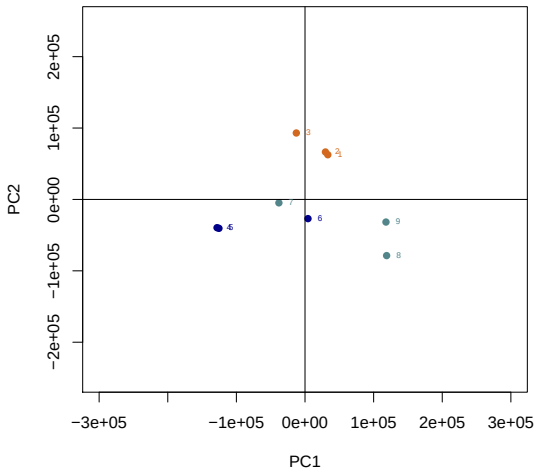
(i) contribution of variables to PC1



(ii) contribution of variables to PC2



(iii) biplot for first two PCs



(iv) Score plot for first two PCs

Figure 5.15: Variables contributing to PC1 and PC2 of data set III (Dunblane).

no correlation between two characteristics such as variable 28 and 104. Since samples 1, 3, 4, 5, 6, 25, 26 and 27 are towards the periphery of the plot (Figure 5.13, plot (iv)), they may focus on a more restricted source, as it was proposed that the samples in the middle of the PCA scores plot are tending to use several different sources of propolis, whereas the

samples nearest the centre of the plot have important compounds from the other outlying samples in moderate amounts.

Regarding Figure 5.14 for Fort William, variables 1, 10, 147, 436 and 188 have the highest positive loadings in PC1 (Figure 5.14, plot (iii), Appendix Table C), thus, samples 1, 2 and 3 will tend to have larger values on these variables in PC1 (Figure 5.14, plot (iv)), as well as variables 142, 318, 488 and 405 being observed to have the highest negative loadings in PC1 (Figure 5.14, plot (iii), Appendix Table C). In the other hand, variables 435, 261 and 265 have the highest negative loadings in PC2 (Figure 5.14, plot (iii), Appendix Table C); thus, samples 7 and 8 will tend to have higher values on these variables in PC2 (Figure 5.14, plot (iv)). The compositions of the Fort William propolis samples appears to be fairly different from the Aberdeenshire samples (see Table 5.4, where the blue colour indicates the common compounds between Aberdeenshire and Fort William, and the green colour indicates the common compounds between Fort William and Dunblane). They differ from each other, but overall the compounds in Table 5.4 are in many cases not the same as the most significant variables in the Aberdeenshire samples. The Fort William samples are rich in compounds putatively identified as sesquiterpene acids. The samples in the middle of the PCA plot are likely to come from several different sources of propolis, whereas the samples on the periphery may focus on a more restricted source (such as samples 1, 2, 3, 7 and 8 in Figure 5.14, plot (iv)).

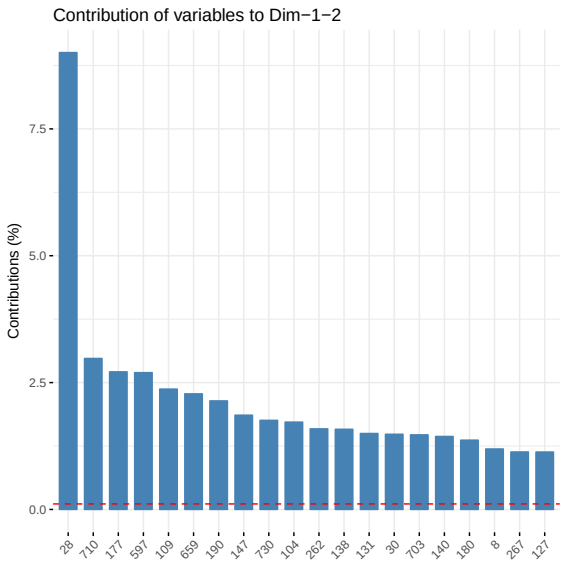
Regarding Figure 5.15 for Dunblane, variables 2, 41 and 33 have the highest positive loadings in PC1 (Figure 5.15, plot (iii), Appendix Table D), thus samples 8 and 9 will tend to have larger values on these variables in PC1 (Figure 5.15, plot (iv)), while variables 1, 4, 3 and 6 are observed to have the highest negative loadings in PC1 (Figure 5.15, plot (iii), Appendix Table D), thus, samples 4 and 5 will tend to have higher values on these variables (Figure 5.15, plot (iv)). On the other hand, variables 5, 14 and 18 have the highest positive loadings in PC2 (Figure 5.15, plot (iii), Appendix Table D), thus samples 1, 2 and 3 will tend to have larger positive values on these variables in PC2 (Figure 5.15, plot (iv)). The samples from Dunblane are different again but closer in character to

the Fort William samples than the Aberdeenshire samples (see Table 5.4; the red colour indicates the common compounds between Aberdeenshire and Dunblane and the green colour indicates the common compounds between Fort William and Dunblane). These highest 20 variables are contributing most to the first two PCs in the data sets I, II and III. We investigate this further in the next chapters.

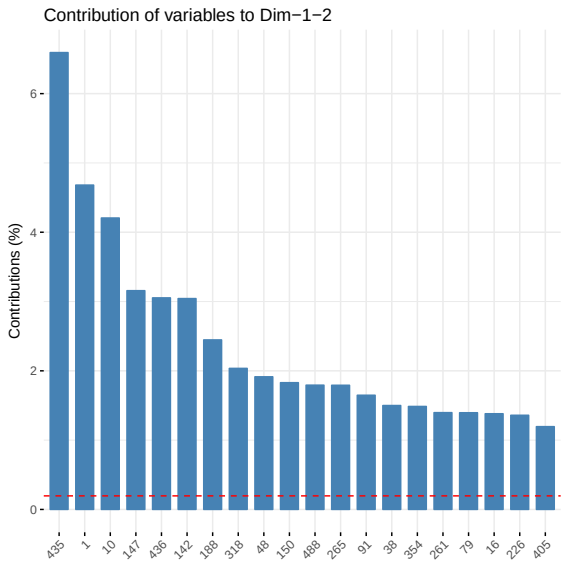
Table 5.4 compares the top 20 variables contributing to PC1 and PC2 for data sets I, II and III (Aberdeenshire, Fort William and Dunblane) which correspond to Figure 5.16 (the total contribution to PC1 and PC2 is obtained with R code "$fviz\_contrib$" from package $factoextra$). In Table 5.4 some compounds such as Methyl pinobanksin (rank number 3 and 17) in the Fort William data set look like duplicates but retention time is different and we have molecules with long alkyl chains that library searches cannot differentiate well between, and if we are drawing the chemical formulae of these compounds they will appear different. It is clear that there are more common compounds between Fort William and Dunblane (shown as green in Table 5.4) and the most differences between Aberdeenshire and Dunblane (shown in red in Table 5.4).

5.5 Applying PCA for the Three Data Sets combined (data set IV)

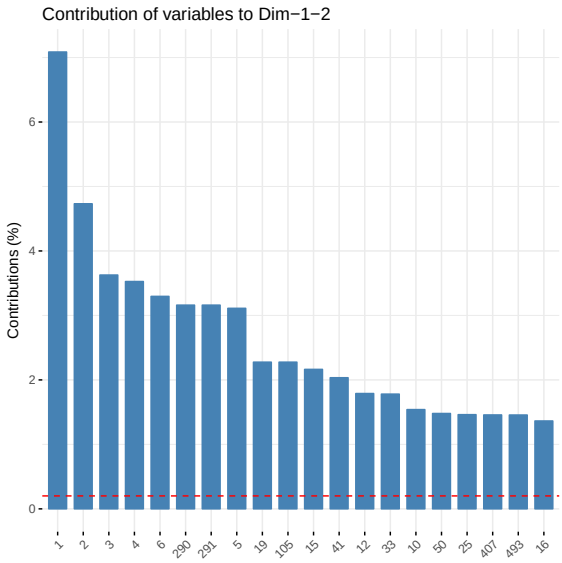
The three data sets (Aberdeenshire, Fort William and Dunblane) are now analysed together. That is, the data set contains the selected 27, 14 (after omitting the 3 identified outliers from data set II) and 9 samples with 921, 511 and 498 variables respectively for the three data sets. Here we want to investigate if the data sets from the three different locations are separated from each other depending on location. We will refer to this data as data set IV. To construct this combined data set, a block diagonal matrix was constructed, where each block contained the data from each location and all other entries were set to 0. This was done because data sets resulted from chemical analysis done at



(i) Aberdeenshire data



(ii) Fort William



(iii) Dunblane

Figure 5.16: Variables contributing to both PC1 and PC2 of data sets I, II and III.

different times, so the variables recorded were not necessarily the same.

As mentioned in Subsection 5.3.2, before doing any PCA analysis it is necessary to test the suitability of the data set IV for PCA. The Gleason-Staelin statistic and the normalised entropy are calculated using equations (5.6) and (5.9) respectively. The value of

Data						
Rank	Aberdeenshire	Retention Time	Fort William	Retention Time	Dunblane	Retention Time
1	*benzoyl hydroxy phenyl acetic acid*	19.7	*coumaric acid*	9.2	benzoyl dihydroxyphenylpropionic acid	18.1
2	*cinnamoyl caffeoyl acetyl glycerol*	18.7	dicaffeoyl acetyl glycerol	14.9	coumaric acid	8.9
3	C23H21O9	17.3	*Methyl pinobanksin*	13.4	caffeic acid hextrieneoate	21.7
4	*dicaffeoyl acetyl glycerol*	15.9	*cinnamoyl caffeoyl acetyl glycerol*	17.7	*prenylated flavonoid*	20.9
5	*coumaroyl feruoyl acetyl glycerol*	19.0	*benzoyl hydroxy phenyl acetic acid*	18.7	*benzoyl hydroxy phenyl acetic acid*	18.6
6	sesquiterpene	15.0	C23H21O9	16.3	benzoyl dihydroxyphenylpropionic acid	17.4
7	dicoumaroyl glycerol	15.7	*benzoyl dihydroxyphenylpropionic acid*	18.3	benzoyl dihydroxyphenylpropionic acid	17.6
8	prenylated flavonoid	22.0	*coumaric acid cinnamyl ether*	22.3	*cinnamoyl caffeoyl acetyl glycerol*	17.6
9	coumaryl acetyl glycerol	12.7	coumaroyl feruoyl acetyl glycerol	16.6	prenylated flavonoid	20.1
10	coumaroyl feruoyl acetyl glycerol	17.6	benzyl coumarate	20.4	prenylated flavonoid	20.3
11	pinocembrin	19.2	Hydroxy phenyl acetyl dihydroxyphenylacetic acid	15.1	phenylacetic acid	9.1
12	C15H13O5	15.4	*C16H11O5*	15.1	Methyl pinobanksin	13.3
13	C24H29O4	25.3	pinocembrin methyl ether	18.3	*dimethyl pinocembrin benzoate*	21.3
14	quercetin hexanoyl ester	19.0	hydroxyphenyl propionic acid	8.9	coumaric acid cinnamyl ether	22.1
15	*Hydroxy phenyl acetyl dihydroxyphenylacetic acid*	16.2	coumaroyl feruoyl acetyl glycerol	18.0	dimethyl flavanol	21.2
16	dimethyl pinocembrin benzoate	22.4	*caffeic acid hextrieneoate*	22.0	C24H29O5	23.9
17	C24H29O4	23.9	*Methyl pinobanksin*	17.9	Methyl pinobanksin	18.5
18	C27H25O6	16.7	*phenylacetic acid*	9.3	C16H11O5	14.9
19	*pinocembrin methyl ether*	19.3	dihydroxylinoleic acid	17.1	Methyl pinobanksin	18.7
20	Caffeic acid pentenyl ester	19.5	Galangin	19.0	prenylated flavonoid	23.7

Table 5.4: Comparison between the 20 components contributing most to the first two PCs in data sets I, II and III; blue indicates common compounds between Aberdeenshire and Fort William, red indicates common compounds between Aberdeenshire and Dunblane, and green indicates common compounds between Fort William and Dunblane.

the Gleason-Staelin statistic using the correlation matrix is 0.39 for data set IV. Moreover, the normalised entropy for data set IV is 0.49. Both statistics confirm beyond any doubt that data set IV is suitable for PCA analysis.

The next step in PCA is to identify the number of principal components to retain for analysis. The principal components of data set IV are shown according to the percentage of the total variation in the data explained, for the first ten principal components, in Figure 5.17. The plot shows that about 66%, 82% and 90% of the total variation is explained by 2, 4 and 7 PCs respectively. Table 5.5 contains the standard deviation, the percentage of the total variance explained and the cumulative percentage of variance explained for the first ten PCs. The detailed results for the variance of the PCs indicate that no more than 3 components need to be retained for further analyses, as they explain most of the variation in the data, about 77.5%. The first two PCs explain about 66%, which is a little low to explain the variation in data set IV. To confirm these findings, as described

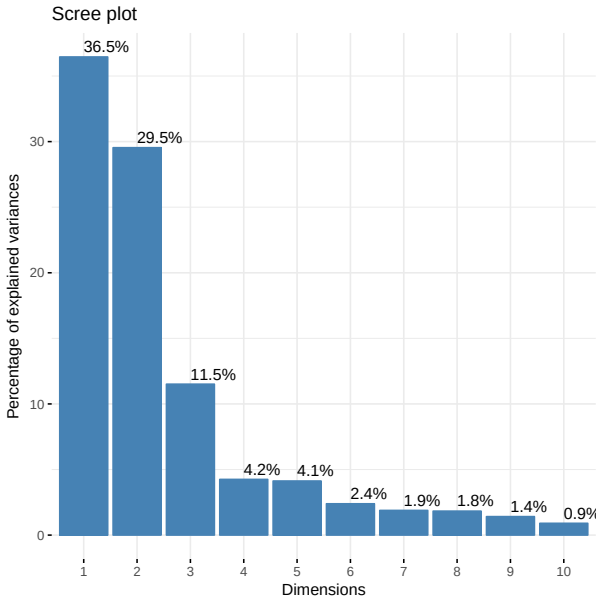


Figure 5.17: Percentages of the total variation in data set IV explained by the first ten PCs.

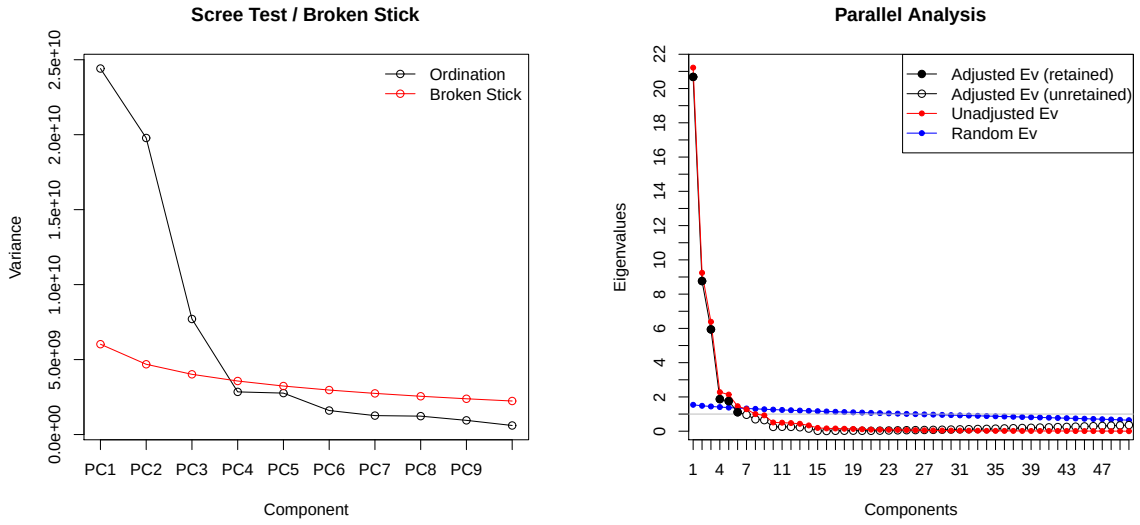
in Section 5.2.4, the broken stick, parallel analysis and scree plot stopping rules will be used to identify the appropriate number of principal components. An illustration of the broken stick model can be seen in Figure 5.18 (left), showing that only three components should be retained, as only three eigenvalues are larger than the expected values of the

PCs	Standard Deviation	% of Variance	Cumulative %
PC1	1.562e+05	36.46	36.46
PC2	1.406e+05	29.54	66.00
PC3	8.779e+04	11.51	77.51
PC4	5.334e+04	4.25	81.76
PC5	5.257e+04	4.13	85.88
PC6	4.002e+04	2.39	88.28
PC7	3.553e+04	1.89	90.16
PC8	3.502e+04	1.83	91.99
PC9	3.076e+04	1.41	93.41
PC10	2.457e+04	0.90	94.31

Table 5.5: Standard deviation, percentage of total variance explained, and cumulative percentage of variance explained for the first ten PCs of data set IV.

broken stick distribution (red line). Cattell′s scree test is also depicted in Figure 5.18 (black line in the left figure), confirming that at most four components should be retained (using one more component after the break in the line (Jackson, 2003)). Another method to confirm how many PCs should be retained is parallel analysis, in Figure 5.18 (right) which suggests that 6 components should be retained. Finally, in Table 5.6, a comparison of the results for a number of stopping rules can be seen. This shows that about 90%, 95% and 99% of the total variation is explained by 7, 11 and 25 PCs respectively. The results stated in Table 5.6 do not show how many PCs should be retained, as there is disagreement between methods. Despite the first three components explaining approximately 77.5% of the total variation, three PCs is most probably the appropriate number of PCs to retain.

From Figure 5.19, the first three PCs can discriminate samples of the data set IV according to the location of each data set. There is overlap of samples from Fort William and Dunblane in the 2D plot (PC2 verses PC1), though PC3 separates them, while in the



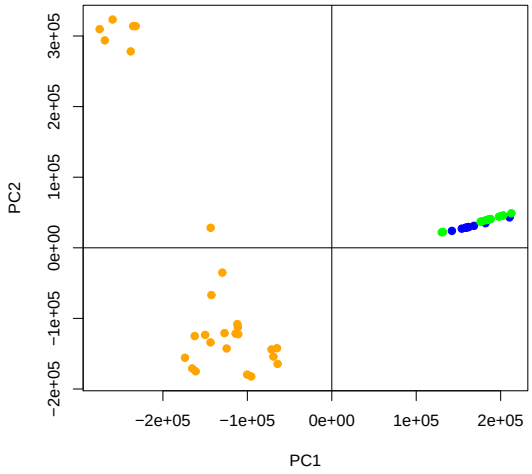
(i) Scree and Broken Stick plots (ii) Parallel Analysis plot

Figure 5.18: Stopping rules for the number of PCs for data set IV.

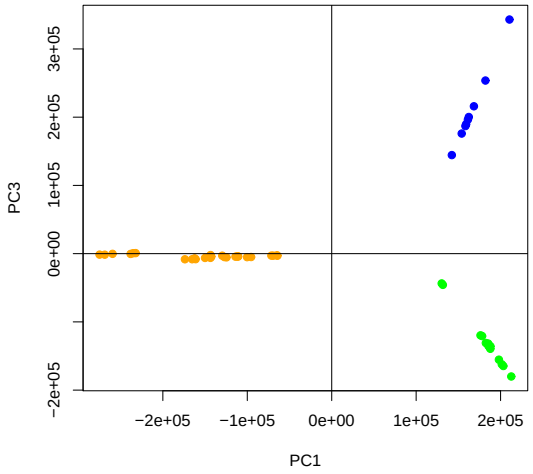
Stopping rule	Number of Components retained for data IV
Parallel Analysis	6
Broken Stick	3
Cattel's Scree Test	4
90% of Variance	7
95% of Variance	11
99% of Variance	25
Information Dimension	7

Table 5.6: Comparison of various stopping rules for the selected data set IV of 50 samples and 921 variables.

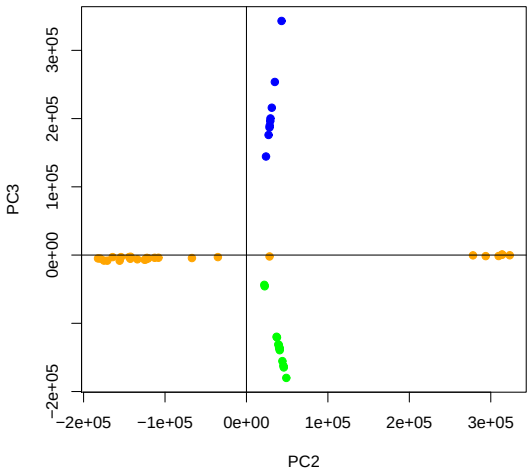
3D plot there are differences between data sets, where the Aberdeenshire data separates completely from Fort William and Dunblane. Moreover, Figure 5.19 indicates that there are many samples having a high positive or negative score on PC1 or PC2, which may be outliers, influencing mainly PC1 or PC2. Diagnostic plots using the score and the orthogonal distance of samples in the data IV can be seen in Figure 5.20. The cut-off values for



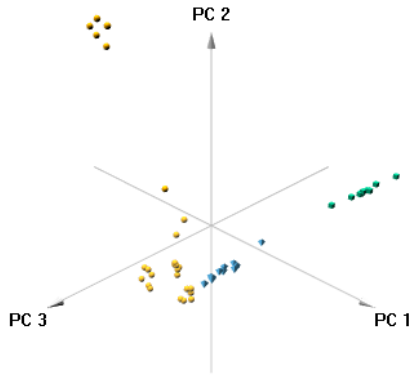
(i) PC1 vs PC2 scores plot.



(ii) PC2 vs PC3 scores plot.



(iii) PC1 vs PC3 scores plot.



(iv) PC1 vs PC2 vs PC3 scores plot.

Figure 5.19: Scores plot for the mean-centred and Pareto-scaled data set IV. The orange colour shows Aberdeenshire, the blue colour shows Fort William, and the green colour shows Dunblane samples.

the score and the orthogonal distance are equal to 24.65423 and 235317.2 respectively. It can be seen that there are samples with score distance higher than the cut-off, namely

samples 4, 5, 6 and, 25, 26 and 27, and ones on the line (sample 49, equal to the cut-off). On the other hand, in the case of the orthogonal distances, it can be seen that there is one point with orthogonal distance equal to the cut-off value (sample 30). The samples 4, 5, 6, 25, 26, 27 and 49 are good leverage outliers, and there are no samples as bad leverage outliers (with high score distance and high orthogonal distance), so there is no need to remove these samples from the data. Therefore, the original data set IV will be used in further investigation.

Figure 5.21 shows the top 20 variables contributing to the principal components for

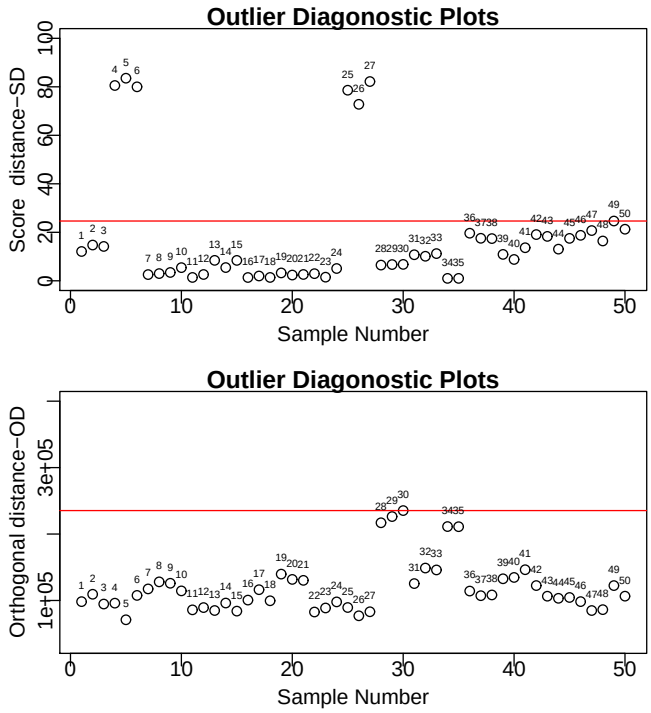


Figure 5.20: Outlier diagnostic plots of data set IV using the score distance (SD) and the orthogonal distance (OD). The numbers in the plots are the sample numbers. The horizontal lines in the two plots represent the cut-off values, such that any point above these lines is a leverage point (top plot) or an orthogonal outlier (bottom plot).

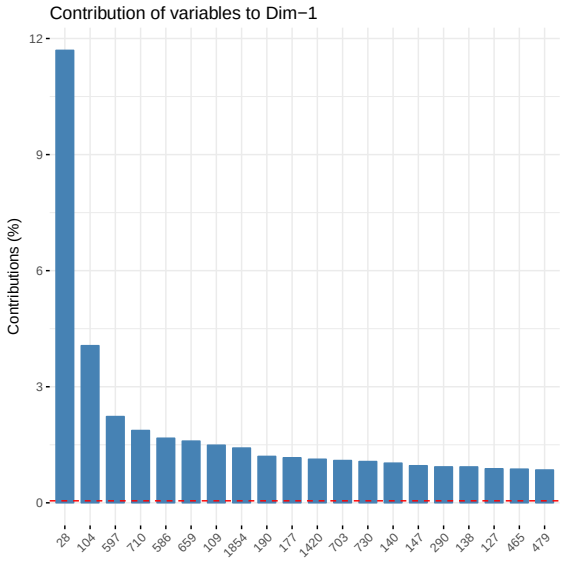
data set IV. The red dashed line on the graph indicates the expected average contribution. If the contribution of the variables were uniform, the expected value would be

$1/length(variables)$ for data set. For a given component, a variable with a contribution larger than this cut-off could be considered as important in contributing to the PCs. Also, the eigenvalues measure the amount of variation retained by each PC. From Figure 5.21 for data set IV, variables 1854 and 1420 have the highest positive loadings in PC1 (Figure 5.21, plot (iii), Appendix Table E), thus samples from Fort William and Dunblane tend to have larger values than samples from Aberdeenshire on these variables in PC1, as well as variables 28, 177 and 109 being observed to have the highest positive loadings in PC2, thus, samples 4, 5, 6, 25, 26 and 27 from the Aberdeenshire data will tend to have larger values on these variables. Samples 1 to 24 from the Aberdeenshire data, except 4, 5 and 6, will tend to have higher values for variables 710, 597, 659 and 104 than other samples. Finally, Table 5.7 shows the top 20 variables contributing to PC1 and PC2 for data set IV, corresponding to the information in Figure 5.22 (the total contribution to PC1 and PC2 is obtained with R code "$fviz\_contrib$" from package $factoextra$).

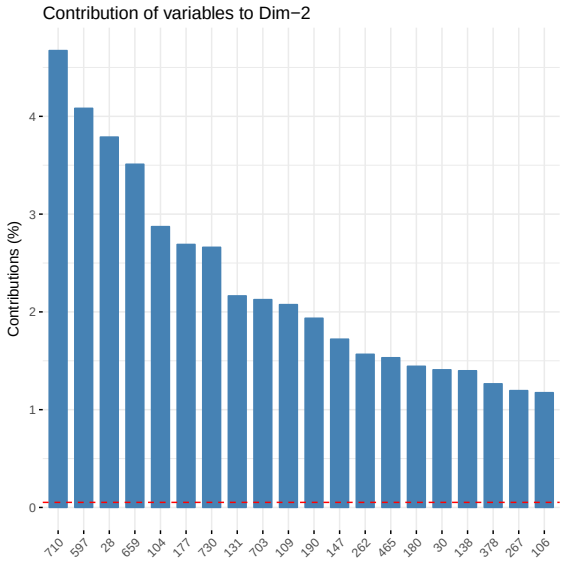
5.6 Applying PCA for Libya Data

Finally, we examine the Libya data using PCA. The Libya data set contains 12 samples with 300 variables. As mentioned in Subsection 5.3.2, before doing any PCA analysis, it is necessary to test the suitability of the Libya data set for PCA. The Gleason-Staelin statistic and the normalised entropy are calculated using equations (5.6) and (5.9) respectively. The value of the Gleason-Staelin statistic using the correlation matrix is 0.59. Moreover, the normalised entropy for the data set is 0.62. Both statistics confirm that the Libya data set is suitable for PCA analysis.

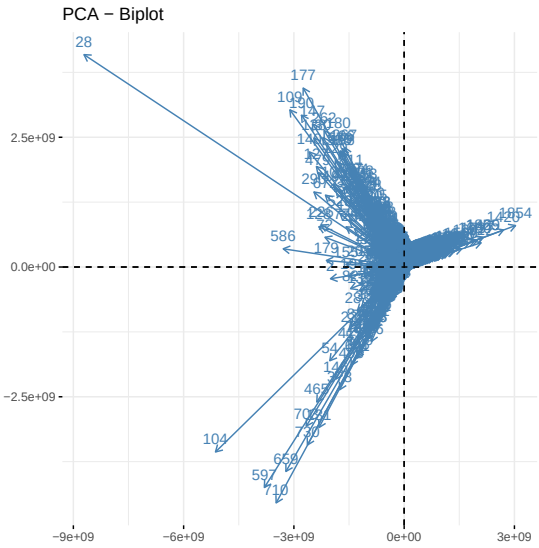
After confirming the suitability of the data, the next step in PCA is to identify the number of principal components to retain for the analyses. The percentage of the total variation in the data explained by the first ten principal components can be seen in Figure 5.23. The plot shows that about 72%, 92% and 95% of the total variation is explained by 2, 5



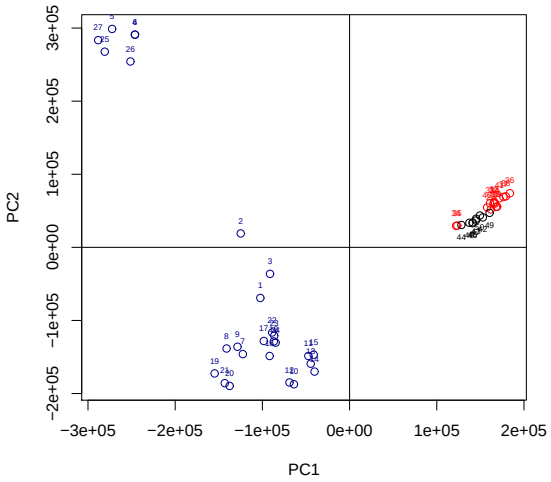
(i) contribution of variables to PC1



(ii) contribution of variables to PC2



(iii) biplot for first two PCs



(iv) Score plot for first two PCs; the blue colour shows Aberdeenshire, red shows Fort William, and black shows Dunblane samples.

Figure 5.21: The top 20 variables contributing to PC1 and PC2 of data set IV, and a biplot and scores plot.

and 6 PCs respectively.

Table 5.8 contains the standard deviation, the percentages of the total variance explained

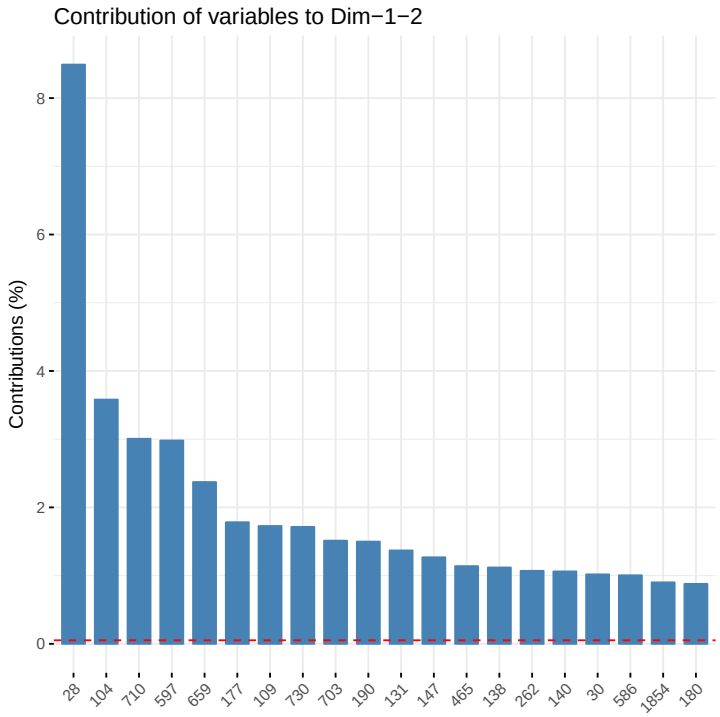


Figure 5.22: The top 20 variables contributing to the first two principal components of data set IV.

and the cumulative percentages of variance for the first ten PCs. The detailed results for the variance of the PCs indicate that no more than 2 components need to be retained for further analyses, as they explain most of the variation in the data, about 72%. To confirm these findings, as described in Section 5.2.4, the broken stick, parallel analysis and scree plot stopping rules will be used to identify the appropriate number of principal components. An illustration of the broken stick model can be seen in Figure 5.24 (left), showing that only one component should be retained, as only one eigenvalue is larger than the expected value of the broken stick distribution (red line). Cattell′s scree test is also depicted in Figure 5.24 (black line in the left figure), confirming that at most two components should be retained (using one more component after the break in the line (Jackson, 2003)). Parallel analysis is shown in Figure 5.24 (right) and suggests that 4 components should be retained. Finally, in Table 5.9, a comparison of the results for a number of stopping rules can be seen. This shows that about 90%, 95% and 99% of the

The Data Set IV		
Rank	name of compound	Retention Time
1	hydroxyphenyl acetic acid	8.0
2	benzyl coumarate	20.4
3	dicaffeoyl acetyl glycerol	14.9
4	cinnamoyl caffeoyl acetyl glycerol	17.7
5	C23H21O9	16.3
6	Methyl pinobanksin	13.4
7	benzoyl hydroxy phenyl acetic acid	18.7
8	coumaroyl feruoyl acetyl glycerol	16.6
9	coumaroyl feruoyl acetyl glycerol	18.0
10	benzoyl dihydroxyphenylpropionic acid	18.3
11	pinocembrin methyl ether	18.3
12	coumaric acid cinnamyl ether	22.3
13	dicoumaroyl glycerol	14.6
14	C16H11O5	15.1
15	Hydroxy phenyl acetyl dihydroxyphenylacetic acid	15.1
16	caffeic acid hextrieneoate	22.0
17	hydroxyphenyl propionic acid	8.9
18	prenylated flavonoid	21.1
19	benzoyl hydroxy phenyl acetic acid	19.72017368
20	Methyl pinobanksin	17.9

Table 5.7: The 20 variables with the highest contribution to PC1 and PC2 for data set IV.

total variation is explained by 4, 6 and 8 PCs respectively. The results stated in Table 5.9 show that 2 or 4 PCs should be retained. Despite the first two components explaining only about 72% of the total variation, two PCs is most probably a sufficient number of PCs to retain.

From Figure 5.25, the most interesting plot of the Libya data set is the score plot for the first two PCs, as these components explain approximately 72% of the total variation. The points are very scattered. Moreover, Figure 5.25 indicates that there are various samples having a high positive or negative score on PC1 or PC2, which seem to be outliers, influencing mainly PC1 or PC2, through point 8 is a bit different from the rest in terms of

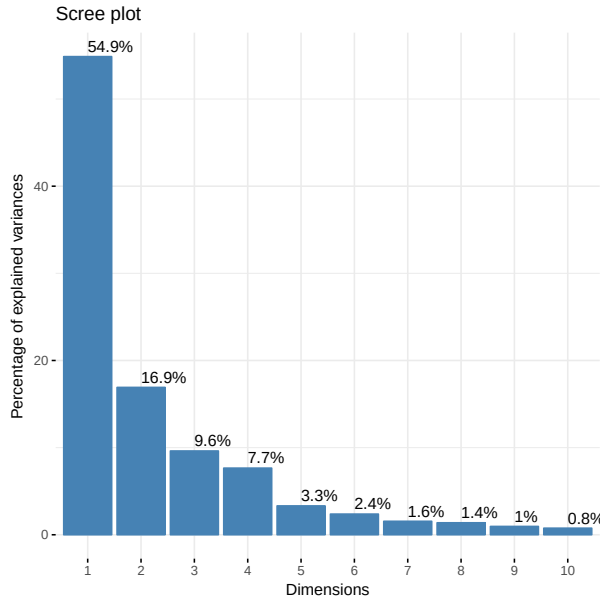
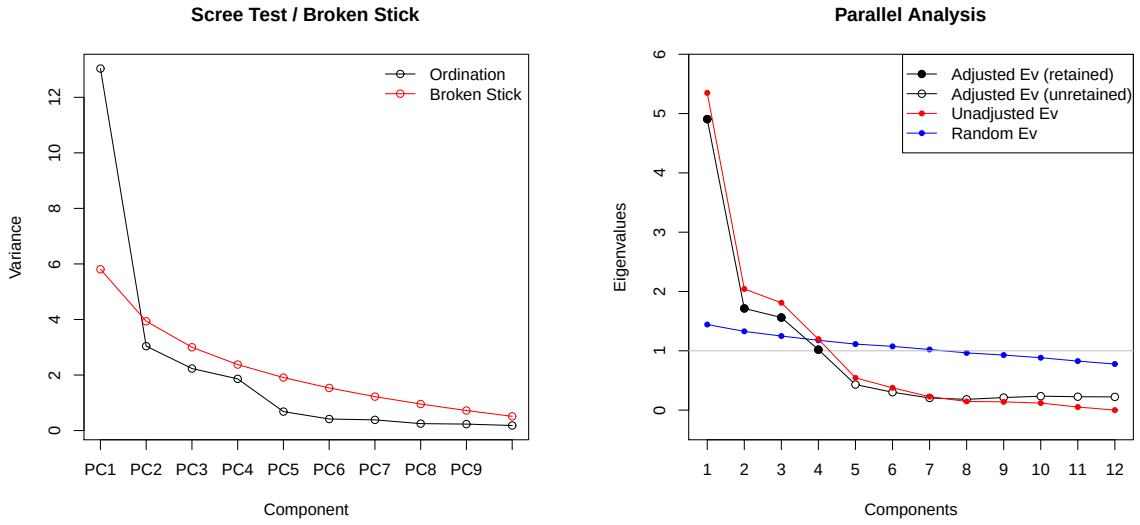


Figure 5.23: Percentages of the total variation in the Libya data explained by the first ten PCs.

PCs	Standard Deviation	% of Variance	Cumulative %
PC1	19.2770	54.87	54.87
PC2	10.7023	16.91	71.79
PC3	8.0711	9.62	81.41
PC4	7.19834	7.65	89.06
PC5	4.73949	3.32	92.38
PC6	4.01229	2.38	94.76
PC7	3.24101	1.55	96.31
PC8	3.05612	1.38	97.69
PC9	2.5503	0.96	98.65
PC10	2.26773	0.76	99.41

Table 5.8: Standard deviation, percentage of total variance explained by, and cumulative percentages of variance for the first ten PCs of the Libya data.



(i) Scree and Broken Stick plots

(ii) Parallel Analysis plot

Figure 5.24: Stopping rules for the number of components of the Libya data.

Stopping rule	Number of Components retained for data IV
Parallel Analysis	4
Brocken Stick	1
Cattel's Scree Test	2
90% of Variance	4
95% of Variance	6
99% of Variance	8
Information Dimension	4

Table 5.9: Comparison of various stopping rules for the Libya data set of 12 samples and 300 variables.

PC3. The samples in this data set seem to be quite varied. Diagnostic plots using the score and the orthogonal distance of samples in the data can be seen in Figure 5.26. The cut-off values for the score and the orthogonal distance are equal to 2.72 and 288859.9 respectively. It can be seen that there are no samples with score distance higher than the cut-off. On the other hand, in the case of the orthogonal distances, it can be seen

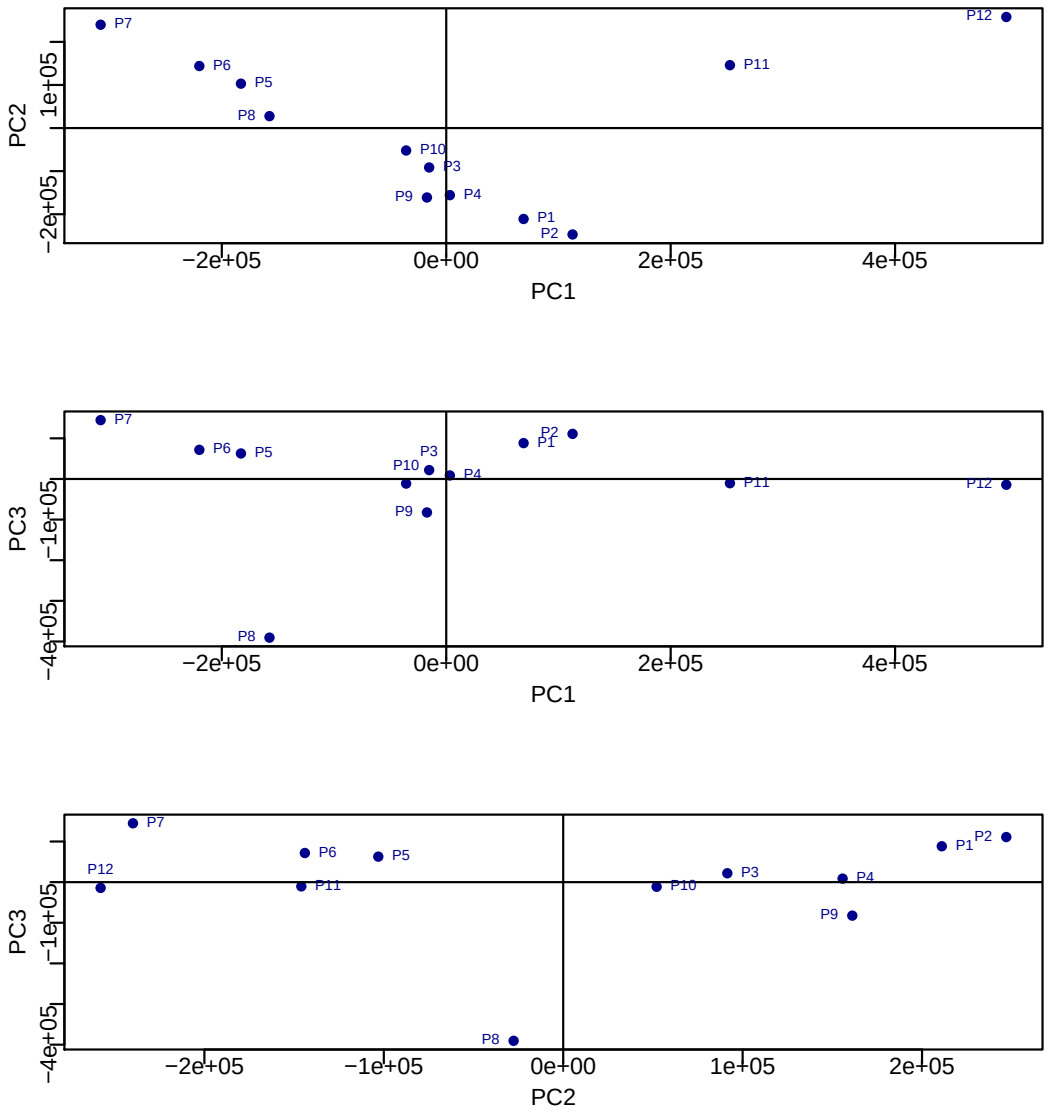


Figure 5.25: Scores plots of the Libya data for the first three PCs, superimposed with numbers representing the different samples, as supplied in the data set.

that there are two points with orthogonal distance higher than the cut-off value (points 11 and 12). These two points are good leverage outliers, and there were no samples as bad leverage outliers (high score distance and high orthogonal distance). Therefore, the original Libya data set will be used in further investigation.

Figure 5.27 shows the top 20 variables contributing to the principal components for the

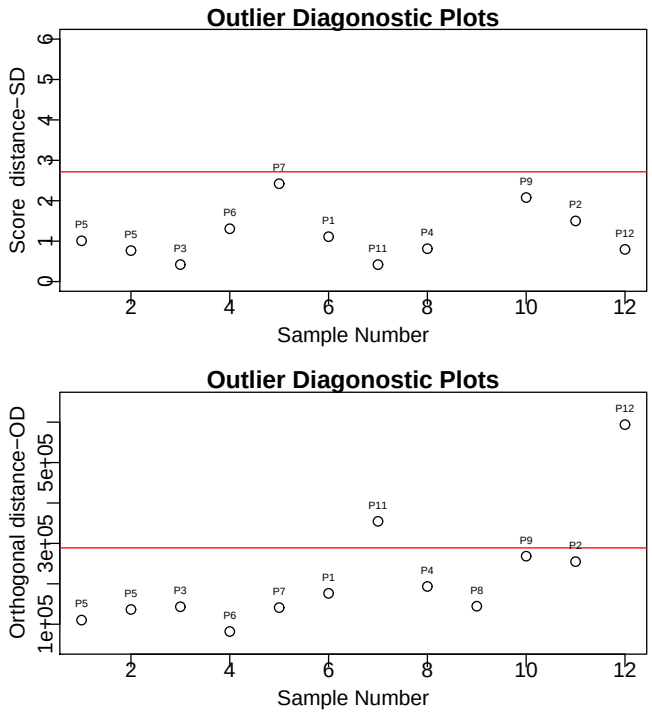
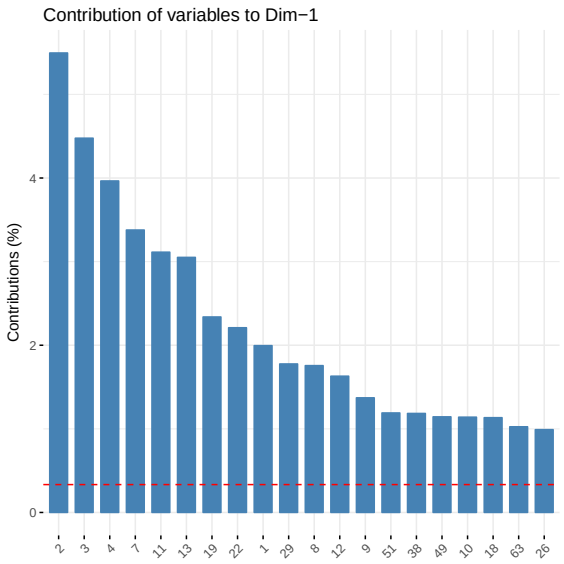
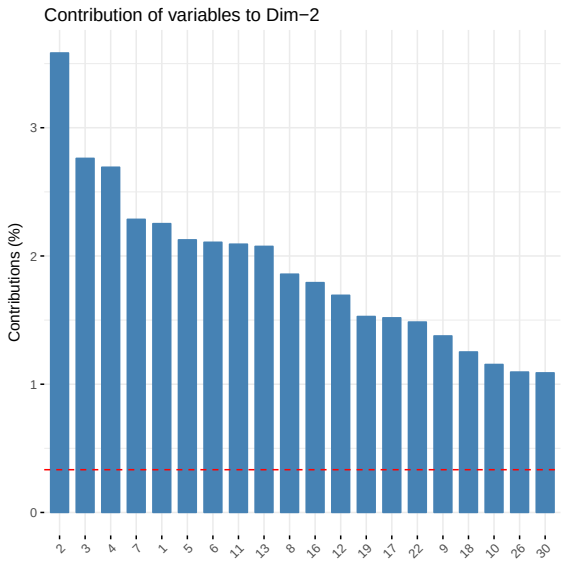


Figure 5.26: Outlier diagnostic plots of the Libya data using the score distance (SD) and the orthogonal distance (OD). The labels in the plots are the numbers of the twelve samples. The horizontal lines in the two plots represent the cut-off values, such that any point above these lines is a leverage point (top plot) or an orthogonal outlier (bottom plot).

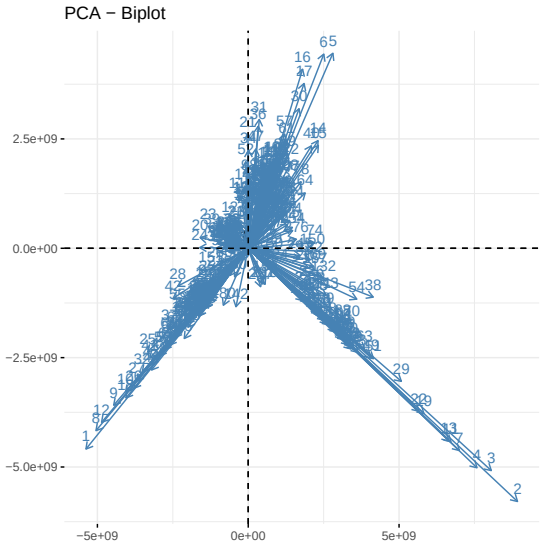
Libya data set. The red dashed line on the graph indicates the expected average contribution. For a given component, a variable with a contribution larger than this cut-off could be considered as important in contributing to the PCs. Also, the eigenvalues measure the amount of variation retained by each PC. From Figure 5.27, variables 2, 3, 4 and 7 have the highest positive loadings in PC1 (Figure 5.27, plot (iii), Appendix Table F), thus samples P1 and P2 tend to have larger values in PC1, as well as variables 5, 6, 16 and 17 being observed to have the highest positive loadings in PC2, thus, samples P11 and P12 will tend to have larger values on these variables. Finally, Figure 5.28 shows the top 20 variables contributing to PC1 and PC2. Table 5.10 compares the top 20 variables con-



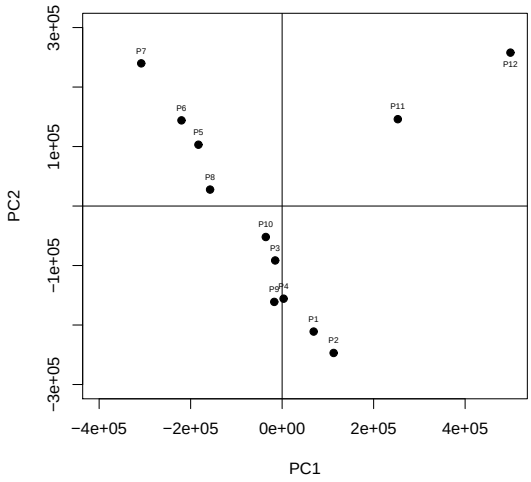
(i) contribution of variables to PC1



(ii) contribution of variables to PC2



(iii) biplot for first two PCs



(iv) Score plot for first two PCs

Figure 5.27: Variables contributing to PC1 and PC2 of the Libya data set.

tributing to PC1 and PC2 for the Libya data set, corresponding to Figure 5.28 (the total contribution to PC1 and PC2, was obtained with R code "$fviz\_contrib$" from package $factoextra$).

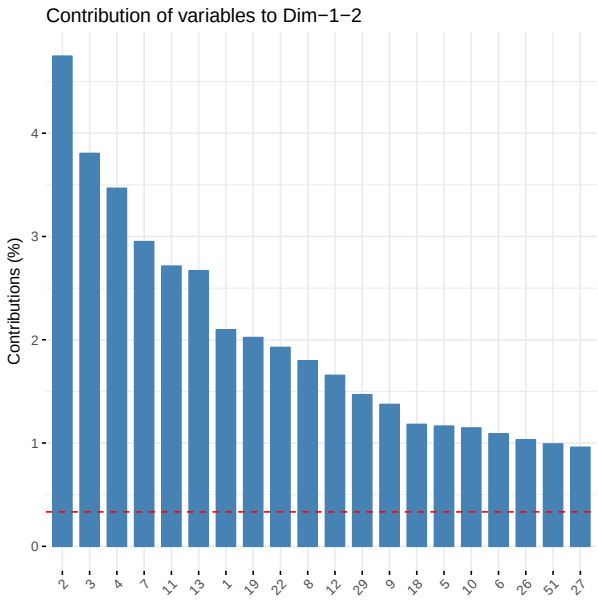


Figure 5.28: The top 20 variables contributing to the first two principal components of the Libya data set.

5.7 Conclusions

In this chapter, PCA was reviewed. Then, the first part of the exploratory analysis covered the application of PCA to data sets I, II, III, IV (Aberdeenshire, Fort William, Dunblane and the data sets combined) and the Libya data set. The data sets I, II, III, IV and Libya were mean-centred and column-scaled by Pareto scaling to make the samples more comparable. Results of the analyses indicated that the first two principal components account for approximately 79%, 84%, 77% and 72% of the total variation in the data sets I, II, III and Libya respectively, and the first three PCs account for approximately 77.5% of the total variation of data set IV (the first two PCs account for 66%, which is low). Therefore the first two or three PCs should be sufficient to separate the samples with respect to every three samples indicating one hive (or colony) in data sets I, II and III and the location of samples in data sets IV and Libya. Various stopping rules, including the broken stick model, scree plot and parallel analysis, were used to confirm that only

Libya Data		
Rank	name of compound	Retention Time
1	Dietrichequinone	14.33
2	Dietrichequinone	15.12
3	Dietrichequinone	14.13
4	C19H24O4	12.86
5	C24H40O3	12.73
6	C24H38O4	12.89
7	C24H38O3	52.59
8	C20H30O3	21.22
9	C21H34O4S	14.43
10	C30H46O4	41.74
11	C24H38O4	42.71
12	C26H44O3	9.74
13	C26H42O3	56.41
14	C20H30O3	36.99
15	C20H22O5	21.43
16	C24H40O3	57.27
17	C20H22O4	24.98
18	C24H36O4	39.97
19	C16H12O6	15.74
20	C24H36O4	37.09

Table 5.10: The 20 compounds contributing most to PC1 and PC2 of the Libya data set.

the first two or three PCs should be retained.

The data sets I, II, III, IV and Libya were examined for the possibility of the existence of potential outliers, and 0, 3, 0, 0 and 0 samples were removed from the data sets I, II, III, IV and Libya respectively, as diagnostic plots showed the omitted points to be bad leverage outliers in data set II (Fort William). The analyses were re-run with the reduced data set II (Fort William), confirming that there was some effect of excluding these samples from the data set on the PCA results. Therefore the new data were used for data set II (Fort William) for further analysis (14 samples instead of 17 samples), while the full original data was used for further analyses of data sets I, III, IV (although data set IV does contain the reduced data set II of 14 samples) and Libya. Loadings plots were used to examine any relationship between samples and variables.

In general, PCA has been quite helpful in obtaining a good idea of the general structure of the data sets I, II, III, IV and Libya and chemical properties of the samples as well as location. It is clear from the analysis that the samples from Libya are much more diverse than these from Scotland. Libya is a very large country and this will reflect different forage sources available to honey bees in different parts of the country. Only samples P5, P6 and P7 from the Southeast of the country gave a distinct group and they were close to the sample P8 from the Southwest. The samples from the coast (see Figure 3.5) did not divide according to longitude and the samples are composed of samples from the East and West of the country, and although P10 was collected from a site close to P11 and P12 it seems to be quite different in composition. The top 20 compounds contributing to the first two PCs are quite different for the Libya data than they are for the Scottish samples.

From the combined data set IV, the Fort William and Dunblane samples are more similar than either of these to the Aberdeenshire samples (Figures 5.19, 5.21 (iv)), but using a third PC dimension (Figure 5.19) it was possible to separate all those locations, therefore there are biochemical differences in the samples from the different locations and also within locations. Table 5.4 did show some compounds in common between the 3 different locations. For Aberdeenshire, the samples do vary. Samples 1, 2, 3, 4, 5, 6, 25, 26 and 27, may relate to more restricted forage sources and the remaining samples to different sources of forage for propolis. For Fort William, samples 4, 5 and 6 were outliers and there was an effect of these samples in the first two PCs. The samples in the middle of the PCA plot are from several different sources of propolis, whereas the samples towards the periphery of the plot focus on more restricted sources, such as samples 1, 2, 3, 7 and 8. For Dunblane, no components can discriminate all samples depending on every three samples from the same hive (or colony). In general, there is variation in composition of samples from any one hive (or colony).

In the next chapter, another unsupervised technique for data exploration and dimension

reduction, multidimensional scaling (MDS), will be reviewed and applied to the propolis data sets, in order to establish if it can be proved more capable of separating the samples or offers any other advantage for dimension reduction of this kind of data.

Chapter 6

Multidimensional Scaling

Having examined PCA in some detail, and applied it to several data sets, in this chapter another data-projection method, with the advantage over PCA that it is flexible and can be used with any dissimilarity measure, is applied to the same data sets as for PCA, to reduce dimensionality, namely, multidimensional scaling (MDS). More specifically, two MDS methods are described in detail and are used, initially the classical MDS, and then, the derived MDS configuration was used as input to the NLM method. The structure of the remainder of this chapter is as follows: In Section 6.2 classical scaling is described and in Section 6.3 metric MDS is reviewed. Sections 6.4, 6.5 and 6.6 are the application of MDS to the Scottish data and the Libya data. Section 6.7 gives conclusions.

6.1 Overview

Various different multivariate statistical methods for multivariate data analysis are encompassed in multidimensional scaling (MDS), such as metric and non-metric MDS methods, unfolding, correspondence analysis and individual differences scaling (Borg et al., 2017; Cox and Cox, 2001). The purpose of MDS scaling is to use a plotted configuration of

points in a dimension that is lower compared to the initial data space, to represent an observed proximity matrix, for easier visualisation or dimensionality reduction (see Figure 6.1[1]). In other words, the dissimilarities among every object pair in a series of n objects serve as input data in MDS. Furthermore, MDS seeks to display such differences as distances among the n points equivalent to the distances among the n objects, in a space of lower dimension (typically two or three dimensions) so that maximum equivalence is established between the obtained distances and the initial dissimilarities (Groenen and Velden, 2004; Izenman, 2008; Williams, 2002).

However, it must be noted that not all MDS methods define the equivalence between

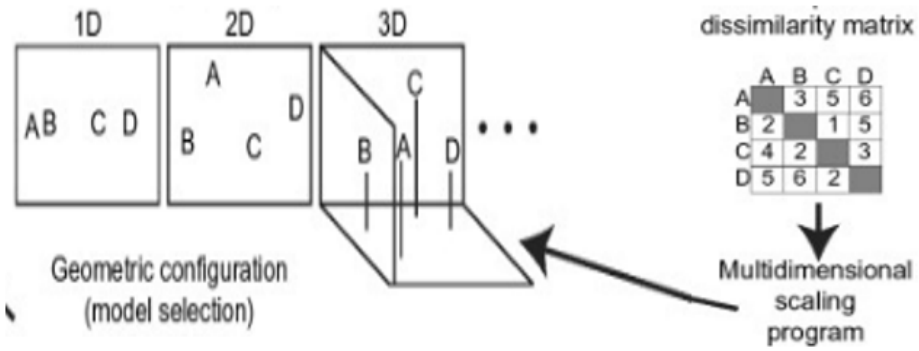


Figure 6.1: Multidimensional scaling analysis.

distances of the points and the object dissimilarities in the same manner. The type of data intended for analysis determines which MDS method should be chosen.

Moreover, the most suitable type of analysis is dictated by how many "modes" and "ways" the input data have. Every series of objects occurring in the data constitutes a "mode" in MDS. For instance, the dissimilarities ρ_{ij} among samples arising from propolis represent one-mode data, while every index in inter-object measurement represents a "way". Hence, due to the presence of two indices, i and j (where y_i and y_j refer to objects), the previously mentioned dissimilarities ρ_{ij} constitute two-way data. As distinguished by Cox and Cox (2001), two-mode, two-way data are typically necessary for correspondence

[1]Source: https://www.sciencedirect.com/topics/medicine-and-dentistry/multidimensional-scaling.

and unfolding analysis, scaling of individual differences facilitates analysis of two-mode, three-way data, while data with a higher number of modes and ways can be managed by other methods.

Additionally, the choice between metric or non-metric MDS depends on the scale of measurement of the dissimilarities. Metric MDS is preferred in cases where the ratio or interval scale is used for measurement of dissimilarities, whilst non-metric MDS is more appropriate in cases of ordinal or nominal data, since its focus is not the actual values but the ranks of the dissimilarities (Cox and Cox, 2001; Izenman, 2008). The focus here is on metric MDS, due to the nature of the metabolomics data.

6.2 Classical Scaling

As algebraic techniques, classical scaling algorithms are employed to map n p-dimensional objects y_i into n points $\acute{y}_i$ in a space of lower dimension to produce maximum equivalence between the initial object dissimilarities (ρ_{ij}) and the distances between points (d_{ij}) (i.e. $d_{ij} \approx \rho_{ij}$) in the new space. Considering y_i with $i = 1, ..., n$ as representing n p-dimensional objects, equation (6.1) defines a dissimilarity (ρ_{ij}) between the object y_i, with coordinates $y_i = (y_{i1}, ..., y_{ip})$, and y_j, with coordinates $y_j = (y_{j1}, ..., y_{jp})$:

$$\rho_{ij} = \left\{ \sum_{k=1}^{p} |y_{ik} - y_{jk}|^S \right\}^{\frac{1}{S}} (S > 0). \tag{6.1}$$

The Euclidean distance, which results from a S value of 2 in equation (6.1), represents the main L_m metric in classical MDS. The Manhattan distance, which results from a S value of 1 in equation (6.1). Other studies (Everitt (1993); Everitt and Rabe-Hesketh (1997); Gordon (1981); Krzanowski and Marriott (1995)) have put forth several other dissimilarity measures such as the Maximum distance or Chebyshev is:

$$\rho_{ij} = \max |y_{ik} - y_{jk}|. \tag{6.2}$$

The Canberra distance is found using equation:

$$\rho_{ij} = \sum_{k=1}^{p} \frac{|y_{ik} - y_{jk}|}{|y_{ik}| + |y_{jk}|}. \tag{6.3}$$

An $(n \times n)$ matrix comprising all pairwise dissimilarities between the n objects is known as a proximity matrix C (i.e. $C = (\rho_{ij})$).

The key stages of the classical MDS algorithm, are as follows (Hothorn and Everitt, 2014; Izenman, 2008; Williams, 2002):

1. The $(n \times n)$ matrix $A = (\alpha_{ij})$ is derived from the $(n \times n)$ proximity matrix $C = (\rho_{ij})$, with

$$\alpha_{ij} = \rho_{ij}^2, \tag{6.4}$$

   e.g. to give squared Euclidean distances.

2. Find the symmetric $(n \times n)$ matrix

$$B = -\frac{1}{2} J_n A J_n,$$

   with

$$J_n = I_n - n^{-1} 1_n 1_n^T,$$

   where I_n represents the $(n \times n)$ identity matrix, n is the number of objects, and 1_n represents the $(n \times 1)$ vector with every element equivalent to 1, and J_n is a centring matrix.

3. Calculation of eigenvalues and eigenvectors of B. Considering the matrix of eigenvalues λ_i of B and the matrix of eigenvectors v_j of B in a column configuration, denoted by $\eta = \text{diag}\ (\lambda_1, ..., \lambda_n)$ and $\mu = (v_1, ..., v_n)$, respectively, the spectral theorem gives the relationship

$$B = \mu \eta \mu^T.$$

4. The obtained lower-dimensional configuration of points $\acute{y}$ represents the t-dimensional configuration of the n input objects y defined by the coordinate matrix

$$\acute{y} = \mu_t \eta_t^{\frac{1}{2}},$$

with μ_t and η_t respectively denoting the matrix of t eigenvectors and the diagonal matrix of the t largest positive eigenvalues of B ($t \leq p$), where t is the specified number of dimensions to be used in the output.

5. Every eigenvalue of matrix B is positive and the t largest eigenvalues define the t-dimensional configuration of optimal fit, if the Euclidean distance is used as the L_m distance metric. If the criterion size given by

$$M_t = \frac{\sum_{i=1}^{t} \lambda_i}{\sum_{i=1}^{n-1} \lambda_i},$$

which is a measure of the percentage of variation accounted for through the use of t dimensions, is large (near to 1) then the fitting is considered suitable.
For cases of non-positive-semi-definite eigenvalues, Cox and Cox (2001) proposed the modified measure

$$M_t = \frac{\sum_{i=1}^{t} \lambda_i}{\sum_{i=1}^{n-1} (positive \;\; eigenvalues)},$$

which should again be large for a good solution. Meanwhile, according to Hothorn and Everitt (2014), Mardia's criterion is

$$\frac{\sum_{i=1}^{t} |\lambda_i|}{\sum_{i=1}^{n-1} |\lambda_i|},$$

and a larger value of this is also desirable.

It is essential that, in order to avoid the loss of key information during the MDS process, the number of dimensions that the obtained configuration should have must be determined in the context of MDS method implementation. Inspection of the eigenvalues of matrix B enables determination of the maximum dimensions necessary. The number of non-zero eigenvalues is a suitable number of dimensions in the case of positive-semi-definite B, as in the context of application of the Euclidean distance metric; under different circumstances, the dimensions depend on how many positive eigenvalues there are. Nevertheless, to ensure practicality and provided that the criteria highlighted above are fulfilled, the first 2 or 3 eigenvalues are typically used, yielding a relatively limited dimensional space for the obtained points and easier visualisation.

6.3 The Metric MDS

In cases where the ratio or interval scale is used for measurement of the data intended for analysis, metric MDS can be employed. If n objects with dissimilarities (ρ_{ij}) are included in the data, it will be necessary to achieve a configuration supporting the relationship

$$d_{ij} \approx f(\rho_{ij}). \tag{6.5}$$

In equation (6.3), the dissimilarities among the points denoting the objects in the point mapping of the initial data space to the space of lower dimension are denoted by d_{ij}, while f represents a continuous parametric monotonic function converting the dissimilarities into distances. There are several choices for f, such as the affine transformation ($d_{ij} = \beta\rho_{ij} + \gamma$), logarithmic transformation ($d_{ij} = \beta \log \rho_{ij} + \gamma$), exponential transformation ($d_{ij} = \beta \exp \rho_{ij} + \gamma$), and power transformation ($d_{ij} = \rho_{ij}^{x}, x > 0$), with the unknown positive coefficients being denoted by β and γ (Hebert et al., 2006; Williams, 2002).

6.3.1 Metric Least - Squares (LS) Scaling

The least squares technique is employed in metric LS scaling for fitting the distances d_{ij} to the transformation $f(\rho_{ij})$ to obtain a configuration of points that can reduce the $STRESS$ function as much as possible (Izenman, 2008), where

$$STRESS = \sum_{i<j} w_{ij}(d_{ij} - f(\rho_{ij}))^2,$$

with w_{ij} representing suitably selected weights. Distances d_{ij} do not necessarily have to be Euclidean. Meanwhile, the dissimilarities are attributed greater weight depending on the selection of weights w_{ij}. For example, greater weight is attributed to small dissimilarities between objects and related points than large dissimilarities if $w_{ij} = \rho_{ij}^{-\frac{1}{2}}$ (Borg et al., 2017; Cox and Cox, 2001). Additionally, the $STRESS$ function serves as a criterion for goodness of fit.

6.3.2 Sammon's Error (STRESS)

A particular version of metric LS scaling, Sammon's non-linear mapping (NLM) is characterised by the weighting system

$$w_{ij} = \frac{1}{\rho_{ij} \sum_{i<j} \rho_{ij}},$$

with f representing the identity function ($f(\rho_{ij}) = \rho_{ij}$) (Cox and Cox, 2001; Izenman, 2008; Sammon, 1969). Under such circumstances, the $STRESS$ function is expressed as (Sammon (1969); Sharaf et al. (1986)):

$$STRESS = \frac{1}{\sum_{i<j} \rho_{ij}^{\phi}} \sum_{i<j} \frac{(\rho_{ij} - d_{ij})^2}{\rho_{ij}^{\phi}}.$$

Since the small ρ_{ij} are retained in this technique, small ρ_{ij} are prioritised over large ρ_{ij} when the distances d_{ij} are fitted, which could be helpful in cases where analysis is geared

towards determining whether the data contain clusters. The prevalence of small or large distances in weighting depends on the power ϕ; for example, small and large distances are attributed identical weights when ϕ has a value of 2, whereas large distances are retained at the expense of small distances when the ϕ value is -2 (Sharaf et al., 1986). Furthermore, an iterative numerical process is employed to solve the series of non-linear least-squares equations making up Sammon's metric $STRESS$ function so as to reduce the $STRESS$ function value as much as possible (Izenman, 2008; Sammon, 1969).

The following steps illustrate how to minimise the $STRESS$ function (Apostolescu and Baran, 2016):

1. Calculate the Euclidean distance from each point X_i to each point X_j, as

$$d_{ij} = \sqrt{\sum_{k=1}^{p}(x_{ik} - x_{jk})^2},$$

   where $i, j = 1, 2, ..., n$ and $k = 1, 2, ..., p$.

2. Use the Singular Value Decomposition (SVD) of the X matrix. SVD decomposes a matrix into a set of rotation and scale matrices, which is used in computing the rank of the matrix X. The general form of SVD decomposition is: $X = USV^T$, where matrix X is $(n \times p)$, matrix U is $(n \times n)$, matrix S is $(n \times p)$ and matrix V is $(p \times p)$. The S matrix is called the singular value matrix and its elements are only nonzero on the diagonal.

3. Estimate the mapping dimension as $s < p$.

4. Calculate a matrix $Y = U^*S$ so that the dimension of Y is $p \times s$.

5. Calculate Euclidean distance from each point Y_i to each point Y_j

$$\rho_{ij} = \sqrt{\sum_{k=1}^{s}(y_{ik} - y_{jk})^2}$$

   where $i = 1, 2, ..., n$ and $j = 1, 2, ..., n$.

6. Calculate the Sammon stress function and minimise it, where

$$STRESS = \frac{1}{\sum_{i<j} \rho_{ij}} \sum_{i<j} \frac{(\rho_{ij} - d_{ij})^2}{\rho_{ij}}.$$

6.4 Application of MDS to Scottish Data Sets I, II and III

The Scottish data sets I, II and III are of type one-mode, two-way, as mentioned previously. In addition, the data consists of continuous variables of quantitative nature measured on the ratio scale due to the nature of the data (metabolite peaks). Therefore, the dissimilarities matrix of the samples contains also quantitative values and metric MDS is the most appropriate to obtain a configuration of points in a lower-dimensional space (Izenman, 2008). An initial configuration will be derived using classical scaling, which will be used as input to the NLM algorithm (Section 6.3.2). The algorithm will attempt to derive a configuration as close as possible to the original, minimising the value of the $STRESS$ function described in the previous section. The data that will be used in the MDS analyses is the same data that were used in Chapter 5, and have been described in detail in Section 3.4.

The first step we will use is the classical scaling solution for data sets I, II and III. Upon exploring the various distance measures that can be used to obtain an initial configuration of points, the criteria for assessing the adequacy of a 2-dimensional solution, indicated that the best distance measure in these cases (Aberdeenshire, Fort William and Dunblane) is the Euclidean distance for both criteria used, as can be seen in Table 6.1. We used the Mardia criterion and also M_2 (Section 6.2). Table 6.1 shows that the Euclidean distance leads to larger (better) criteria values than any of the other methods used. Also, the Canberra distance gives less good results as assessed using the Mardia criterion. The Manhattan distance gives the second best results. The Canberra method is poor when

Data Sets	Aberdeenshire		Fort William		Dunblane	
Metric	M_2	Mardia	M_2	Mardia	M_2	Mardia
Euclidean	0.790	0.790	0.838	0.838	0.769	0.769
Manhattan	0.786	0.751	0.835	0.823	0.755	0.751
Maximum	0.730	0.719	0.809	0.779	0.682	0.666
Canberra	0.610	0.440	0.815	0.481	0.717	0.434

Table 6.1: Values of M_k and Mardia criteria for various Minkowski metrics in classical scaling for data sets I, II and III (k = 2).

the Mardia criterion is used.

Using a 2-dimensional solution is justified by the fact that both criteria for most metrics used indicate that a high proportion of the data variation is explained by using 2 dimensions. Therefore, a 2-dimensional space should be sufficient in these cases. As the Euclidean distance metric is the most commonly used in MDS, and for both criteria its value is 0.790 for Aberdeenshire, 0.838 for Fort William and 0.769 for Dunblane, suggesting that the fit is very good, it seems that it is the most appropriate to use in classical MDS. Results of these metrics can be compared to those from the second best metric, Manhattan, which are similar.

The 2-dimensional configuration derived from the classical scaling using these two distance metrics can be seen in Figure 6.2. It is clear that the plots for the Euclidean and Manhattan distances are both similar to each other and also similar to these of the PCA for the first two PCs in Figures 5.6, 5.10 and 5.11 for data sets I, II and III respectively. That is expected, as in classical MDS using the Euclidean distance results in the same scores derived from PCA, as seen in the left panel of Figure 6.2 (Brereton, 2009). The use of the Manhattan metric results in the configuration seen in the right panel of Figure 6.2, which are similar to these of the Euclidean distance. The configurations from the Euclidean and Manhattan distances squeeze the points towards the bottom side of the plot with respect to coordinate 2 for data set I. However, in all configurations there are

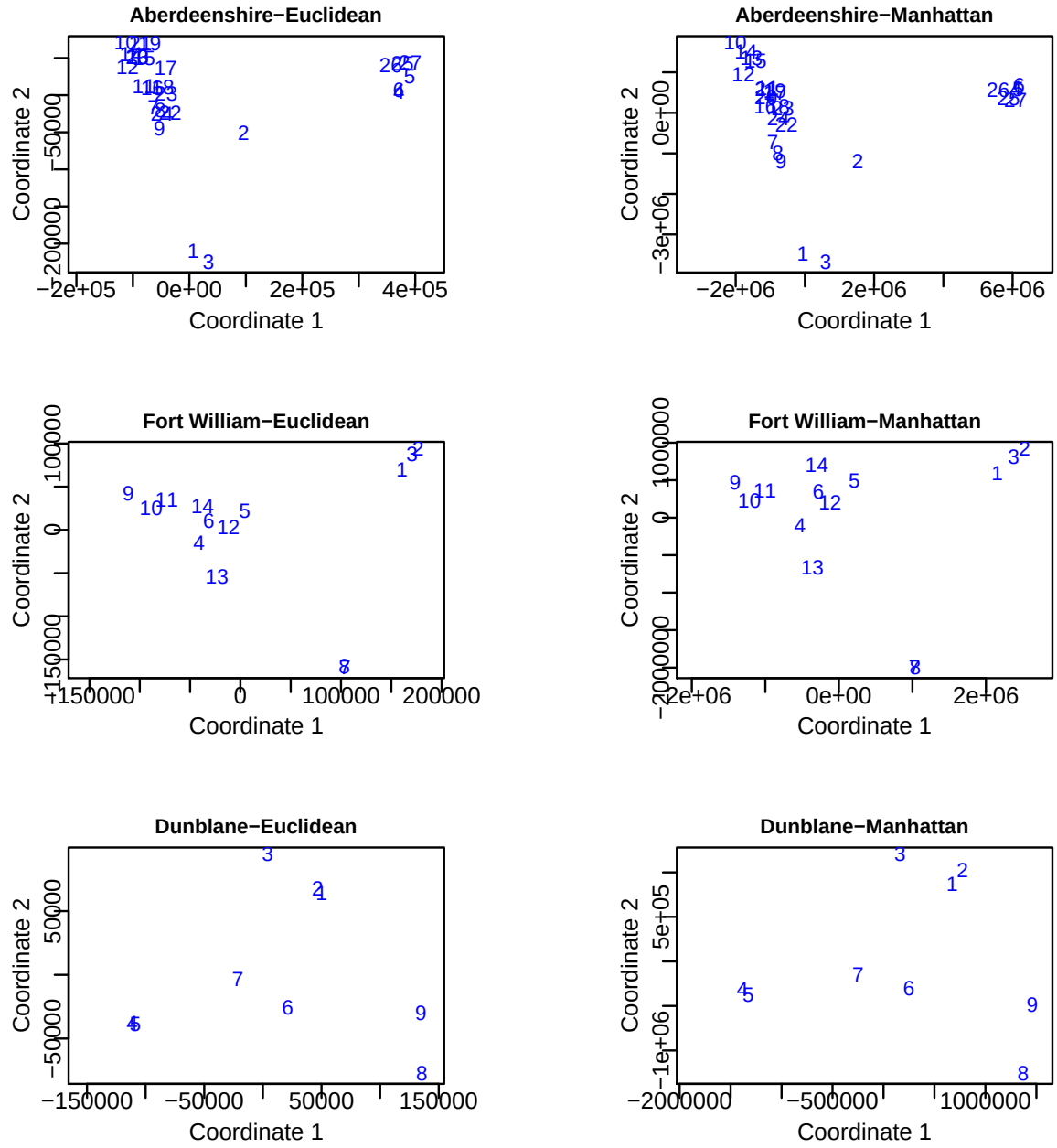


Figure 6.2: Two-dimensional solution of classical MDS using the Euclidean (left plot) and the Manhattan (right plot) distance metrics for data sets I, II and III. The numbers in the plots are the original labels of the samples.

obvious similarities to those of the PCA for the first two PCs of the data sets I, II and III.

A spanning tree is useful in MDS analysis, as it can provide a graphical way of highlighting any possible distortion in the MDS solution. This type of tree is defined as a tree spanning N_s multi-dimensional points (samples). This is any set of straight line segments joining pairs of points such that:

- No closed loops occur,
- Every point is visited at least once,
- The tree has paths between any pairs of points.

The sum of the lengths of the tree's segments is defined as the length of the tree. The minimum spanning tree (MST) is defined as the spanning tree with the minimum length (Hothorn and Everitt, 2014). The links of the minimum spanning tree can be superimposed on the 2-dimensional MDS configuration. Any distortions in the MDS solution are then identified when any nearby points on the scores plot are not connected by a direct line segment of the MST in the above MDS solution.

Figure 6.3 illustrates the minimum spanning tree for the derived MDS configurations above. From the minimum spanning trees, it is clear that there are distortions in both models. For example, samples 1 and 3 in the Euclidean model for data set I, as well in the Manhattan model, among others, appear to be quite close in the scores plot but they are not linked directly in the minimum spanning tree, and similarity in data set II. For example, samples 11 and 14 in the Euclidean model for data set II, as well as 6 and 14 in the Manhattan model, among others, appear to be quite close in the scores plot but they are not linked directly in the minimum spanning tree. In the case of data set III, samples 6 and 7 in the Euclidean model for data set III, as well as samples 8 and 9 in the Manhattan model, among others, appear to be quite close in the scores plot but they are not linked directly in the minimum spanning tree.

In general, classical MDS has confirmed the results of PCA. The Euclidean MDS model provides a two-dimensional configuration which is similar to that of the Manhattan MDS

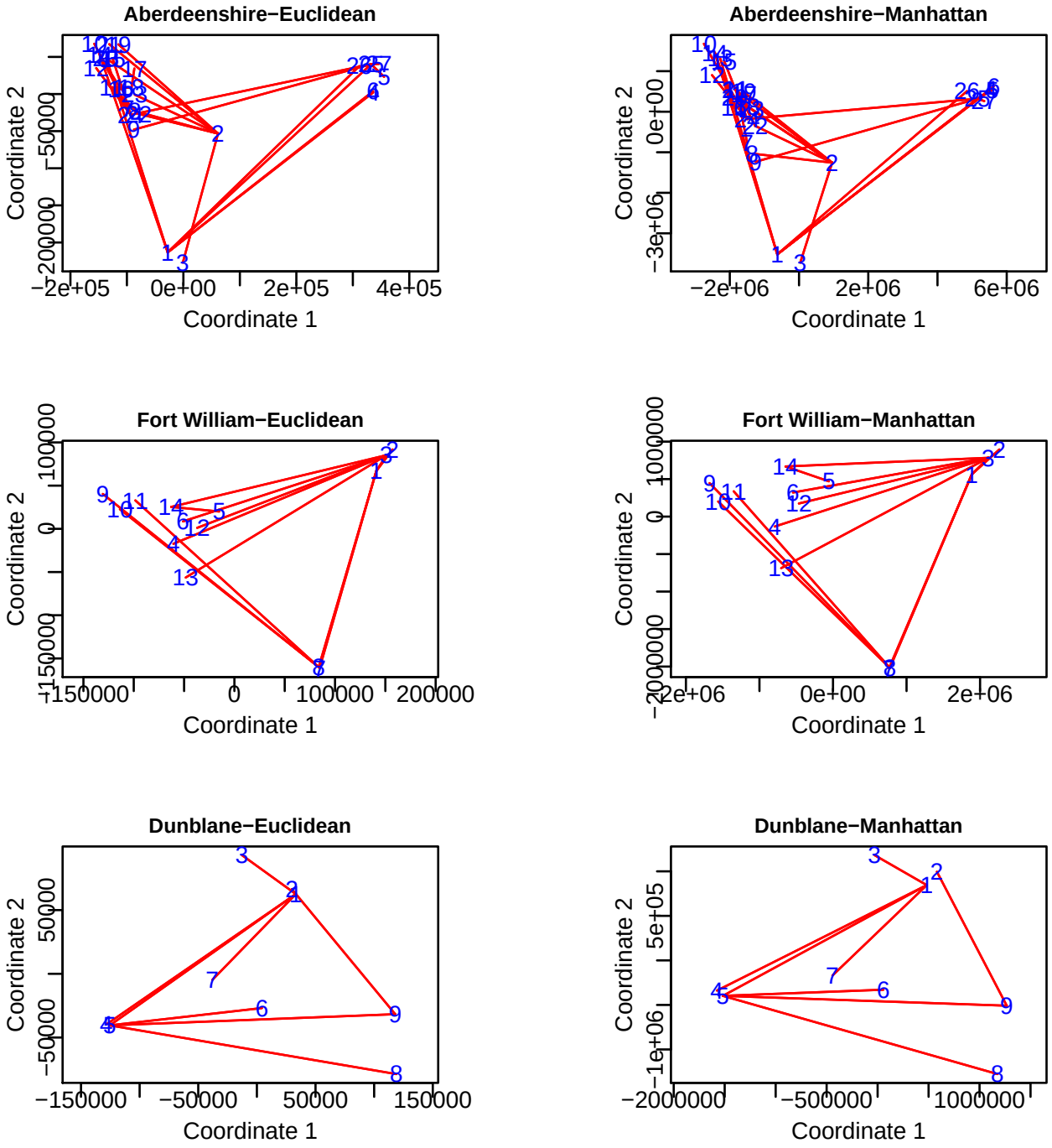


Figure 6.3: Minimum spanning tree for the two best MDS configurations for data sets I, II and III. The numbers in the plots are the numbers of the samples.

model. However, so far MDS has not provided any additional information for the grouping of the samples to that obtained by PCA.

In the next step, an alternative method of implementing non-linear MDS, Sammon's non-linear mapping (NLM) will be applied to data sets I, II and III, to investigate whether NLM can improve the results obtained from the classical MDS analysis.
The initial configurations derived by classical scaling using the Euclidean and Manhattan distance metrics, will be used as input to the NLM algorithm. The optimal NLM models are derived when the minimum values of the $STRESS$ function are 0.01367, 0.01328 and 0.02486 after 50, 30 and 40 iterations and 0.02491, 0.01241 and 0.02035 after 60, 40 and 30 iterations, for the Euclidean and the Manhattan NLM models respectively, for data sets I, II and III respectively. These results are again similar for the Euclidean and Manhattan distances. Figure 6.4 illustrates the final configurations for the two derived optimal NLM models. Comparing the NLM configurations to those obtained from classical MDS (Figure 6.2), it can clearly be seen that in the cases of the Euclidean model and Manhattan model, there are no great difference in the distances between the samples in the two MDS models and in the actual topology of the two Euclidean and Manhattan configurations, as most of the samples are located at approximately the same place in both models of data sets I, II and III. On the other hand, in the case of the Manhattan models of data set I, the compression-like, tight clustering effect that occurs in the classical MDS model has been eliminated in the NLM model, and therefore there is a considerable difference in the between-samples distances of the formerly compressed samples. The NLM configurations are similar in their topology as in the classical MDS models, except that the NLM for Manhattan model for data set I is still clustered but the clustering is looser. As in the case of MDS, the two NLM models do not provide any additional information for grouping of samples to that obtained by PCA.

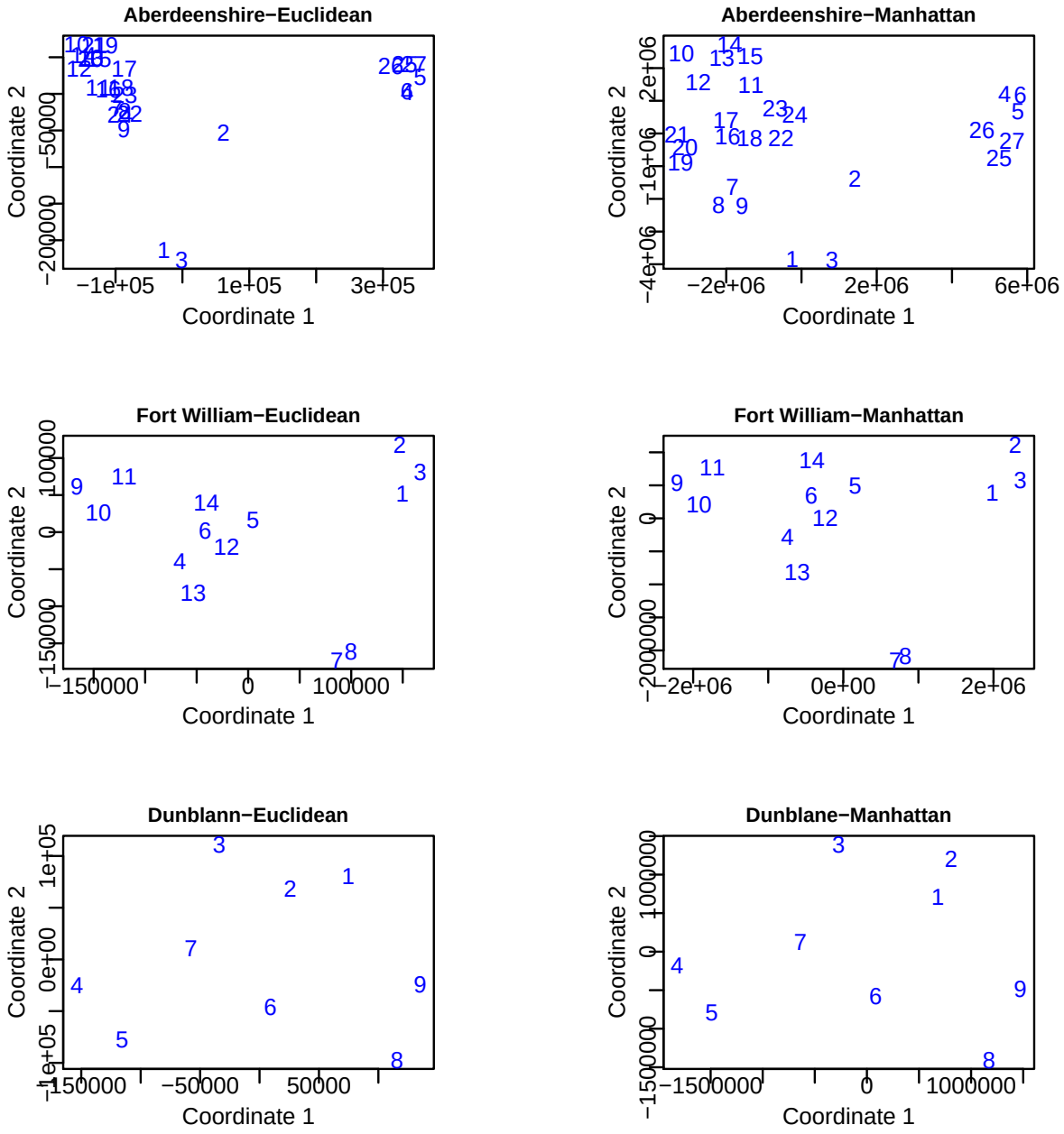


Figure 6.4: Two-dimensional solution of NLM MDS using the Euclidean (left plot) and the Maximum (right plot) distance metrics for data sets I, II and III. As initial configurations, the classical MDS models depicted graphically in Figure 6.2 have been used. The numbers in the plots are the sample numbers.

6.5 Application of MDS to the three Scottish Data Sets combined (IV)

The three data sets (Aberdeenshire, Fort William and Dunblane) are now considered together as one data set, using MDS, as was done in the PCA. That is, data set IV contains the selected 27, 14 and 9 samples with 921, 511 and 498 variables respectively. The three data sets have been mean-centred and column-scaled by Pareto scaling before analysis. Therefore, the dissimilarities matrix of the samples again contains quantitative values and metric MDS is the most appropriate type of MDS to obtain a configuration of points in a lower-dimensional space.

Upon exploring the various distance measures that can be used to obtain an initial configuration of points, the criteria for clustering the adequacy of a 2-dimensional solution indicated that the best distance measure in this case is again the Euclidean distance, having a value for both criteria of 0.660, as can be seen in Table 6.2. The Manhattan distance again is the next best. The Canberra distance is poor again with the Mardia criterion. Using a 2-dimensional solution is justified as both criteria for all metrics indicate that a

Metric	M_2	Mardia
Euclidean	0.660	0.660
Manhattan	0.642	0.640
Maximum	0.637	0.633
Canberra	0.621	0.320

Table 6.2: M_k and Mardia criteria for various Minkowski metrics in classical scaling for data sets IV (k = 2).

high proportion of the data variation is explained by using 2 dimensions. Therefore, a 2-dimensional space should be sufficient in this case.

As the Euclidean distance metric is the most commonly used in MDS and for both criteria

its value is the same and the highest, being approximately 0.660, and suggesting that the fit is good, it seems that it is the most appropriate to use in classical MDS. Results of this metric will again be compared to those from the second best metric, Manhattan. The 2-dimensional configuration derived from the classical scaling using these two distance metrics can be seen in Figure 6.5. It is clear that the plot for the Euclidean and Manhattan distance is similar to that of the PCA for the first two PCs in Figure 5.21 (iv) for data set IV, with the only difference being that the first coordinate in the MDS plot is reflected, as seen in the top panel of Figure 6.5. The use of the Manhattan metric results in the configuration seen in the bottom panel of Figure 6.5. The two patterns of clustering are similar using the Euclidean or Manhattan distance. However, in both configurations there are no obvious differences in groupings of the samples compared to the PCA result. Figure 6.6 illustrates the minimum spanning tree for the derived MDS configurations above. From the minimum spanning trees, it is clear that there are distortions in both models. For example, samples 2 and 3 in the Euclidean model as well as in the Manhattan model, among others, appear to be quite close in the scores plot but they are not linked directly in the minimum spanning tree.

In general, classical MDS has confirmed the results of PCA. The Euclidean MDS model provides a two-dimensional configuration which is similar to the Manhattan MDS model. However, so far MDS has not provided any additional information for the grouping of the samples to that obtained by PCA.

In the next step, an alternative method of implementing non-linear MDS, Sammon's non-linear mapping will again be applied to data set IV, to investigate if NLM can improve the results obtained from the classical MDS analysis. The initial configurations derived by classical scaling using the Euclidean and Manhattan distance metrics will be used as input to the NLM algorithm. The optimal NLM models are derived when the minimum values of the $STRESS$ function are 0.06694 after 30 iterations and 0.03023 after 50 iterations, for the Euclidean and the Manhattan distances respectively. Figure 6.7 illustrates the final configurations for the two derived optimal NLM models. Comparing the NLM

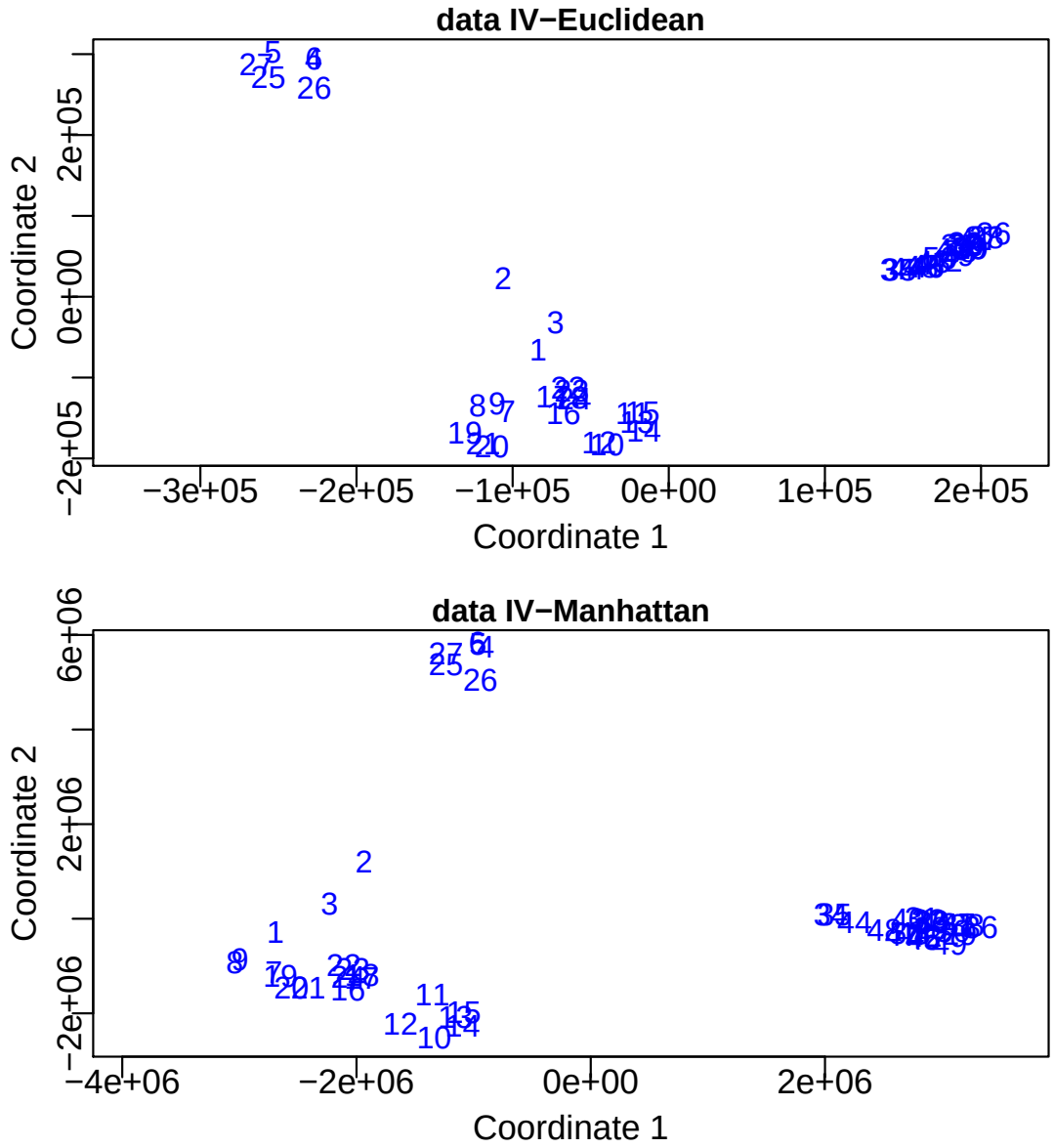


Figure 6.5: Two-dimensional solution of classical MDS using the Euclidean (upper plot) and the Manhattan (bottom plot) distance metrics for data set IV. The numbers in the plots are labels of the samples.

configurations to those obtained from classical MDS (Figure 6.5), it can clearly be seen that in the cases of the Euclidean and Manhattan models, there are great differences in the distances between the samples in the two MDS models and in the actual topology of the two Euclidean and Manhattan based configurations. The clusters are much the

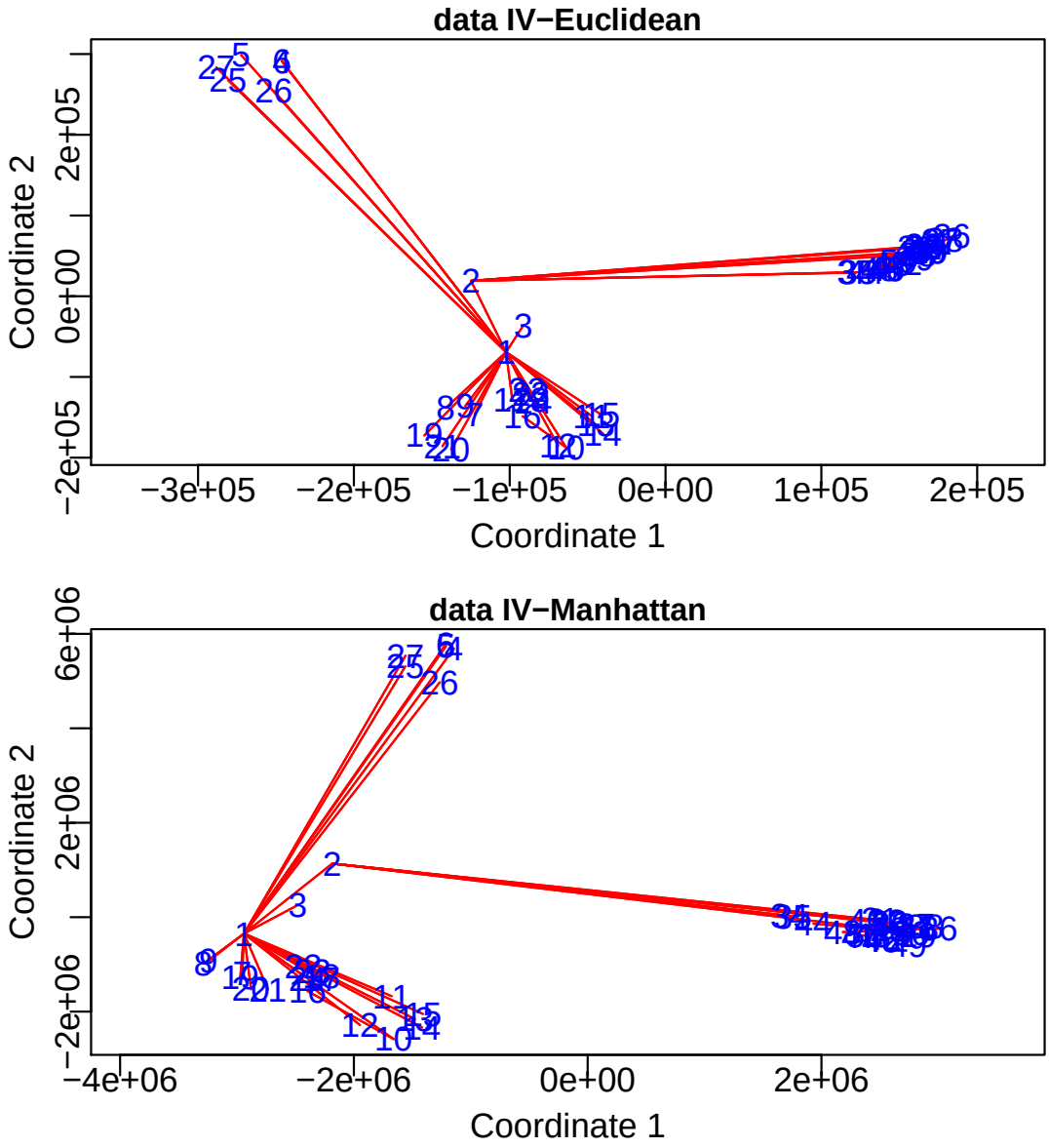


Figure 6.6: Minimum spanning tree for the two MDS configurations for data set IV. The numbers in the plots are labels of the samples.

same but the points are much more widely spaced in the NLM model using Euclidean and Manhattan distances. In the cases of the Euclidean and Manhattan models, the compression-like effect that occurs in the classical MDS model has been eliminated in the NLM model, and therefore there is a considerable difference in the between-samples distances of the formerly compressed samples. The NLM configurations are quite different

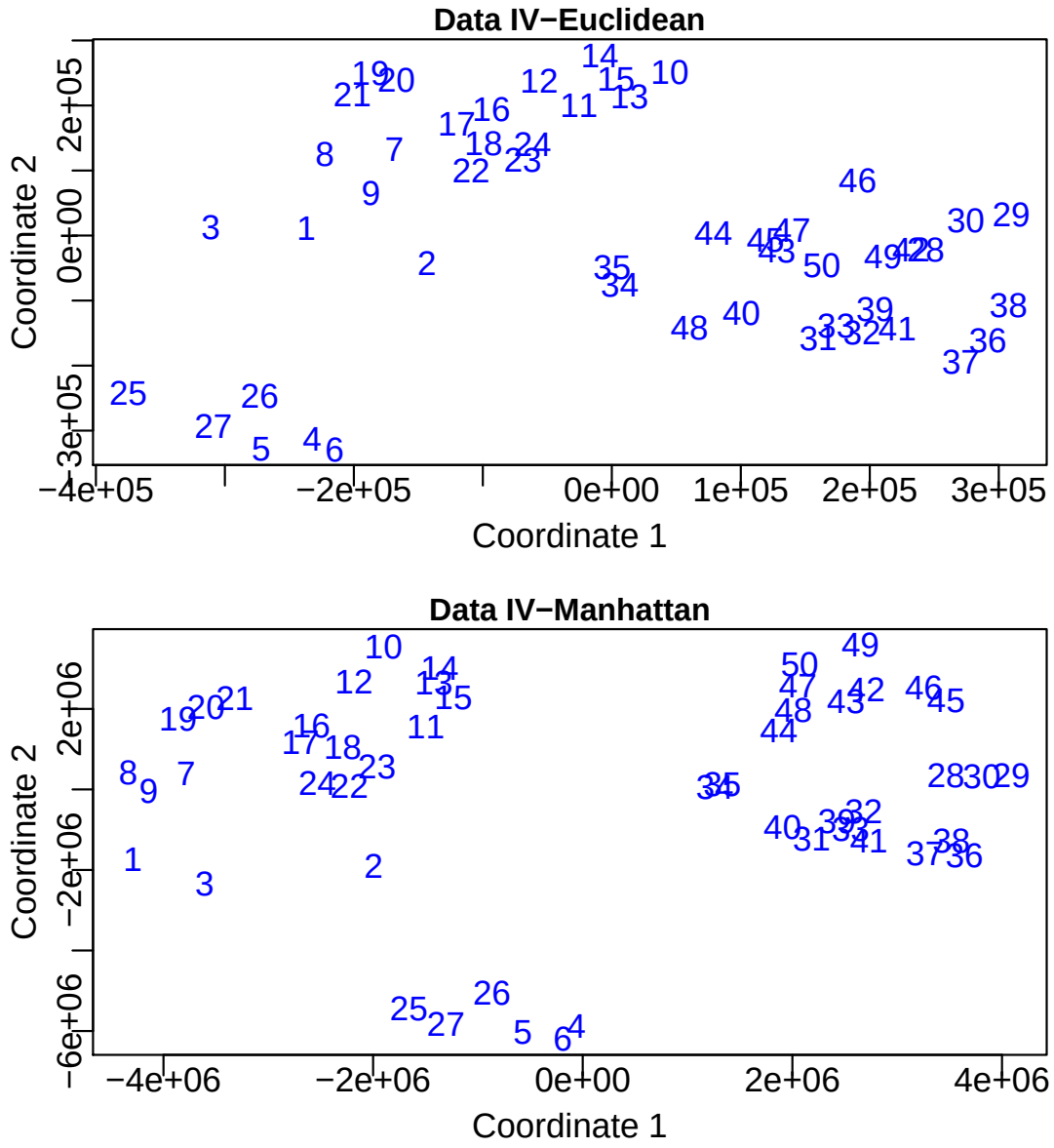


Figure 6.7: Two-dimensional solution of NLM MDS using the Euclidean (upper plot) and the Manhattan (bottom plot) distance metrics for data set IV. As initial configurations, the classical MDS models depicted graphically in Figure 6.5 have been used. The numbers in the plots are the sample numbers.

in their topology than in the classical MDS models. In general, there is no rotation or reflection of the samples in the two NLM models, compared to the classical MDS configurations. As in the case of MDS, the two NLM models do not provide any additional

information for grouping of samples to that obtained by PCA.

6.6 Application of MDS to the Libya Data

A similar analysis was carried out on the Libya data. The best distance measure in this case is the Euclidean distance, having a value of both criteria 0.718, as can be seen in Table 6.3, for a 2-dimensional solution. The results are a bit different from those for the Scottish data but the Euclidean distance still gives good results. Using a 2-dimensional

Metric	M_2	Mardia
Euclidean	0.718	0.718
Manhattan	0.505	0.490
Maximum	0.709	0.706
Canberra	0.698	0.545

Table 6.3: M_k and Mardia criteria for various Minkowski metrics in classical scaling for Libya data (k = 2).

solution is justified by the fact that both criteria for all metrics indicate that a fairly high proportion of the data variation is explained by using 2 dimensions. Therefore, a 2-dimensional space should be sufficient in this case. For both criteria the value is the same using Euclidean distance, being approximately 0.718, and suggesting that the fit is good, so it seems that it is the most appropriate to use in classical MDS.

Results of this metric will be compared to those from the second best metric, which is Maximum in this case. The 2-dimensional configuration derived from the classical scaling using these two distance metrics can be seen in Figure 6.8. It is clear that the plot for the Euclidean distance is similar to that of the PCA for the first two PCs in Figure 5.25 for the Libya data, as seen in the top panel of Figure 6.8. The two panels are rather different. The use of the Maximum metric results in the configuration seen in the bottom panel of

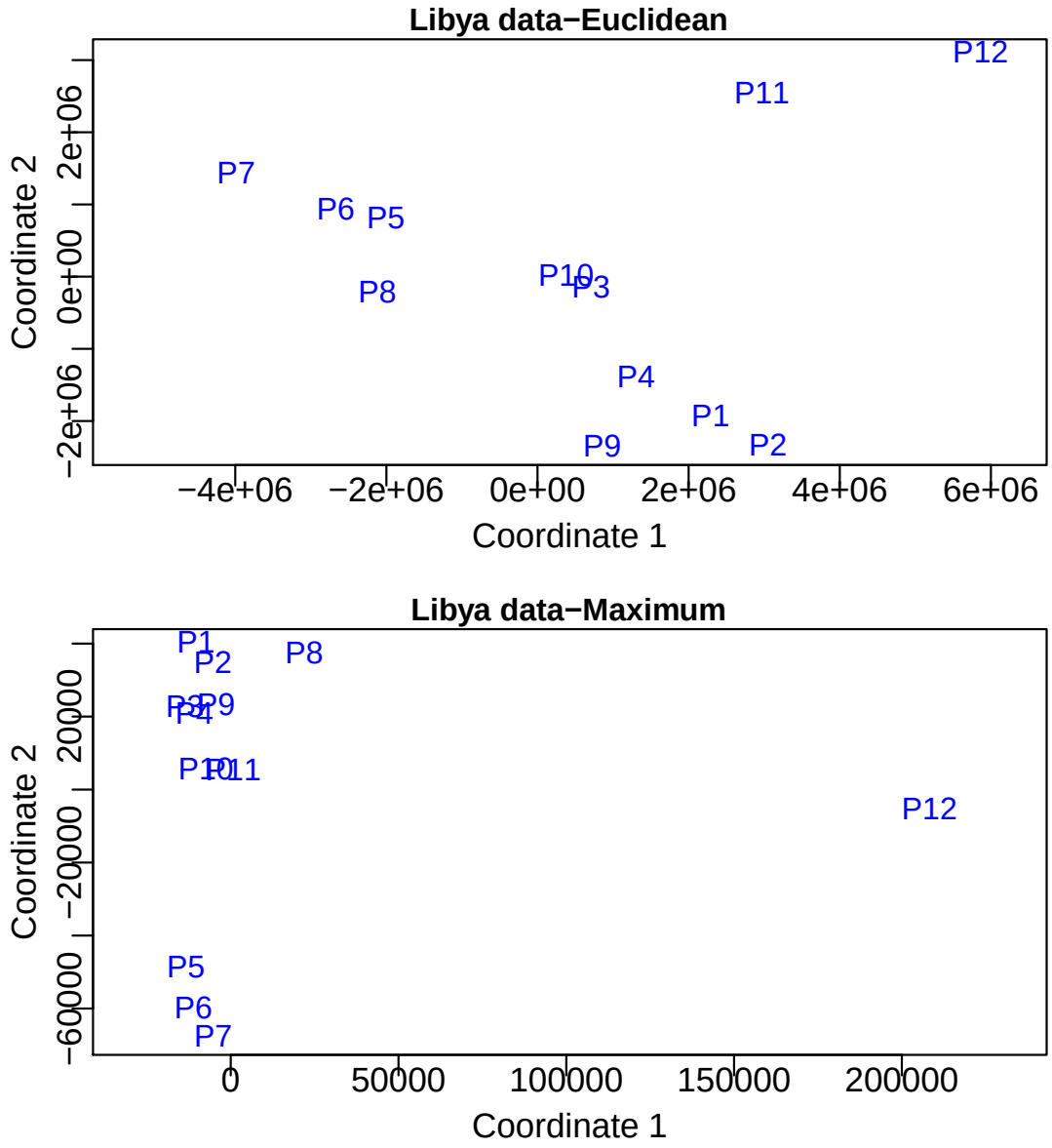


Figure 6.8: Two-dimensional solution of classical MDS using the Euclidean (upper plot) and the Maximum (bottom plot) distance metrics for Libya data. The numbers in the plots are labels of the samples.

Figure 6.8. In the case of the Maximum distance there is some clustering of the samples, as P5, P6 and P7 are similar and P12 is different from the rest. The Maximum distance squeezes the points towards the left side of the plot and towards zero with respect to coordinate 1 for those samples compared to the Euclidean result in Figure 6.8.

Figure 6.9 illustrates the minimum spanning tree for the derived MDS configurations above. From the minimum spanning trees, it is clear that there are distortions in both

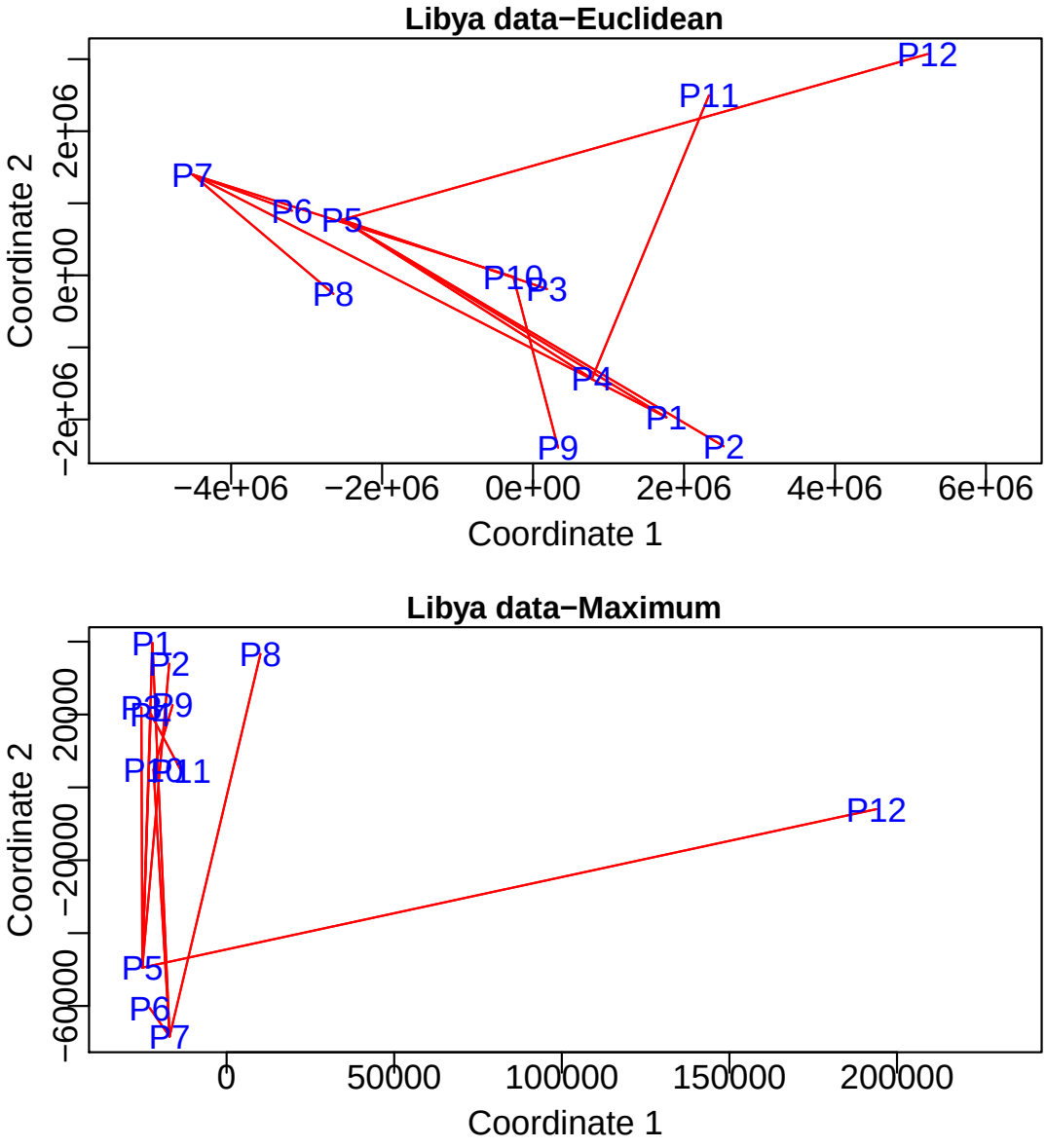


Figure 6.9: Minimum spanning tree for the two MDS configurations for Libya data set. The numbers in the plots are the numbers of the samples.

models. For example, samples P11 and P12 in the Euclidean model for Libya data, as well as P5 and P6 in the Maximum model, among others, appear to be quite close in the scores plot but they are not linked directly in the minimum spanning tree.

In general, classical MDS has confirmed the results of PCA. The Euclidean MDS model provides a two-dimensional configuration which is easier to read than the Maximum MDS model. However, so far MDS has not provided any additional information for the grouping of the samples to that obtained by PCA.

In the next step, Sammon's non-linear mapping will be applied to the Libya data set, to investigate if NLM can improve the results from the classical MDS analysis. The initial configurations derived by classical scaling using the Euclidean and Maximum distance metrics will be used as input to the NLM algorithm. The optimal NLM models are derived when the minimum values of the $STRESS$ function is 0.01524 after 50 iterations and 0.05513 after 110 iterations, for the Euclidean and the Maximum distance respectively. The first is much better. Figure 6.10 illustrates the final configurations for the two derived optimal NLM models. Comparing the NLM configurations to those obtained from classical MDS, it can clearly be seen that in the case of the Euclidean model, there is little difference in the distances between the samples in the two MDS models and in the actual topology of the two Euclidean configurations, as most of the samples are located at approximately the same place in both models. On the other hand, in the case of the two Maximum models, the compression-like effect that occurs in the classical MDS model has been eliminated in the NLM model, and therefore there is a considerable difference in the between-sample distances of the formerly compressed samples. The classical MDS models are much closer in their topology than in the NLM configurations.

6.7 Conclusions

Mathematically and conceptually, there are close correspondences between MDS and other methods used to reduce the dimensionality of complex data, such as Principal Components Analysis (PCA) and factor analysis. PCA is more focused on the dimensions themselves, and seeks to maximise explained variance, whereas MDS is more focused on relations

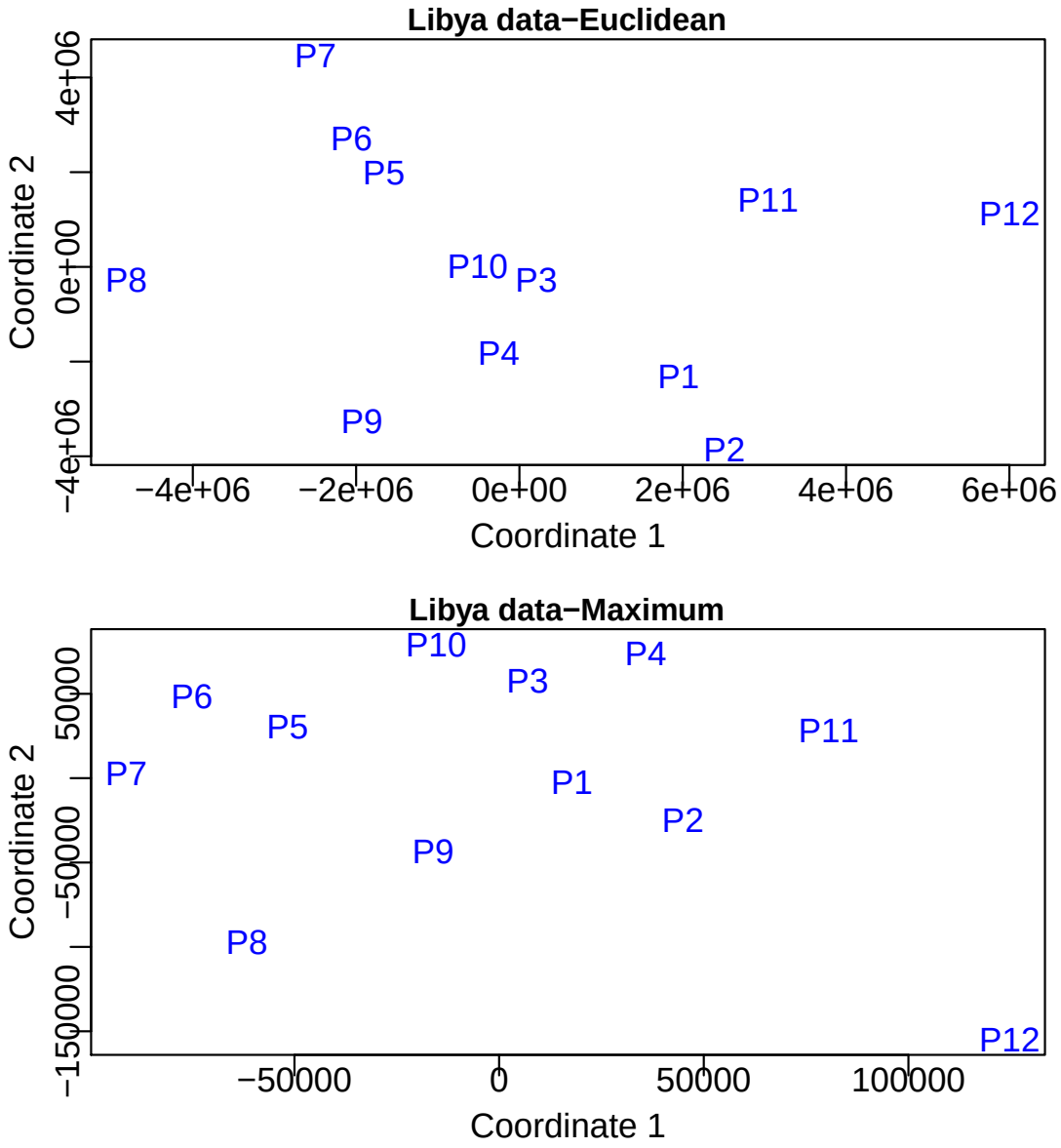


Figure 6.10: Two-dimensional solution of NLM MDS using the Euclidean (upper plot) and the Maximum (bottom plot) distance metrics for Libya data set. As initial configurations, the classical MDS models depicted graphically in Figure 6.5 have been used. The labels in the plots are the sample labels.

among the scaled objects. MDS projects p-dimensional data points onto a (commonly) 2-dimensional space such that similar objects in the p-dimensional space will be close together in the two-dimensional plot, while PCA projects a multidimensional space onto the

directions of maximum variability using the covariance or correlation matrix to analyse the correlations between data points and variables.

In this chapter, another data-projection method, with the advantage over PCA that it is flexible and can be used with any dissimilarity measure, is applied to the same propolis data sets as for PCA, for clustering purposes, namely, multidimensional scaling. More specifically, two MDS methods were described in detail and used, initially the classical MDS, and then the derived MDS configuration was used as input to the NLM method of MDS.

In the case of the initial configuration, results using four different distance metrics, Euclidean, Manhattan, Maximum and Canberra, were compared with the help of two criteria, M_2 and Mardia's criterion. Considering the results of the criteria, the Euclidean metric was the best, giving a consistently good fit of the original distances of the samples to the corresponding two-dimensional MDS space. The pattern recognition capability of the classical MDS models was tested both by examining the graphical representations of the models' configurations of the samples.

Results proved to be very consistent to those of PCA, but overall classical MDS did not improve the PCA findings or add more information to them.

In general, the Euclidean model proved to be similar to the second best metric, which was the Manhattan distance for data sets I, II, III and IV, and the Maximum distance for the Libya data, in term of the model fit.

Applying the NLM method to the data sets I, II, III and Libya, using the derived classical MDS models as the initial configuration, showed that only slight differences are observed between the classical MDS and the NLM results, when the Euclidean distance metric is used. On the other hand, in general, the compression-like effect of the points that had been observed in the classical MDS model has been remedied in the NLM and the samples are broadly spread in the NLM models.

The two MDS models were capable of reproducing quite successfully the findings of PCA, in terms of clustering the data, but they did not provide any further information on the

potential clustering of the samples and in general neither did the two NLM models.

The next chapters, Chapters 7 and 8, describe in detail two of the essential unsupervised classification techniques, for types of data such as metabolomics data, in the areas of hierarchical clustering, and partitioning methods such as hard clustering, in an attempt to identify suitable clustering models for these data and interpret the results of applying these to the various data sets studied. These methods, as they are explicitly designed to identify groupings present in the data sets I, II, III, IV and Libya, are expected to confirm the findings of PCA in Chapter 5.

Chapter 7

Cluster Analysis

After reducing the dimensionality of the data sets by PCA (and MDS), we consider techniques for clustering of samples of data into different groups. These techniques are unsupervised, with no samples in the data belonging to any pre-defined clusters and no prior knowledge of the number of clusters that are required to identify any similarity between samples. This chapter describes chapter analysis in general, and then focuses on the very commonly used practical method of hierarchical cluster analysis (HCA), described in Section 7.4, and applied to the data sets in Section 7.6. The focus is on the agglomerative clustering approach. Other techniques of clustering are considered in Chapters 8 and 9.

7.1 Overview

Cluster analysis, also known as clustering, refers to numerous statistical techniques used to structure or partition data into different groups or clusters of similar samples according to their characteristics. These techniques are not the same as supervised classification techniques, even though they refer to classification, which is a broader area of statistics. Cluster analysis can be used in two ways. One is for subdividing the data for analysis,

termed by Krzanowski and Marriott (1995) as dissection, or identifying natural groups in the given data. Theodoridis and Koutroumbas (2003) asserted that cluster analysis can be used in several applications, including group prediction, data reduction, and hypothesis generation and testing. It is also used to determine natural groupings in metabolomics data.

There are several categories of clustering algorithms, such as genetic clustering algorithms, sequential algorithms, competitive clustering algorithms, and cost function optimisation algorithms. Furthermore, there are subcategories of some clustering techniques, such as hierarchical clustering algorithms, that are further split into divisive algorithms and agglomerative nesting algorithms.

This chapter and Chapters 8 and 9 will examine clustering techniques that have been determined as the most appropriate for metabolomics data (Adams, 2007; Gordon, 1981; Lindon et al., 2001), including partitioning methods such as hierarchical agglomerative clustering, hard clustering algorithms and competitive learning algorithms. These techniques will then be applied to the metabolomics data sets used previously in Chapters 4, 5 and 6. This chapter is divided as follows: Section 7.2 will explain the different factors that need to be considered when applying cluster analysis; Section 7.3 will examine proximity measures that are used to represent data sets in cluster analysis, and Section 7.4 will discuss hierarchical clustering using the agglomerative nesting method. Chapter 8 will define the partitioning methods used with the k-means hard clustering algorithm, and chapter 9 will describe competitive learning algorithms, specifically SOM, a less used approach for metabolomics data, that are also tested here on the metabolomics data.

7.2 Considerations of Clustering

As shown by Theodoridis and Koutroumbas (2003), there are six aspects to consider in a cluster analysis: variable selection, proximity measures, clustering procedure, clustering

algorithm, quality of the clusters and clustering interpretation.

- **The selected variables:** The first step for cluster analysis is establishing which variables in the data are potentially relevant to the research questions. The variables selected may be a significant factor in deciding the proximity measure, clustering procedure and clustering algorithm to be used for the analysis. The goal of the research and the reason for applying cluster analysis should direct the researcher to suitable variables for the analysis. Further, before the data is used, it may need to be pre-processed and pre-treated.
- **Proximity measures:** The proximity measure indicates the similarity between pairs of objects, and what is suitable will depend on the nature of the data. The next Section, 7.3, will provide a detailed explanation of different proximity measures used in cluster analysis.
- **Clustering procedure:** The third step for cluster analysis is choice of the clustering procedure, such as Hierarchical methods or Partitioning methods. We will discuss these in Section 7.4.
- **Clustering algorithm:** Once the proximity measure and the clustering proceeder are selected, a clustering algorithm must be determined to obtain clusters from the data. In this chapter, clustering algorithms considered and used are Single, Complete, Average, McQuitty and Ward's linkage.
- **Quality of the clusters:** After performing cluster analysis, a measure such as the Silhouette Coefficient must be applied to assess the quality of the results. Section 7.5 will provide a detailed explanation of the Silhouette Coefficient.
- **Cluster explanation:** In this final step, the results of analysis are interpreted. For clearer interpretation, results of other statistical techniques, such as from PCA, may have to be used as well as cluster analysis, for example for visualisation of results.

It is also important to assess whether the data has a clustering tendency, that is, whether the data is suitable for cluster analysis. This can be achieved using different methods. The clustering solutions can be affected by the selected variables, clustering algorithms, or proximity measures.

7.3 Proximity Measures

Clustering indicates that objects in a data set are similar or dissimilar to each other, and thus a proximity or difference measure, to express closeness of objects, is required for clustering a set of objects into natural groups. The approach used for clustering objects into groups depends on the way the objects are presented for analysis. Gordon (1996) put forth two methods of representing data: the profile matrix and proximity matrices.

Profile Matrix, also known as the Pattern Matrix

The profile matrix is an $(n \times p)$ data or input matrix X containing elements x_{ik}, where x_{ik} is the observed value on the i^{th} object's k^{th} variable $(i = 1, ..., n, k = 1, ..., p)$. A majority of clustering techniques use this data representation as the input for clustering. In this thesis, x_{ik} is the intensity value that is noted in the i^{th} sample for the k^{th} metabolite (variable).

Proximity Matrices

Below are the two types of proximity matrices:

- **Dissimilarity Matrix**

An $(n \times n)$ matrix containing elements d_{ij} is a dissimilarity matrix D if the dissimilarity between the i^{th} and j^{th} objects $(i, j = 1, ..., n)$ is d_{ij}. Further, d is a dissimilarity coefficient that is a function mapping $\Phi \times \Phi$ to the real line $\mathbb{R}$, where Φ indicates the set of objects that are to be classified (Gordon, 1981; Lukasova, 1979). The properties of d are the following:

$$\begin{aligned} & d_{ij} \geqslant 0; \ \ \forall \ i, j \in \Phi, \\ & d_{ii} = 0; \ \ \forall \ i \in \Phi, \\ & d_{ij} = d_{ji}; \ \ \forall \ i, j \in \Phi \ \ \text{(symmetric)}. \end{aligned} \tag{7.1}$$

Further, as stated by Everitt (1993) and Kaufman and Rousseeuw (2009), if d satisfies the following

$$d_{ij} \leq d_{ih} + d_{hj} \ \ \forall i, j, h \in \Phi \ \ \text{(triangle inequality)} \tag{7.2}$$

then d can be considered as a distance function. Though Gordon (1981) and Kaufman and Rousseeuw (2009) have asserted that for a measure to be a dissimilarity, it need not satisfy (7.1) and (7.2), both these equations must be applicable for a distance measure. The following equation is the Minkowski metric, the most notable distance measure.

$$d_{ij}^{(q)} = \left\{ \sum_{k=1}^{p} |x_{ik} - x_{jk}|^q \right\}^{\frac{1}{q}} (q > 0). \tag{7.3}$$

Equation (7.3) demonstrates the city block (Manhattan) and the Euclidean metric for q = 1 and q = 2, respectively. Other studies (Everitt (1993); Everitt and Rabe-Hesketh (1997); Gordon (1981); Krzanowski and Marriott (1995)) have put forth several other dissimilarity measures, both with and without distance measures. The Chebyshev distance or maximum metric between a pair of objects is:

$$d_{ij} = \max |x_{ik} - x_{jk}|. \tag{7.4}$$

The Canberra distance is found using equation:

$$d_{ij} = \sum_{k=1}^{p} \frac{|x_{ik} - x_{jk}|}{|x_{ik}| + |x_{jk}|}. \tag{7.5}$$

- **Similarity Matrix**

An $(n \times n)$ matrix S is a similarity matrix if it consists of elements s_{ij}, wherein the similarity coefficient s_{ij} indicates the similarity between i^{th} and the j^{th} objects $(i, j = 1, ..., n)$. This similarity coefficient s_{ij} implies the closeness of objects i and j, indicating values ranging from 0 to 1, with 0 signifying complete dissimilarity between objects i and j and 1 signifying maximum similarity. Kaufman and Rousseeuw (2009) observed that for a similarity function, the conditions given below must be satisfied:

$$\begin{aligned} 0 \leq & s_{ij} \leq 1, \ \ \forall i, j \in 0, 1, \ldots n, \\ & s_{ii} = 1, \ \ \forall i \in 0, 1, \ldots, n, \\ & s_{ij} = s_{ji}, \ \ \forall i, j \in 0, 1, \ldots, n \ \ \text{(symmetric)}. \end{aligned} \tag{7.6}$$

Gordon (1981); Krzanowski and Marriott (1995); Everitt (1993); and Kaufman and Rousseeuw (2009) determined that similarities can be changed to dissimilarities if an appropriate transformation is used. Selecting the proximity measure for a clustering study largely depends on the data that is used in the analysis as well as the types of variables included in the data set.

7.4 Hierarchical Clustering Methods

This is a traditional statistical approach to cluster analysis, typically used with numerical data.

7.4.1 Overview

Hierarchical clustering algorithms are an important and commonly used clustering approach that requires multiple steps for establishing clusters. In this technique, every object in the data set is designated to only one cluster during each algorithm step, and thus this technique is a *hard* or *crisp* clustering method. Many studies have applied hierarchical cluster analysis (HCA), including for developing models for toxicology of drugs, such as in Harrigan et al. (2004) for classifying control rats and rats subjected to bacterial lipopolysaccharide and ranitidine for inducing hepatotoxicity to develop a predictive idiosyncratic toxicity model, as well as in Seltmann et al. (1994) for clinically isolating the Salmonella enteridis bacterium. HCA has also been used for metabolite profiling, such as in Griffin et al. (2000) for comparing and clustering metabolic profiles of the kidneys and urine of three wild mammals and a laboratory rat, acquired using 1H NMR spectroscopy, and in Want et al. (2006) for identifying similar metabolite features among several serum extraction methods applied to LC-MS generated metabolic profiles of human serum. David Watson has used HCA to cluster propolis samples in many of his studies, such as in Watson et al. (2006), who analysed 43 propolis samples collected in different parts of the world (Africa, Asia, Brazil, Europe and the Solomon Islands). The results showed chemical variation parallel to the different geographical origins of the propolis. The two important types of HCA algorithms are the agglomerative nesting and the divisive algorithms. The former is explained in detail in Subsection 7.4.2, as it is used in this thesis and a brief mention of the latter can be found in Subsection 7.4.3. Section 7.6 explains the HCA application on data sets I, II and III.

7.4.2 Agglomerative Nesting Algorithms

Everitt (1993) stated that the analysis in these algorithms begins with n clusters that contain one sample each, ending with all samples combined in one cluster. During every

step, the algorithm identifies the closest pair of different clusters using a pre-specified dissimilarity criterion, and then combines these clusters and decreases the number of clusters by one. The process ends after all the samples are combined into one cluster, that is, there is only one cluster left. The cluster similarity measures and the selected distance measure determine the result of these techniques.

Agglomerative nesting techniques are typically used in HCA studies, especially in those that concern mass spectrometry (MS) generated data such as the study by Mariey et al. (2001). They are also used in NMR metabolomics studies, including studies that identify similarities among metabolic profiles that are produced by 1HNMR spectroscopy of urine samples from control rats and rats subject to different doses of model compounds, to investigate toxicity and metabolic effects at various pre-specified times, as in Beckonert et al. (2003).

There are different ways (linkage methods) to define the closeness of any two clusters, using the specified distance or dissimilarity measure. Below are the commonly used ones:

Single Linkage

This is one of the simplest methods, and is also known as the nearest neighbour algorithm. As shown in Figure 7.1[1], the distance between two clusters is shown by the closest pair of samples, in the two clusters, wherein every pair consists of only one sample from each group. The dissimilarity between two clusters is as follows:

$$d_{C_pC_q} = \min_{\substack{i \in C_p \\ j \in C_q}} d_{ij},$$

where C_p and C_q indicate any two clusters, and i and j are samples from these clusters. This algorithm tends to identify elongated clusters, as clusters derived from single linkage typically form at low dissimilarities in the dissimilarity dendrogram, which illustrates the sequence of points and clusters joining together.

[1]Sources: https://www.saedsayad.com/clustering hierarchical.htm

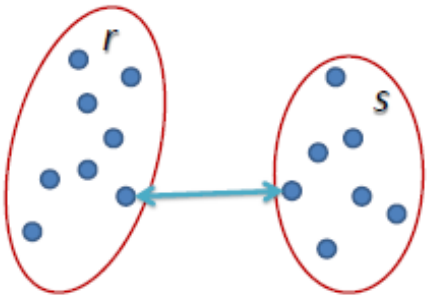


Figure 7.1: Illustrated example of single linkage method

Complete Linkage

Complete linkage is the exact opposite of single linkage. In this method, also called the farthest neighbour algorithm, the distance between two clusters is taken as that of the farthest pair of samples, consisting of one sample taken from each cluster, as shown in Figure 7.2[2]. The equation below gives the dissimilarity between two clusters:

$$d_{C_pC_q} = \max_{\substack{i \in C_p \\ j \in C_q}} d_{ij},$$

where C_p, C_q, i and j are defined as they were in single linkage. However, in contrast to single linkage, the clusters obtained in complete linkage are created at high dissimilarities in the dissimilarity dendrogram. This method is more suitable for finding spherical, compact, and small clusters, and is thus the preferred method if the data contains compact clusters.

Average linkage

The distance between two clusters in this method is determined by the average of the dissimilarities between pairs of samples in the two clusters, with each pair consisting of a

[2]Sources: https://www.saedsayad.com/clustering hierarchical.htm

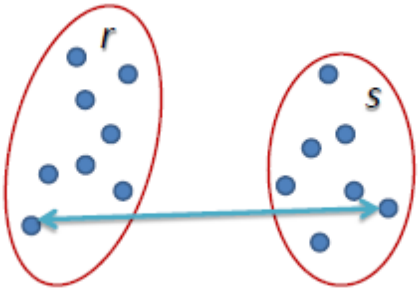


Figure 7.2: Illustrated example of complete linkage method

sample from each cluster. The equation below presents the distance between two clusters:

$$d_{C_f C_q} = \frac{1}{|C_f||C_q|} \sum_{f \in C_f} \sum_{q \in C_q} d(f, q) \tag{7.7}$$

where C_f and C_q indicate any two clusters where $|C_f|$ and $|C_q|$ are number of samples in clusters C_f and C_q, respectively, as stated by Mullner (2011). Kaufman and Rousseeuw (2009) asserted that as this method provides relatively spherical clusters, it can be considered as a compromise between the single and complete linkage methods, which are two extremes.

McQuitty Method

Similar to the average linkage method, the McQuitty method is a variant of this, which also provides relatively spherical clusters, offering a compromise between the two extreme single and complete linkage methods (Kaufman and Rousseeuw, 2009). The equation below presents this method's measure of dissimilarity between two clusters:

$$d_{C_f C_q} = \frac{1}{2} d_{C_i C_q} + \frac{1}{2} d_{C_j C_q}, \tag{7.8}$$

where C_f is the newly formed cluster where f is $(i \cup j)$ and C_q is the old cluster, as stated by Mullner (2011).

Ward's Method

Everitt (1993) and Krzanowski and Marriott (1995) observed that there was loss of information when two clusters were fused. Ward's method introduces a measure of cluster tightness to minimise this loss. The distance between two clusters is given by

$$d_{C_f C_q} = \sqrt{\frac{|f||q|}{|f|+|q|}} \|\bar{C}_f - \bar{C}_q\| \tag{7.9}$$

where $\bar{C}_f$ and $\bar{C}_q$ are the centroids of each of the two clusters, and $|f|$ and $|q|$ are the number of samples in clusters C_f and C_q, respectively.

The results of any of the above mentioned algorithms are usually represented graphically by means of a plot called a dendrogram (Jambu, 1978).

Algorithm for agglomerative nesting clustering

Izenman (2008) devised the following algorithm for agglomerative nesting clustering:

1. Input $\Omega = \{x_i, i = 1, 2, ..., N_c\}$ a set of multivariate samples, in which N_c indicates the number of clusters (and each cluster is a single point at the start).

2. Measure the $(N_c \times N_c)$ dissimilarity matrix $D = (d_{ij})$ between the N_c clusters, wherein $d_{ij} = d(x_i, x_j)$, $i, j = 1, 2, ..., N_c$ and the pre-selected dissimilarity measure is d.

3. Determine the smallest dissimilarity, from the dissimilarity matrix, for example d_{CiCj}, and form a new cluster C_{ij} by combining clusters C_i and C_j.

4. Use a pre-selected agglomerative method to calculate the dissimilarities between the new cluster C_{ij} and other clusters $C_k \neq C_i, C_j$, for example $d_{C_{ij}C_k}$.

5. Create a new $((N_c - 1) \times (N_c - 1))$ dissimilarity matrix, say $D^{(2)}$, removing from matrix D rows and columns C_i and C_j and adding a new row and column C_{ij}, using the computed dissimilarities in step 4.

6. Repeat steps 3, 4, and 5 $N_c - 1$ times such that the dissimilarity matrix $D^{(i)}$ at the i^{th} step is a symmetric matrix of size $((N_c - i + 1) \times (N_c - i + 1))$, where $i = 1, 2, ..., N_c$. $(i = N_c)$, and $D^{(N_c)} = 0$ at the final step because all clusters will have combined into one.

7. Output will specify which clusters have been merged at every step, depicted through a dissimilarity dendrogram, and the dissimilarity value or height of each merge.

Methodology

A banner plot can be considered as a horizontal barplot depicting the agglomerative clustering graphically. The values on the x-axis of the plot are the heights (levels) at which a merge of observations or clusters occurs, values from the minimum distance for the very first merge to the level of the value of the very last (final) merge. The overall width of (the red part of) a banner plot is important as it gives an idea of the amount of structure that has been found by the algorithm. When the between-cluster dissimilarities (and consequently the highest level) are much larger than the within-cluster dissimilarities, there is a clear cluster structure in the data, and the widths of the (red) bars in the banner are longer as objects merge earlier on.

The agglomerative coefficient (AC) can be calculated from such a plot, by taking the average of all the normalised widths of the bars in the banner (Kaufman and Rousseeuw, 2009). The labels on the y-axis on the right side of the plot correspond to a permutation of the original observations, such that the creation of a dendrogram with this ordering and merge information does not have any crossings of the branches. A banner can be plotted using the R function *bannerplot*() of package *cluster*.

The agglomerative coefficient (AC), will be used to assess whether HCA finds natural structure in the data or not. The agglomerative coefficient is defined as:

$$AC = 1 - \frac{d_{average}}{d_{final}}, \tag{7.10}$$

where $d_{average}$ is the average distance at which each object merges with one or more objects for the first time, and d_{final} is the distance at which all the objects are merged into one cluster. The AC will be calculated using R function $coef()$ of package $cluster$. This coefficient is a dimensionless quantity with values between 0 and 1. If the AC for a specific agglomerative analysis is small, then no clusters exist in the data. Hence the data consist of one big cluster. The closer to 1 the value of AC is, the clearer the clustering structure of the data is, i.e. the better the agglomerative method worked to identify clusters. However, the AC value can be affected by the existence of outliers in the data, so that it is necessary when AC is large to examine also the graphical output of the clustering analysis, such as dendrograms and silhouette plots, to ensure that the value of AC is representative of the clustering structure of the data.

To confirm the findings, two other statistics will be computed for all 20 methods, i.e. the Cophenetic correlation and the Gower distance (Borcard et al., 2011). The Cophenetic correlation is related to the dendrogram, which describes a hierarchical clustering method (see for example Figure 7.4). More specifically, the Cophenetic distance between two items in a dendrogram is defined as the distance at which the two items are joined to the same group. For a pair of items, starting from one of them, climbing up the dendrogram to the first node which leads down to the second item, the level of this node is the Cophenetic distance between the two items. Consequently, a Cophenetic matrix is a matrix which contains the Cophenetic distances between all pairs of items. It is then possible to compute a Pearson's r correlation, which is called the Cophenetic correlation, between the original dissimilarity matrix of a hierarchical clustering method and the Cophenetic matrix. The method with the highest Cophenetic correlation can be considered as the agglomerative method which produced the best clustering method for the distance matrix of the original data. An essential aspect of this statistic is that it depends strongly on the clustering method, independently of the data available for analysis.

Another measure of goodness of fit between the matrices is the Gower distance. This statistic is defined as the sum of squared differences between the values in the two matrices

(Legendre and Legendre, 1998). That is,

$$D_{Gower} = \sum_{i,j} (original(d_{ij}) - cophenetic(d_{ij}))^2, \tag{7.11}$$

The smaller the value of this statistic, the better the fit of the method to the original data. Similarly to the Cophenetic correlation, the Gower distance requires the results for comparison to be from the same original distance matrix. In addition, it is not necessarily true that both statistics (Cophenetic correlation and the Gower distance) will indicate the same clustering method as the best.

7.4.3 Divisive Clustering Algorithms

This approach works in reverse to agglomerative clustering. In this technique at the start, there is only one cluster that contains all samples from the data. At every step of the algorithm, the number of clusters is increased by one, so that at the end there are n clusters, each containing one sample. Thus, there are $2^{n-1} - 1$ non-trivial ways to classify the samples into two clusters for n samples present in the data. Therefore it is computationally infeasible to examine all possible divisions, even for cases with a moderate number of samples. As these algorithms require far more calculations than agglomerative methods do (Tamilselvi et al., 2015), they are not so popular, and these will not be described in detail or used in the application of HCA to the propolis metabolomics data.

7.5 The Silhouette Coefficient

The quality of the solution from a clustering algorithm is the quality of the derived partition, which must satisfy several requirements: the structure of the data; whether the 'within' cluster dissimilarities are smaller than the 'between' cluster dissimilarities; the number of 'natural' clusters in the data; and how well the samples are classified. This

can be assessed an different ways.

Rousseeuw (1987) developed a statistic to address these requirements named the silhouette coefficient. If d_{icj} indicates the average dissimilarity of any object i in the data to its cluster c_j, in which any two clusters c_j and c_i contain completely different objects, then

$$\beta_i = \min_{c_i \neq c_j} d_{ic_j} \tag{7.12}$$

and the cluster, for example c_k, that satisfies Equation (7.12) is the neighbour of object i. If object i was assigned to a cluster different than its cluster c_i, then the neighbouring cluster c_k would be the second best option. The following presents object i 's *silhouette width*, with values ranging from $-1 \leq s_i \leq 1$:

$$s_i = \frac{\beta_i - \alpha_i}{\max\{\alpha_i, \beta_i\}}$$

where α_i is the average dissimilarity of object i to all other objects in its own cluster, c_i. Values of s_i closer to 1 indicate that object i is well-clustered, whereas the object is misclassified if the value is closer to -1. The closer the value is to 0, the more unclear it is as to which cluster object i belongs, as object i remains between the assigned and neighbouring cluster. An average silhouette width can then be defined for each cluster.

The silhouette coefficient has been defined as the maximum of the average silhouette widths by Kaufman and Rousseeuw (2009), as it can be considered as a measure of the amount of structure that the clustering algorithm reveals. They further interpreted the silhouette coefficient values, as follows: the silhouette coefficient value ≥ 0.25 indicates no substantial structure; a value between 0.26 and 0.50 indicates weak and possibly artificial structure; a value between 0.51 and 0.70 indicates a reasonable structure; and a value between 0.71 and 1.00 indicates well-structured clustering. The clustering algorithm does not affect the silhouette coefficient, as it depends on the proximity matrix and the derived partition of objects.

7.6 Application of HCA to Data Sets I, II and III

Here we use agglomerative HCA on the metabolomics data sets for the propolis samples, using different distance measures and several linkage methods, to compare the results.

7.6.1 Overview

The data sets I, II and III (Aberdeenshire, Fort William and Dunblane respectively) that will be used in the hierarchical clustering analyses are the same as in the analyses of Chapters 5 and 6. That is, the data set contains the selected 27, 14 and 9 samples with 921, 511 and 498 variables respectively. The distance matrix of the samples will be computed using four different distance measures, i.e. Euclidean, Manhattan, Maximum and Canberra, for comparison purposes. Five different agglomerative nesting methods will be used in order to perform the HCA. These include Single linkage, Complete linkage, Average linkage, the Ward's method and the McQuitty method. The aim is to examine which method or methods are most successful on this kind of data.

To facilitate identification of the best clustering method for data sets I, II and III among the 20 combinations mentioned in the previous paragraph, various statistics will be computed and plotting tools will be used to compare the results of the clustering analyses. These tools include banner plots, the agglomerative coefficient, the Cophenetic correlation, the Gower distance and the silhouette coefficient and plot. In addition, the optimal number of clusters will be identified with the help of plotting tools such as graphs of silhouette widths and fusion levels. An important consideration in these analyses is that the above mentioned tools may prove sufficient to show the best hierarchical clustering method with regards to the available data.

7.6.2 Comparison of Hierarchical Clustering Results of the Data Sets I, II and III

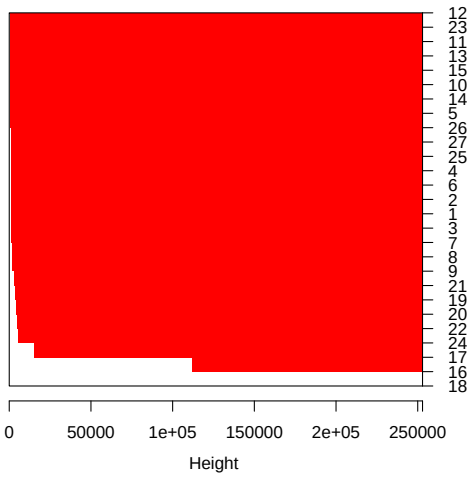
Some banner plots are shown in Figure 7.3 for these data sets. Table 7.1 gives the agglomerative coefficient values obtained from the analyses of the 20 hierarchical clustering methods described previously for data sets I, II and III respectively. From Table 7.1 it is clear that the AC of data sets I, II and III has the highest value, 0.997, 0.958 and 0.923 respectively for the method obtained by the Canberra distance metric using Complete linkage. The second best method for data sets I, II and III, obtained by the same distance metric for the Ward's linkage method, has agglomerative coefficient 0.991 (which is close to the best method's value), 0.927 and 0.901 respectively. In general, the Canberra metric seems to give the best results of all the considered metrics, and Ward's method provides the best results of the linkage methods.

In general, for data sets I, II and III, the Canberra - Complete approach gives the best results of the methods tried. The banner plots for this combination are shown in Figure 7.3.

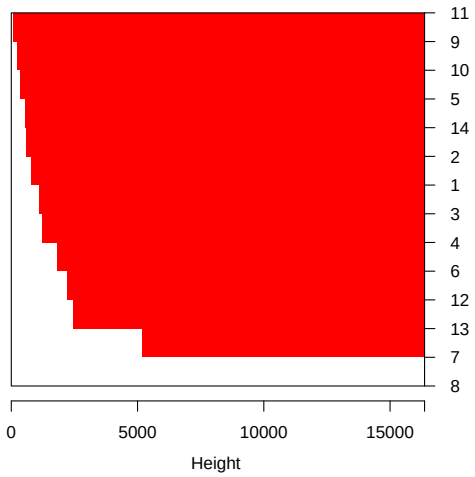
To confirm the findings, two other statistics will be computed for all 20 methods, i.e. the Cophenetic correlation and the Gower distance. Table 7.2 gives the Cophenetic correlation values for all 20 hierarchical clustering methods for data sets I, II and III. From Table 7.2, for the Aberdeenshire data the combination Euclidean - Average linkage has the largest Cophenetic correlation of 0.973 among all 20 methods (shown in blue in Table 7.2). Canberra - Single linkage has the lowest (worst) Cophenetic correlation of 0.122 among all 20 methods. The clustering method Canberra - Complete indicated with the largest AC value has Cophenetic correlation of only 0.159, shown as red in Table 7.2.

Therefore according to Cophenetic correlation, the appropriate method seems to be Euclidean - Average, as for the Aberdeenshire data.

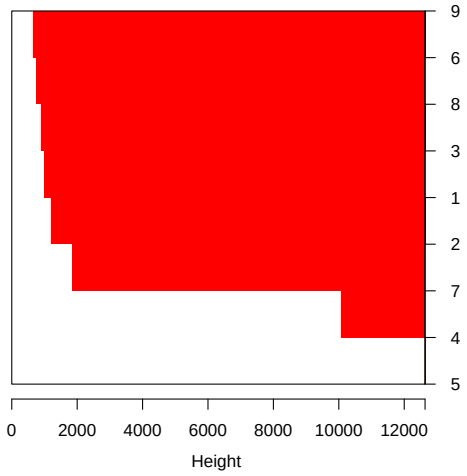
From Table 7.2 for the Fort William data, the Euclidean - Average linkage has the largest



(i) Aberdeenshire data (data set I)



(ii) Fort William data (data set II)



(iii) Dunblane data (data set III)

Figure 7.3: Banner plots of the partition of data sets I, II and III derived by Complete linkage using the Canberra distance metric. Height corresponds to the level of merge for a pair of observations, while the labels on the y-axis of the plot are the numbers of the samples in the data set.

Metric	Euclidean	Manhattan	Maximum	Canberra
Single	0.691	0.753	0.592	0.622
Complete	0.825	0.860	0.867	0.997
Average	0.802	0.838	0.814	0.914
Ward	0.930	0.942	0.925	0.991
McQuitty	0.786	0.828	0.788	0.900

(a) Aberdeenshire data (data set I)

Metric	Euclidean	Manhattan	Maximum	Canberra
Single	0.757	0.749	0.725	0.592
Complete	0.838	0.857	0.846	0.958
Average	0.805	0.821	0.784	0.786
Ward	0.902	0.909	0.890	0.927
McQuitty	0.813	0.825	0.811	0.798

(b) Fort William data (data set II)

Metric	Euclidean	Manhattan	Maximum	Canberra
Single	0.362	0.374	0.538	0.391
Complete	0.631	0.650	0.629	0.923
Average	0.531	0.558	0.555	0.771
Ward	0.692	0.703	0.692	0.901
McQuitty	0.566	0.561	0.565	0.768

(c) Dunblane data (data set III)

Table 7.1: Agglomerative coefficients for data set I, II and III. The clustering method with the largest agglomerative coefficient is shown in red. The underlined values are the next best.

Metric	Euclidean	Manhattan	Maximum	Canberra
Single	0.965	0.970	0.787	0.122
Complete	0.968	0.956	0.892	0.159
Average	0.973	0.972	0.899	0.162
Ward	0.963	0.960	0.869	0.143
McQuitty	0.972	0.970	0.870	0.135

(a) Aberdeenshire data (data set I)

Metric	Euclidean	Manhattan	Maximum	Canberra
Single	0.945	0.918	0.860	0.275
Complete	0.947	0.928	0.859	0.449
Average	0.954	0.943	0.876	0.447
Ward	0.938	0.923	0.866	0.451
McQuitty	0.953	0.942	0.871	0.446

(b) Fort William data (data set II)

Metric	Euclidean	Manhattan	Maximum	Canberra
Single	0.753	0.737	0.749	0.268
Complete	0.780	0.782	0.852	0.405
Average	0.801	0.787	0.884	0.410
Ward	0.798	0.756	0.847	0.404
McQuitty	0.777	0.787	0.883	0.409

(c) Dunblane data (data set III)

Table 7.2: Pearson's r Cophenetic correlation for the 20 hierarchical clustering methods for data sets I, II and III. The colours red and blue show the two clustering methods selected using as criteria the agglomerative coefficient and Cophenetic correlation respectively.

Cophenetic correlation of 0.954 among all 20 methods (shown in blue in Table 7.2). Canberra - Single has the lowest (worst) Cophenetic correlation of 0.275 among all 20 methods, shown as red in Table 7.2. Therefore according to this statistic, the appropriate method seems to be Euclidean - Average, as for the Fort William data.

From Table 7.2 for the Dunblane data, the combination Maximum - Average linkage has the largest Cophenetic correlation of 0.884 among all 20 methods (shown in blue in Table 7.2). Canberra - Single has the lowest (worst) Cophenetic correlation of 0.268 among all 20 methods. The clustering method Canberra - Complete indicated with the largest AC value has Cophenetic correlation of 0.405, shown as red in Table 7.2. Therefore according to this statistic, the appropriate method seems to be the Maximum - Average. In general the Canberra distance gives the poorest results in the terms of the Cophenetic correlation. The other distances all give good results in general.

From Table 7.2 for data sets I, II and III note that Manhattan - Average has the largest Cophenetic correlation among all five Manhattan methods (0.972, 0.943 and 0.787 respectively) and very close to the largest Cophenetic correlation for data set I. The Maximum - Average linkage has the largest (or nearly the largest) Cophenetic correlation among all five Maximum methods, and the Canberra - Average method has the largest Cophenetic correlation among all five linkage methods with the Canberra metric, and also this applies to the Euclidean metric.

In general, Average linkage has the largest Cophenetic correlation (or nearly the largest) among all five linkage methods.

Figures 7.4, 7.5 and 7.6 show the dendrograms of the Euclidean - Average linkage method for Aberdeenshire and Fort William, and the Maximum - Average method for the Dunblane data, because these are the best methods according to Pearson's r Cophenetic correlation for the 20 hierarchical clustering methods. The dendrograms make it easy to identify any patterns in their clustering solutions.

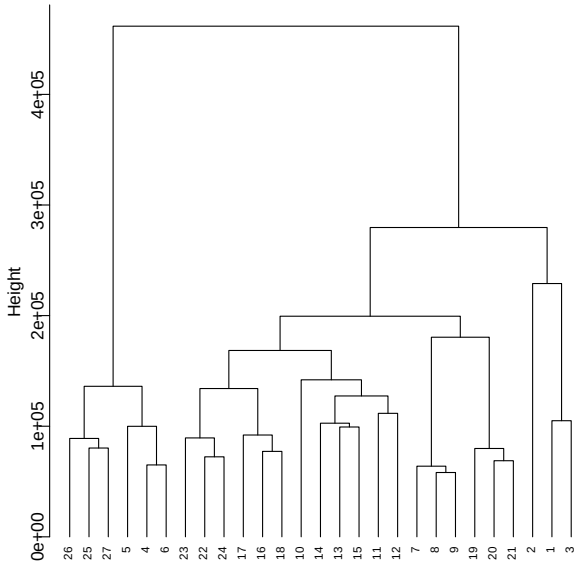


Figure 7.4: Dendrogram for the cluster partition derived by the Euclidean-Average linkage clustering method for the Aberdeenshire data. The labels at the end-leafs of the tree are the original numbers of the samples in the data set.

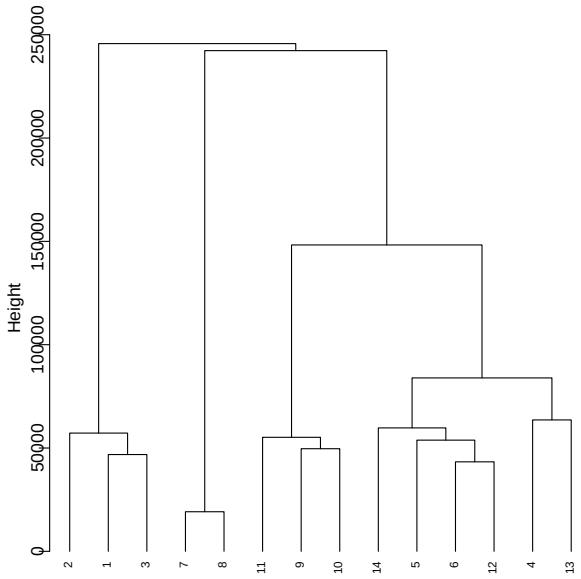


Figure 7.5: Dendrogram for the cluster partition derived by the Euclidean-Average linkage clustering method for the Fort William data. The labels at the end-leafs of the tree are the original numbers of the samples in the data set.

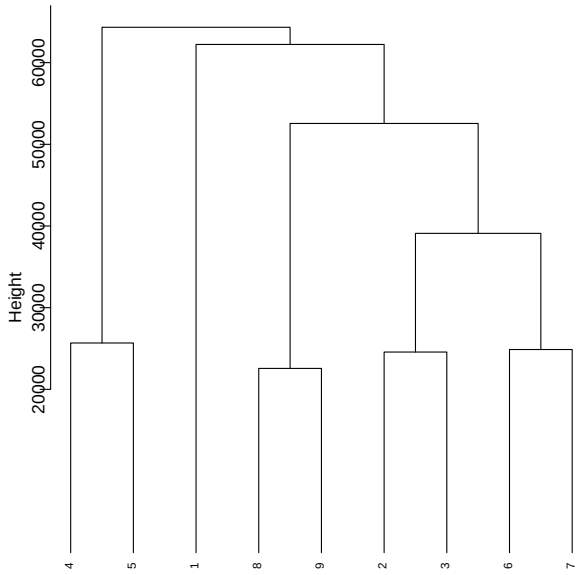


Figure 7.6: Dendrogram for the cluster partition derived by the Maximum-Average linkage clustering method for the Dunblane data. The labels at the end-leafs of the tree are the original numbers of the samples in the data set.

The Gower distance values for all clustering methods can be seen in Table 7.3 for data sets I, II and III. From Table 7.3, for Aberdeenshire, Fort William and Dunblane data respectively, the Gower distance value for Canberra - Average is the smallest compared to all other methods and each and every metric. The Euclidean - Average method had the highest Cophenetic correlation of 0.973 and 0.954 for data sets I and II (Aberdeenshire and Fort William) respectively, and the Maximum - Average had the highest Cophenetic correlation for data set III (Dunblane) (as seen in Table 7.2) of 0.884. So far the best methods in terms of the agglomerative coefficient are given by the Canberra - Complete method, which has the highest AC of 0.997 for Aberdeenshire, 0.958 for Fort William and 0.923 for Dunblane. These results are summarised in Table 7.4. However, further investigation is needed to confirm which of the three methods, Canberra - Complete, Euclidean - Average and Canberra - Average gives the best fit of the Aberdeenshire and Fort William data, and which the three methods of Canberra - Complete, Maximum -

Metric	Euclidean	Manhattan	Maximum	Canberra
Single	4.7039	1130.6650	1.2022	0.1414
Complete	1.5867	662.9914	0.8076	9.5343
Average	0.3696	101.1034	0.1798	0.1299
Ward	101.5430	26290.9900	7.6812	1.2413
McQuitty	0.5375	132.5807	0.2554	0.1306

(a) Aberdeenshire data (data set I)

Metric	Euclidean	Manhattan	Maximum	Canberra
Single	0.2459	55.6075	0.0309	0.0005
Complete	0.2327	57.5462	0.0573	0.0087
Average	0.0524	10.6052	0.0120	0.0002
Ward	3.3412	571.5541	0.2404	0.0013
McQuitty	0.0578	10.9969	0.0157	0.0003

(b) Fort William data (data set II)

Metric	Euclidean	Manhattan	Maximum	Canberra
Single	0.0914	14.5410	0.0063	0.0005
Complete	0.0877	13.2677	0.0050	0.0020
Average	0.0279	4.4877	0.0018	0.0002
Ward	0.2585	45.7569	0.0228	0.0010
McQuitty	0.0337	4.4967	0.0018	0.0003

(c) Dunblane data (data set III)

Table 7.3: Gower distance for the 20 hierarchical clustering methods for data sets I, II and III. In red is shown the clustering method with the smallest Gower distance value. The values shown in the table have all been divided by 10^{12} for ease of displaying them.

Average and Canberra - Average gives the best fit for the Dunblane data.

Table 7.4: The different methods determined as the best clustering methods for the three data sets using 3 different criteria.

Data	Agglomerative coefficient	Cophenetic correlation	Gower distance
Aberdeenshire	Canberra-Complete	Euclidean-Average	Canberra-Average
Fort William	Canberra-Complete	Euclidean-Average	Canberra-Average
Dunblane	Canberra-Complete	Maximum-Average	Canberra-Average

7.6.3 Identification of the Optimal Number of Clusters of Data Sets I, II and III

An important part of the clustering procedure is to decide at what level to cut the dendrogram of a clustering solution. This decision can be taken either subjectively by choosing the number of clusters from visual inspection of the dendrogram, or chosen to satisfy some criteria.

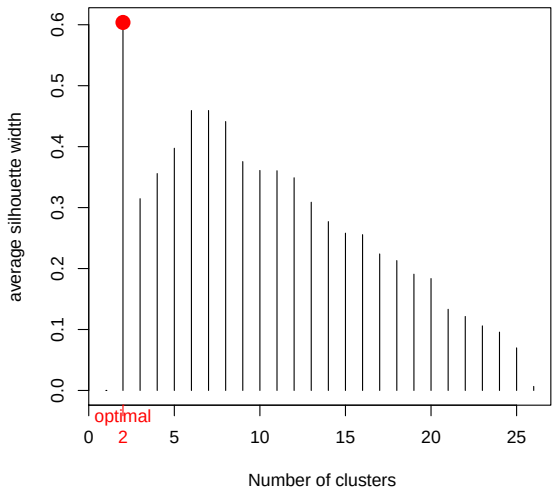
We will use a command *NbClust* in *R* to compare between several indices to determine the value of k to use in the analysis for the three clustering methods mentioned previously in Table 7.4. The index to be calculated can be one of the following: "kl", "ch", "hartigan", "ccc", "scott", "marriot", "trcovw", "tracew", "friedman", "rubin", "cindex", "db", "silhouette", "duda", "pseudot2", "beale", "ratkowsky", "ball", "ptbiserial", "gap", "frey", "mcclain", "gamma", "gplus", "tau", "dunn", "hubert", "sdindex", "dindex", "sdbw", "all" (all indices except GAP, Gamma, Gplus and Tau), "alllong" (all indices with Gap, Gamma, Gplus and Tau included) (Charrad et al., 2014). Applying this command to data sets I, II and III, according to the majority rule in Table 7.5, the best numbers of clusters are 2, 4 and 4 for the data sets I, II and III respectively.

Silhouette widths and plots of the fusion level values are two methods which can be used

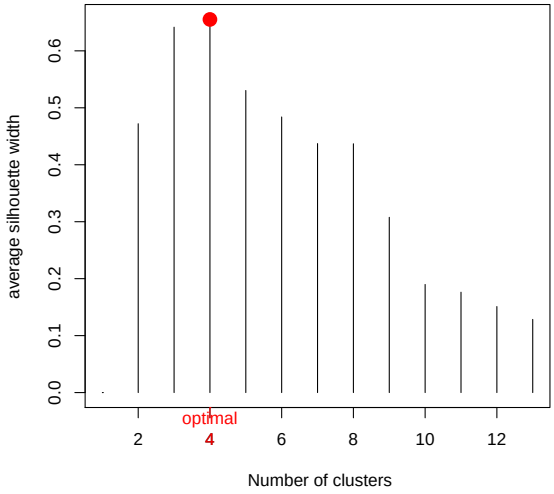
Aberdeenshire		
Canberra-Complete	Euclidean-Average	Canberra-Average
7 proposed 2 as the best number of clusters	7 proposed 2 as the best number of clusters	7 proposed 2 as the best number of clusters
1 proposed 3 as the best number of clusters	7 proposed 3 as the best number of clusters	1 proposed 3 as the best number of clusters
5 proposed 4 as the best number of clusters	4 proposed 4 as the best number of clusters	4 proposed 4 as the best number of clusters
2 proposed 5 as the best number of clusters	4 proposed 6 as the best number of clusters	5 proposed 5 as the best number of clusters
3 proposed 6 as the best number of clusters	1 proposed 10 as the best number of clusters	1 proposed 6 as the best number of clusters
1 proposed 7 as the best number of clusters		1 proposed 7 as the best number of clusters
2 proposed 9 as the best number of clusters		3 proposed 9 as the best number of clusters
2 proposed 10 as the best number of clusters		1 proposed 10 as the best number of clusters
Fort William		
Canberra-Complete	Euclidean-Average	Canberra-Average
2 proposed 2 as the best number of clusters	3 proposed 2 as the best number of clusters	4 proposed 2 as the best number of clusters
1 proposed 3 as the best number of clusters	13 proposed 4 as the best number of clusters	1 proposed 3 as the best number of clusters
10 proposed 4 as the best number of clusters	2 proposed 5 as the best number of clusters	8 proposed 4 as the best number of clusters
3 proposed 6 as the best number of clusters	2 proposed 9 as the best number of clusters	1 proposed 7 as the best number of clusters
1 proposed 8 as the best number of clusters	3 proposed 10 as the best number of clusters	6 proposed 9 as the best number of clusters
6 proposed 10 as the best number of clusters		3 proposed 10 as the best number of clusters
Dunblane		
Canberra-Complete	Euclidean-Average	Canberra-Average
1 proposed 3 as the best number of clusters	3 proposed 3 as the best number of clusters	2 proposed 3 as the best number of clusters
6 proposed 4 as the best number of clusters	6 proposed 4 as the best number of clusters	7 proposed 4 as the best number of clusters
3 proposed 5 as the best number of clusters	3 proposed 5 as the best number of clusters	1 proposed 5 as the best number of clusters
2 proposed 6 as the best number of clusters	1 proposed 6 as the best number of clusters	2 proposed 6 as the best number of clusters
11 proposed 7 as the best number of clusters	10 proposed 7 as the best number of clusters	11 proposed 7 as the best number of clusters

Table 7.5: The optimal number of clusters for data sets I, II and III, using 23 different criteria and 3 clustering methods.

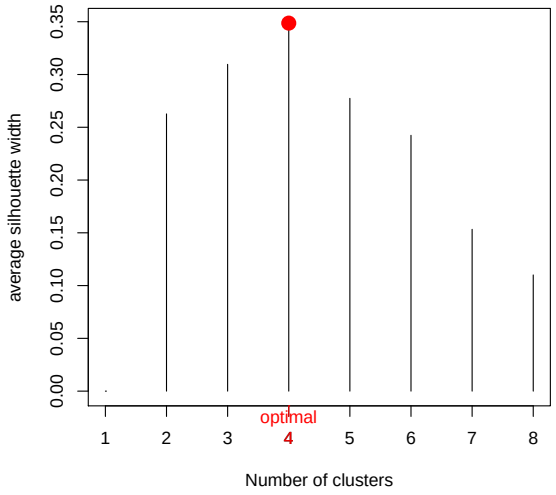
to define criteria for the appropriate number of clusters. As has already been described in detail in Section 7.5, the silhouette width is a measure of the degree of membership of an item to its cluster. This measure can be computed and the obtained values drawn in a bar plot for all possible numbers of clusters in a clustering solution. The *R* function *silhouette*() of package *cluster* will be used to obtain such a plot for the clustering solutions of Aberdeenshire, Fort William and Dunblane data. This plot will be drawn for the clustering methods for the data being discussed, namely the Euclidean - Average method, for Aberdeenshire and Fort William, and the Maximum - Average method for Dunblane, which have been identified as those methods with most potential as the best-fitting clustering using the Cophenetic correlation in Table 7.4.



(i) Aberdeenshire data (data set I)

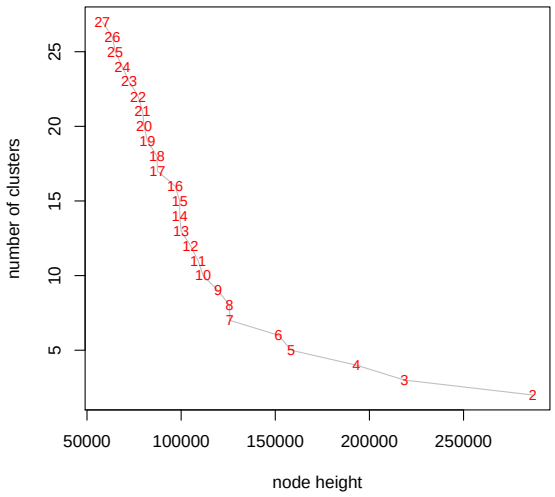


(ii) Fort William data (data set II)

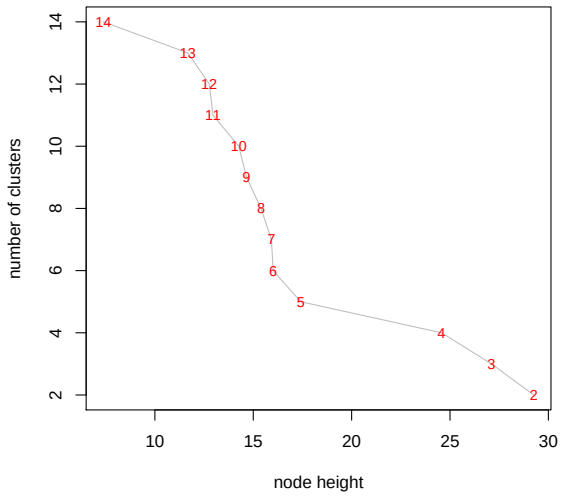


(iii) Dunblane data (data set III)

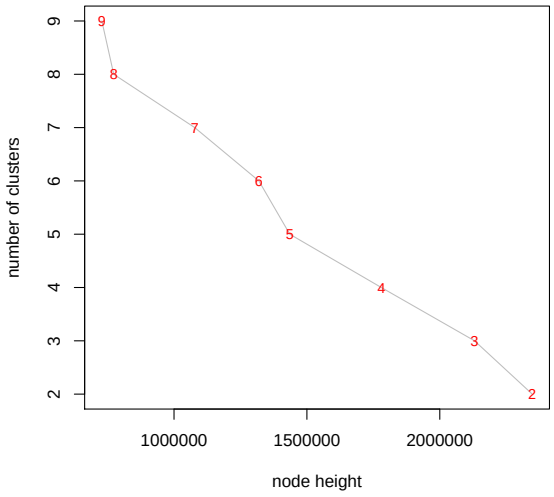
Figure 7.7: Average silhouette widths for partitions of 2-27 and 2-14 clusters for data sets I and II using Euclidean - Average clustering and 2-9 clusters for data set III using Maximum - Average clustering (the best method using Cophenetic correlation in Table 7.4). The optimal number of clusters is indicated in red in each case.



(i) Aberdeenshire data (data set I)



(ii) Fort William data (data set II)



(iii) Dunblane data (data set III)

Figure 7.8: Graphs of the fusion level values of the corresponding dendrograms for the best clustering methods in terms of Cophenetic correlation of data sets I, II and III (in Table 7.4). The numbers in red are the number of clusters obtained at specific node heights using Euclidean - Average for data sets I and II and Maximum - Average for data set III.

Figure 7.7 illustrates the average silhouette widths for all partitions, from 2-27, 2-14 and 2-9 clusters for the data sets I, II and III, for the three clustering methods mentioned previously (in Table 7.4). It is clear that in all methods the optimal number of clusters is 2 for Aberdeenshire, 4 for Fort William and 4 for Dunblane.

The fusion level values of a dendrogram can also be plotted, and from this plot, the optimal number of clusters can be identified. A fusion level value is the distance at which a merge or fusion between two branches of a dendrogram occurs. Figure 7.8 shows the fusion level values corresponding to the dendrograms of the best clustering methods from Cophenetic correlation (in Table 7.4) of data sets I and II. Reading the graphs from right to left for the Aberdeenshire data (2 clusters to 27 clusters) and (2 clusters to 14 clusters) for the Fort William data, it can be seen that for Euclidean - Average there is a large jump after the two-clusters fusion and the four-clusters fusion for data sets I and II respectively. Therefore, plotting tools indicate that the optimal number of clusters is 2 and 4 for data sets I and II.

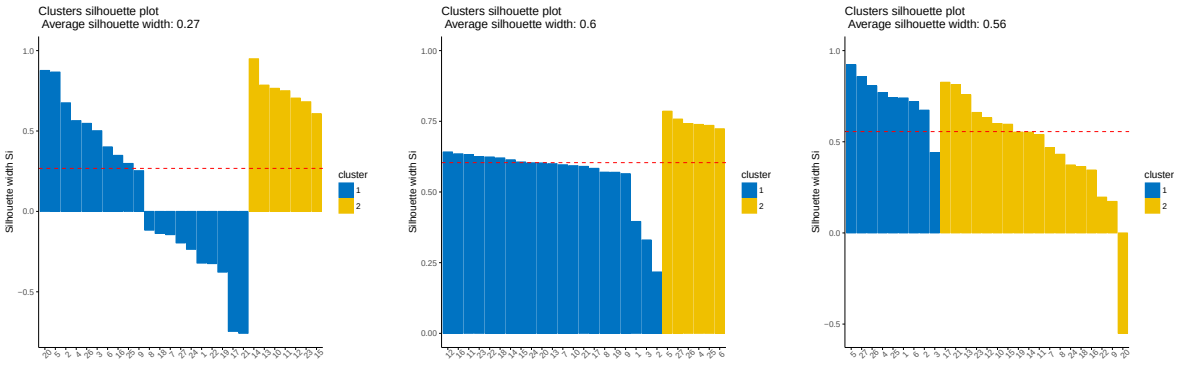
From Figure 7.8, reading the graphs from right to left for the Dunblane data (2 clusters to 9 clusters) for the best clustering methods of Cophenetic correlation (in Table 7.4) of data set III, it can be seen that for Maximum - Average there is a large jump after the three and four-clusters fusion (it is not clear which is the best). Therefore, this plotting tool indicates that the optimal number of clusters is 3 or 4, and from Table 7.5, the majority vote for the number of clusters is 4 for data set III.

7.6.4 Identification of the Best Method for data sets I,II and III

Although dendrograms (e.g. Figure 7.4) and heat maps (e.g. Figure 7.11) illustrate the clustering results achieved by the application of a clustering method to data sets I, II and III, another type of graphical tool, the silhouette plot (based on the silhouette widths) can show how well each and every sample has been assigned to its respective cluster after the classification process, i.e. to what degree a sample is a member of its cluster. I will

discuss and determine which result obtained in Table 7.4 should be retained for further analyses of data sets I, II and III, as follows:

- **Aberdeenshire data (I)**



(i) Canberra-Complete method (ii) Euclidean-Average method (iii) Canberra-Average method

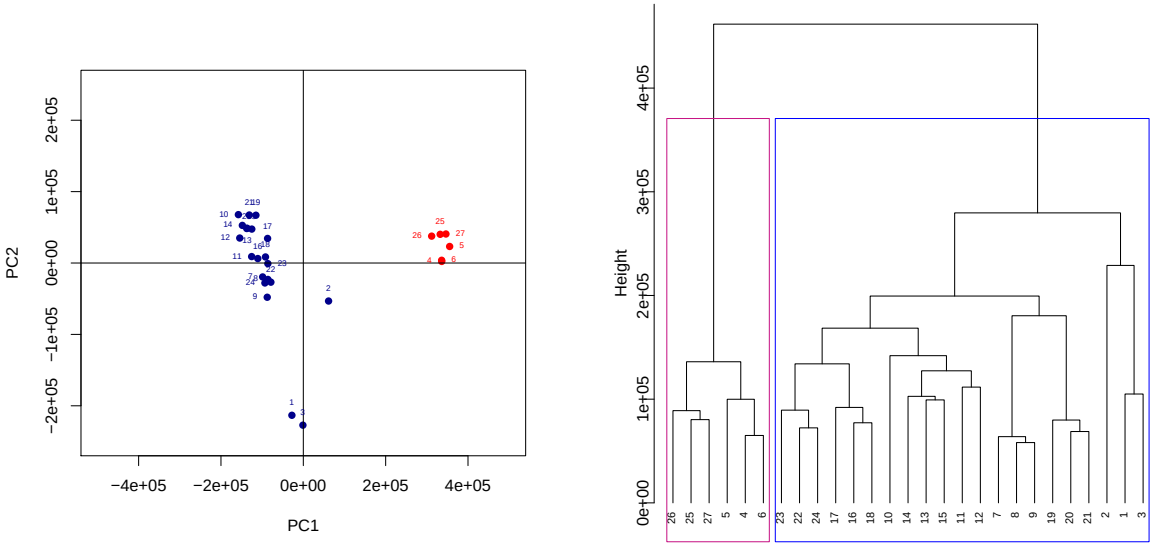
Figure 7.9: Silhouette plot for the 2-cluster partition derived by three clustering methods for the Aberdeenshire data.

The silhouette plots for the three clustering methods for the Aberdeenshire data can be seen in Figure 7.9 for the Canberra - Complete, Euclidean - Average and Canberra - Average methods. The silhouette width values of all samples can be seen in the silhouette plot as bars. It is, therefore, clear, which samples lie well within their cluster. The wider the silhouette bar for a sample is, the larger the silhouette value for this sample and the better the sample lies in the cluster, that is, the within cluster dissimilarity of the sample is much smaller than the smallest dissimilarity of the sample to other clusters. On the other hand, the Average Silhouette widths for the clusters differ considerably for the three methods, as the number of samples in the two clusters are not the same. The clustering methods have Average Silhouette widths for the entire data of 0.27, 0.60 and 0.56 respectively (which is also the Silhouette coefficient for the three methods respectively), therefore there

is a difference between them, as the method that has 0.60 (Euclidean - Average) is better than the other methods (Canberra - Complete) and (Canberra - Average). In addition, the silhouette plot for method Canberra - Complete shows that there are 10 samples clearly misclassified (highly negative silhouette widths) as members of cluster 1, while according to the silhouette plot these should have been members of cluster 2. Model Canberra - Average is definitely more balanced, with only 1 sample misclassified in cluster 2 (actually belonging to cluster 1) and having in general smaller negative silhouette widths than the misclassified samples in method Canberra - Complete. Euclidean - Average is definitely the best method, without any samples misclassified. The findings and the information obtained by the silhouette plots practically mean that only the Euclidean - Average method should be retained for further analyses of Aberdeenshire data, as it is seen to be the best.

To illustrate the clustering solution derived by the Euclidean - Average method for Aberdeenshire, a number of graphical tools will be used. A two-dimensional projection of the clustering solution can be seen in Figure 7.10 (i). The first two principal component scores (according to the results from Chapter 5) can be seen superimposed with the partition derived by the Euclidean - Average method. In the scores plots, colours represent the samples clustered to the cluster such that points labelled correspond to the sample numbers. From Figure 7.10 (i) it can be seen that the scores plot show clearly that there is no clear separation between samples according to each group of three samples coming from the same hive. However, the samples (hives) were located on two different sites (colonies) where sample numbers 4, 5, 6 and 25, 26, 27 were on Site 1, and the rest of the hives were on site 2 (Saleh et al., 2015). In the plot, there is a clear separation based on the site, since 4, 5, 6 and 25, 26, 27 are in the same cluster and are separated from the other samples.

Figure 7.10 (i) shows the most clearly clustered samples, which were obtained from the twenty-seven hives sampled in Aberdeenshire, based on the Euclidean - Average linkage clustering method. To some extent samples 1-3 are also different from the



(i) Scores plots (ii) Dendrogram

Figure 7.10: (i) Scores plots of the first two PCs, superimposed with the 2-cluster partition derived by the Euclidean - Average clustering method for Aberdeenshire data. Blue and red points represent samples in the first and second cluster. The labels of the points in the plot correspond to sample numbers. (ii) Dendrogram for the cluster partition derived by the Euclidean - Average linkage clustering method of Aberdeenshire data. The labels at the end-leafs of the tree are the numbers of the samples in the data set. The blue and red rectangles show the two clusters.

others.

A dendrogram for the clustering partition derived by the Euclidean - Average method for Aberdeenshire can be seen in Figure 7.10 (ii). The labels at the end-leafs of the tree correspond to the numbers of samples. A dendrogram can also be represented, perhaps more accurately, by a heat map, a square matrix of coloured pixels such that the colour intensity represents the similarity among the samples. The heat map of the distance matrix re-ordered according to the dendrogram of Figure 7.10 (ii) can be seen in Figure 7.11. The re-ordering of the heat map sorts the matrix such that most of the darker values representing high similarities are located closer

to the main diagonal.

Despite the optimal number of clusters being 2, it might be useful to examine

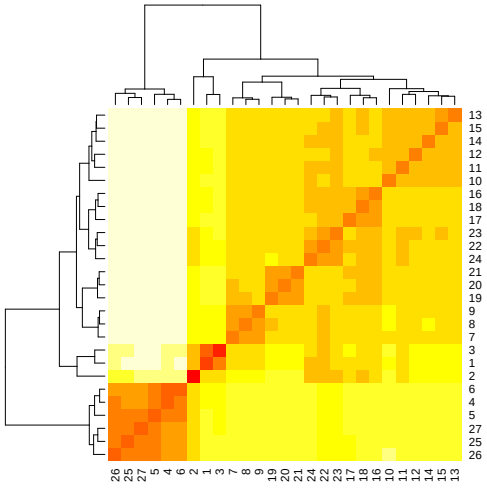


Figure 7.11: Heat map of the distance matrix of the Euclidean - Average clustering method according to the dendrogram of Figure 7.10 (ii) for Aberdeenshire. The colour intensity represents the similarity among the samples, such that the darker the colour the closer the similarity.

whether partitions of larger number of clusters than 2 can provide an insight into the discrimination based on location or compounds. Figure 7.12 shows the results of the partitions of 3-6 clusters, obtained from clustering method Euclidean - Average, for the Aberdeenshire data. The colours of the points in the scores plots correspond to the clusters in each partition. In the 3-cluster partition, the right cluster (red points) is the cluster of size 6, which remains unchanged up to and including the 6-clusters partition, and the top cluster (black points) is the cluster of size 3, which changed in the 4-cluster partition, then is unchanged until the 6-clusters partition. The left most cluster (green points) is the largest cluster in all partitions, which keeps being broken into smaller partitions until the 6-cluster partition. As in the 2-cluster partition, Figure 7.12 shows that there is no clear discrimination between

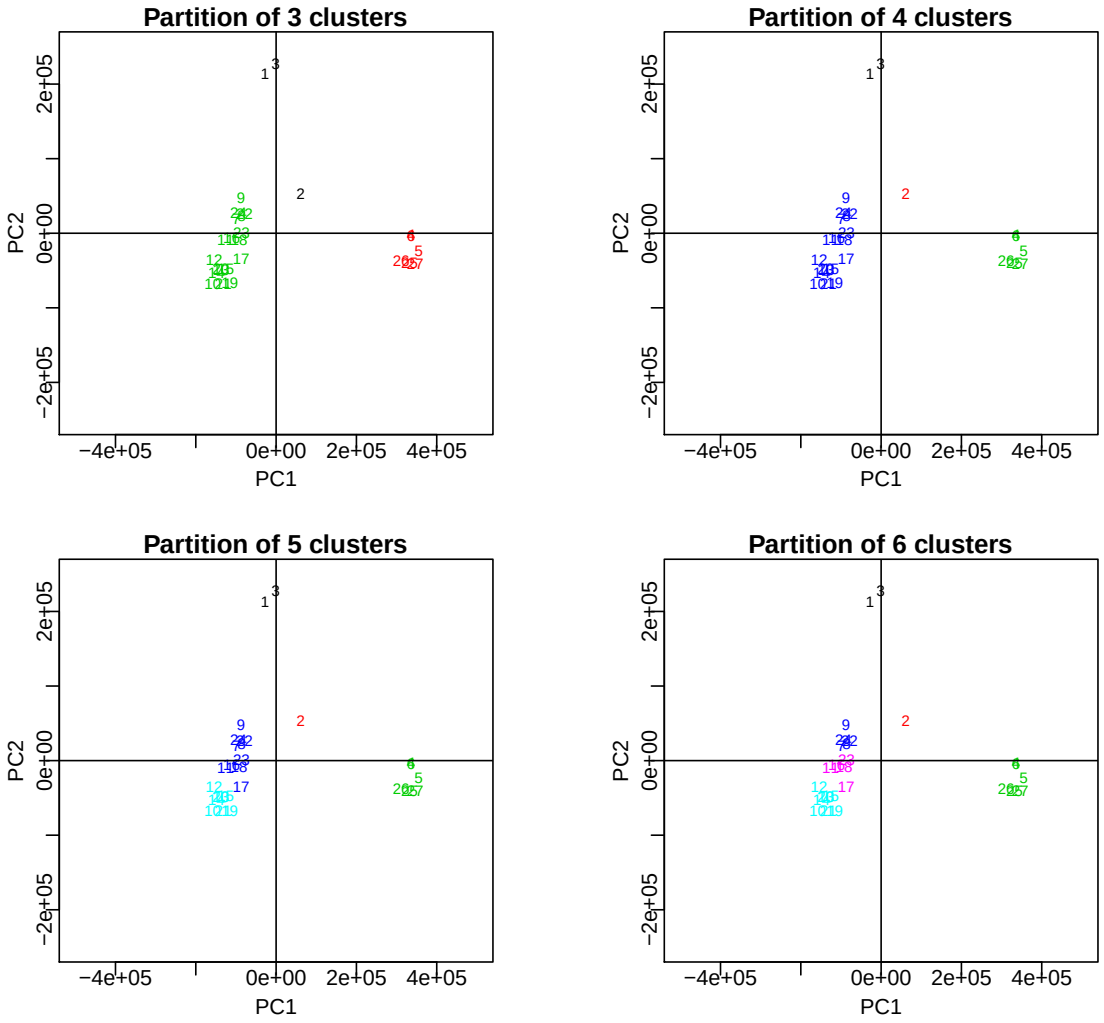


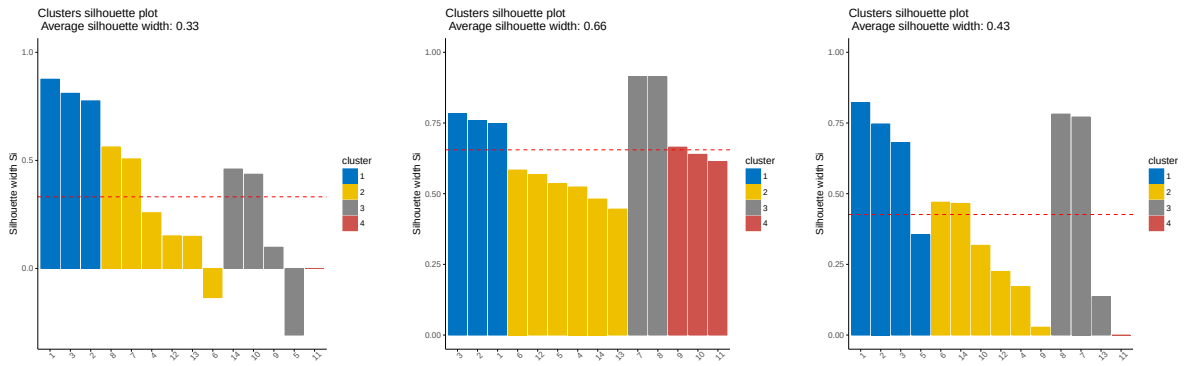
Figure 7.12: Scores plots of the first two PCs, superimposed with the cluster partitions for 3-6 clusters for Aberdeenshire data, derived by the Euclidean - Average clustering method. The labels of the points in the plots correspond to the sample numbers, and the colours indicate different clusters.

samples depending on individual hive in any of the four partitions (where each group of 3 samples belongs to the same hive) but again points 4, 5, 6 and 25, 26, 27 are shown as different from the others in the space given by the first 2 PCs.

The most important compounds distinguishing hives 4, 5, 6 and 25, 26, 27 from the rest are flavonoids and in particular methylated flavonoids (Saleh et al., 2015).

From the second group (coloured blue in Figure 7.10 (i)) samples 19, 20, 21, which are separated from 4, 5, 6 and 25, 26, 27, have glycerol esters of phenylpropanoid compounds as important compounds. Sample numbers 10, 11, 12 and 13, 14, 15 are characterised by esters of pinobanksin. In general, the second group has esters as important compounds.

- **Fort William data (II)**



(i) Canberra-Complete method (ii) Euclidean-Average method (iii) Canberra-Average method

Figure 7.13: Silhouette plot for the 4-cluster partition derived by three clustering methods for Fort William data.

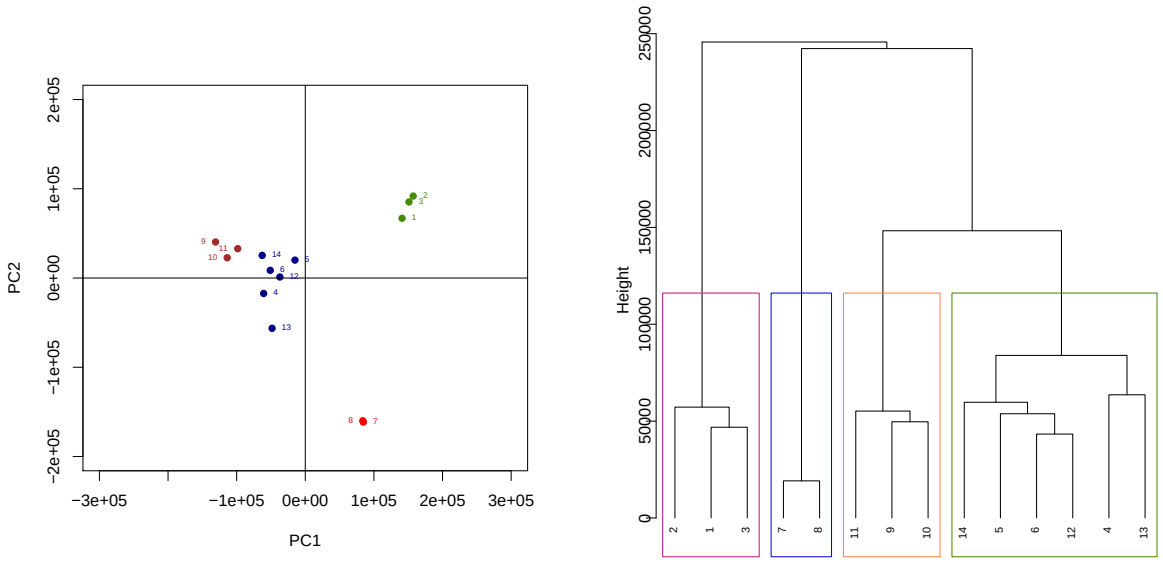
The silhouette plots for the three clustering methods for Fort William data can be seen in Figure 7.13 for the Canberra - Complete, Euclidean - Average and Canberra - Average methods. Silhouette width values of all samples can be seen in the silhouette plot as bars. The Average Silhouette widths for the clusters differ considerably between the three methods, as the number of samples in the four clusters are not the same. The three clustering methods have Average Silhouette width for the entire data set of 0.33, 0.66 and 0.43 respectively (which is also the Silhouette coefficient), therefore there is a difference between them as the method that has a width of 0.66 (Euclidean - Average) is better than the other methods (Canberra - Complete) and (Canberra - Average). In addition, the silhouette plot for method Canberra

- Complete shows that there are 2 samples clearly misclassified (highly negative Silhouette widths) as members of cluster 2 and 3 respectively. Model Canberra - Average is definitely more balanced, without any samples misclassified. Model Euclidean - Average is definitely the best method, without any samples misclassified, and it has the best Average Silhouette width of 0.66. The findings practically mean that only the Euclidean - Average method should be retained for further analyses of the Fort William data.

To illustrate the clustering solution derived by the Euclidean - Average method for Fort William, again a number of graphical tools will be used. A two-dimensional projection of the clustering solution can be seen in Figure 7.14 (i). The first two principal component scores can be seen superimposed with the partition derived by the Euclidean - Average method. From Figure 7.14 (i), the scores plots show clearly that there is some discrimination between samples in Fort William, according to the hive replicates, as samples such as 1, 2 and 3 cluster near each other. Figure 7.14 (i) shows that the fourteen samples of propolis from Fort William could be classified into four groups by hierarchical cluster analysis without splitting the replicates. In this case, the samples 4, 5 and 6 had the most average composition and the same samples are duplicate samples taken from the same hive.

The composition of these propolis samples appears to be fairly different from the Aberdeenshire samples. They differ from each other, but overall the compounds in Table 5.4 are in many cases not the same as most of the Aberdeenshire samples. The Fort William samples are rich in compounds putatively identified as sesquiterpene acids. Since the third group (blue colour) is towards the centre of the plot, that means it has important compounds from the other outlying groups in more moderate amounts. Thus the samples (hives) in the middle of the PCA plot are likely to come from several different sources of propolis, whereas the groups towards the periphery of the plot may focus on more restricted sources, such as samples 1, 2, 3, 7, and 8.

A dendrogram for the partition derived by the Euclidean - Average clustering



(i) Scores plots (ii) Dendrogram

Figure 7.14: (i) Scores plots of the first two PCs, superimposed with the 4-cluster partition derived by the Euclidean - Average clustering method for the Fort William data. Coloured points represent the samples in the clusters. The labels of the points in the plot correspond to sample numbers. (ii) Dendrogram for the cluster partition derived by the Euclidean - Average linkage clustering method for the Fort William data. The labels at the end-leafs of the tree are the numbers of the samples in the data set. The coloured rectangles mark the clusters chosen.

method can be seen in Figure 7.14 (ii). The heat map of the distance matrix re-ordered according to the dendrogram of Figure 7.14 (ii) can be seen in Figure 7.15.

Figure 7.16 illustrates the results derived from 2, 3, 5 and 6 cluster partitions for the Fort William data, for comparison. In the 2-cluster partition, the right cluster (black points) is the cluster of size 3, which remains unchanged up to and including the 6-clusters partition, and the left cluster (red points) is the cluster of size 11, which keeps being broken into smaller partitions until the 6-cluster partition, where the left-most partition is broken for the sixth cluster partition. Figures 7.14 (i) and 7.16 show clearly that there is discrimination between groups of samples from the

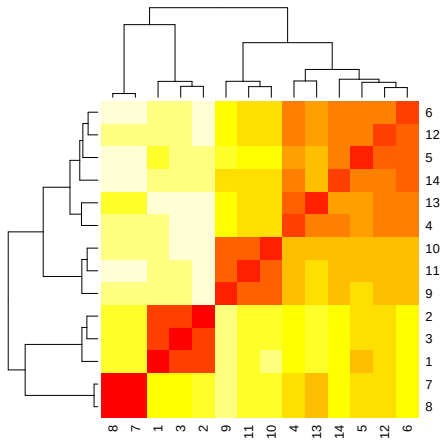


Figure 7.15: Heat map of the distance matrix of the Euclidean - Average clustering method according to the dendrogram of Figure 7.14 (ii) for Fort William. The colour intensity represents the similarity among the samples, such that the darker the colour the closer the similarity.

same hive in the second, third and fourth partition solutions. Note that samples 7 and 8 in this data set come from the same hive, but sample 9 comes from a different hive, the same one as samples 10 and 11. Conversely, also the 5 and 6 cluster partitions do not keep all samples from the same hive together in the same cluster.

- **Dunblane data (III)**

The silhouette plots for the three clustering methods for the Dunblane data can be seen in Figures 7.17 for the Canberra - Complete, Maximum - Average and Canberra - Average methods. Silhouette width values for all samples can be seen in the silhouette plot as bars. The Average Silhouette widths for the clusters again differ considerably between the methods, as the numbers of samples in the four clusters are not the same. The three clustering methods have Average Silhouette widths for the entire data set of 0.35, 0.40 and 0.45 respectively (which are also the Silhouette

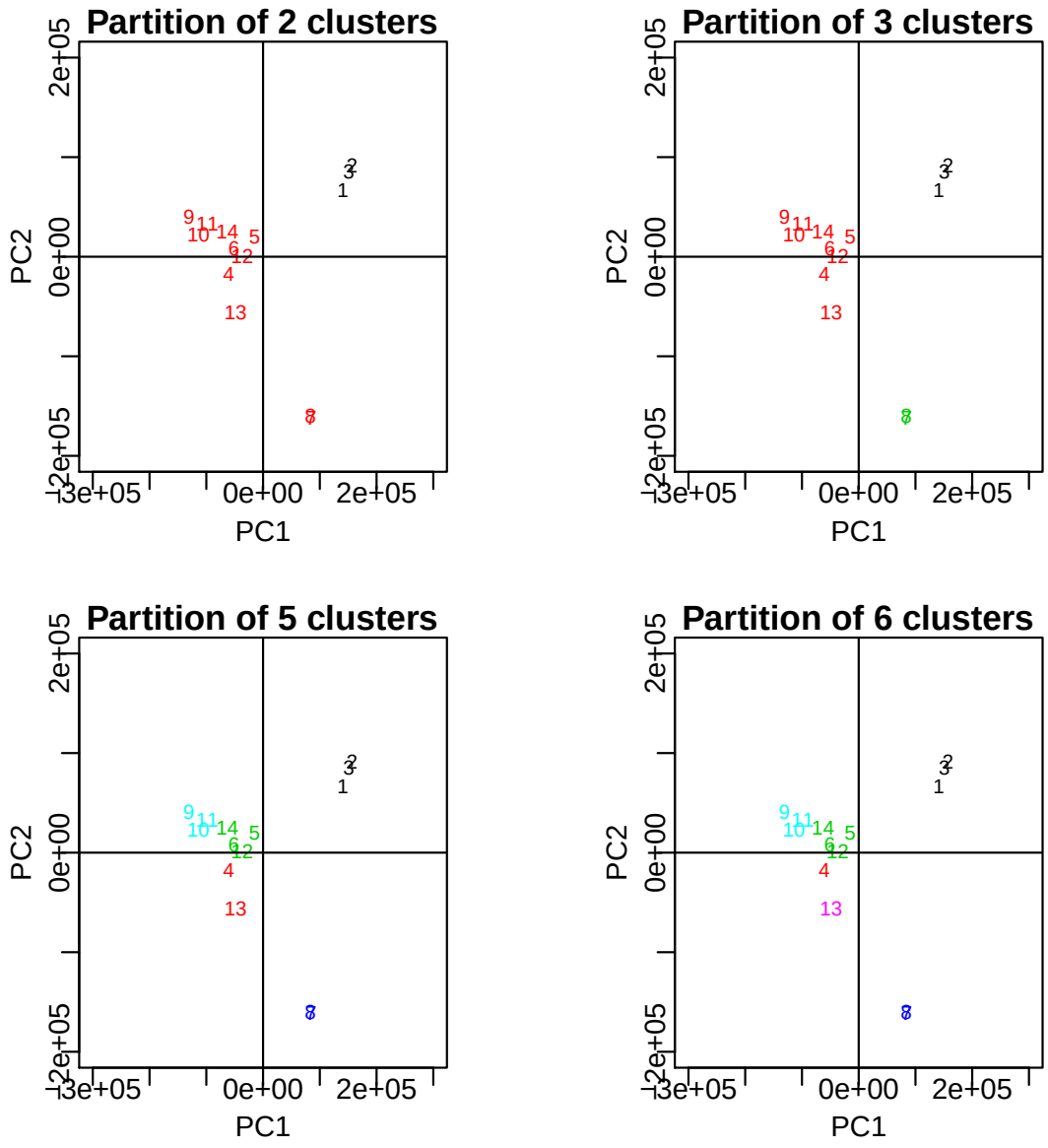
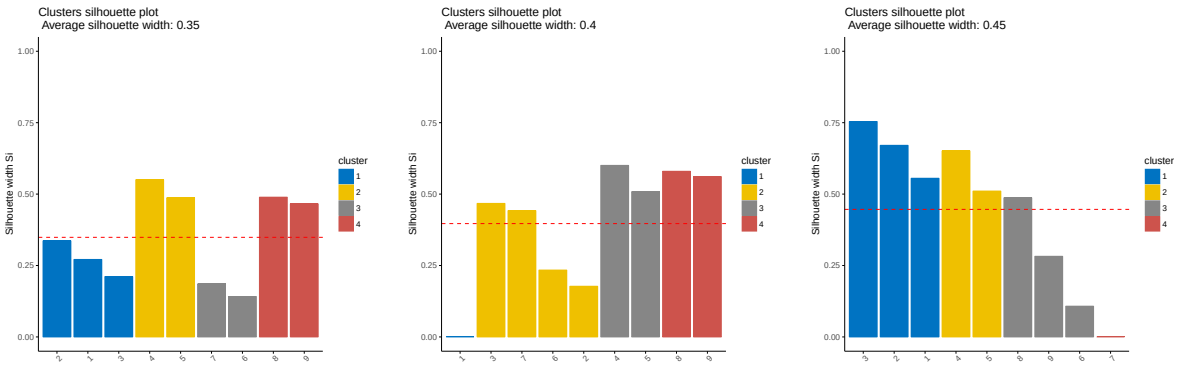


Figure 7.16: Scores plots of the first two PCs, superimposed with the cluster partitions for 2-6 clusters for the Fort William data, derived by the Euclidean - Average clustering method. The labels of the points in the plots correspond to the numbers of samples, and the colours indicate different clusters.

coefficients). The method that has a width of 0.45 (Canberra - Average) is better than the other methods (Canberra - Complete) and (Maximum - Average). In addition, the silhouette plot for the three methods shows that there are no samples



(i) Canberra-Complete method (ii) Maximum-Average method (iii) Canberra-Average method

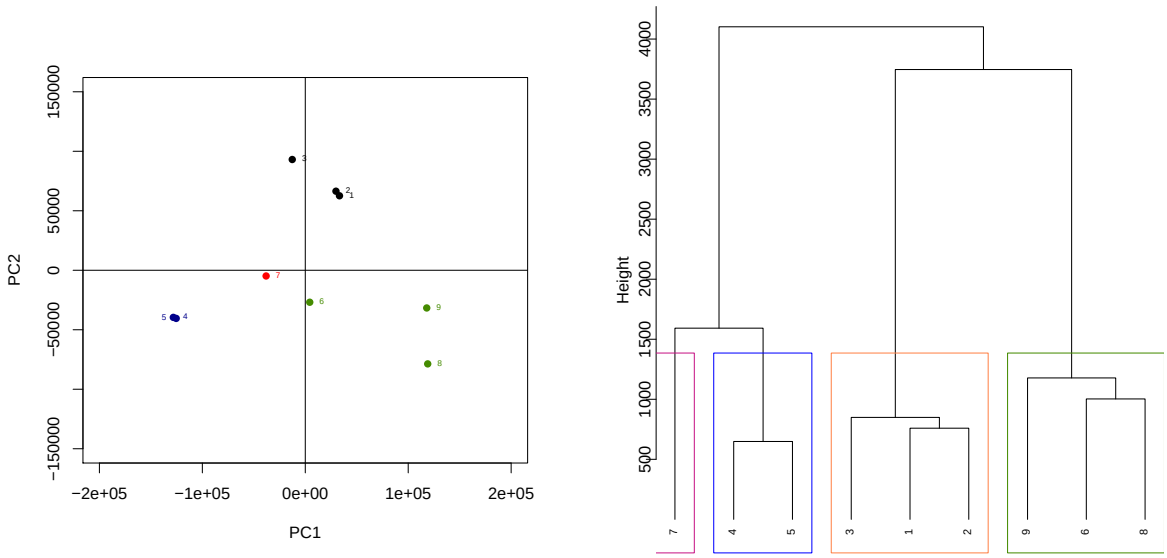
Figure 7.17: Silhouette plot for the 2-cluster partition derived by three clustering methods for the Dunblane data.

misclassified (highly negative Silhouette widths). The findings practically mean that only the Canberra - Average method should be retained for further analyses of the Dunblane data, as it is deemed to be the best.

To illustrate the clustering solution derived by Canberra - Average for the Dunblane data set, a two-dimensional projection of the clustering solution can be seen in Figure 7.18 (i). The first two principal component scores (according to the results from Chapter 5) can be seen superimposed with the partition derived by the Canberra - Average method. From Figure 7.18 (i), it can be seen that the scores plots show clearly that each set of three samples belonging to one hive is not clearly separated from the others, for example, samples 1, 2 and 3 belong to the same hive, and samples 4, 5 and 6 to the same hive and so on. Sample number 7 clusters on its own, but it belongs to the group of samples numbered 7, 8, 9. Also, sample number 6 is separated from its group (numbered 4, 5, 6) and is grouped with samples 8 and 9. In general, the samples from one hive are far from each other.

Since samples 6 and 7 lie towards the centre of the plot, that means that they

have important compounds from the other outlying samples in moderate amounts. Thus it might be proposed that the samples 6 and 7 in the middle of the PCA plot are tending to use several different sources of propolis, whereas the samples towards the periphery of the plot may focus on more restricted sources.



(i) Scores plots (ii) Dendrogram

Figure 7.18: (i) Scores plots of the first two PCs, superimposed with the 4-cluster partition derived by the Canberra - Average clustering method for the Dunblane data. Coloured points represent the samples in the clusters. The labels of the points in the plot correspond to the sample numbers. (ii) Dendrogram for the 4-cluster partition derived by the Canberra - Average linkage clustering method. The labels at the end-leafs of the tree are the sample numbers in the Dunblane data set.

A dendrogram for the clustering partition derived by the Canberra - Average clustering method can be seen in Figure 7.18 (ii). The labels at the end-leafs of the tree correspond to sample numbers. The heat map of the distance matrix re-ordered according to the dendrogram of Figure 7.18 (ii) can be seen in Figure 7.19. Figure 7.18 (ii) shows the most clearly clustered samples, which were obtained from nine of the hives sampled in

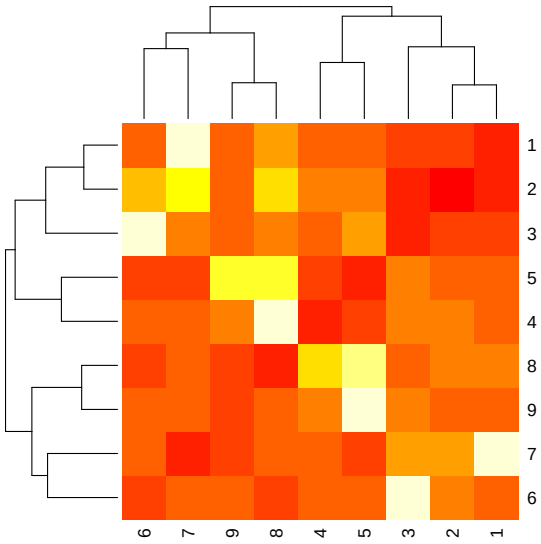


Figure 7.19: Heat map of the distance matrix of the Canberra - Average clustering method according to the dendrogram of Figure 7.18 (ii) for Dunblane. The colour intensity represents the similarity among the samples, such that the darker the colour the closer the similarity.

Dunblane. The samples from Dunblane are different again but are closer in character to the Fort William samples than the Aberdeenshire samples (see Table 5.4). Despite the optimal number of clusters being 4, it might be useful to examine whether partitions of a smaller number of clusters than 4 can provide an insight into the discrimination based on location or compounds. The heat map illustrates the clustering result achieved by application of a clustering method to the data. Figure 7.19 shows the heat map, in which most of the darker values representing high similarities are located closer to the main diagonal. Figure 7.20 illustrates the scores plots for the partition of the Dunblane data. The colours of the points in the scores plots correspond to the clusters. It shows clearly that there is no clear discrimination between samples in the two and three cluster partitions depending on the sets of three samples belonging to a hive.

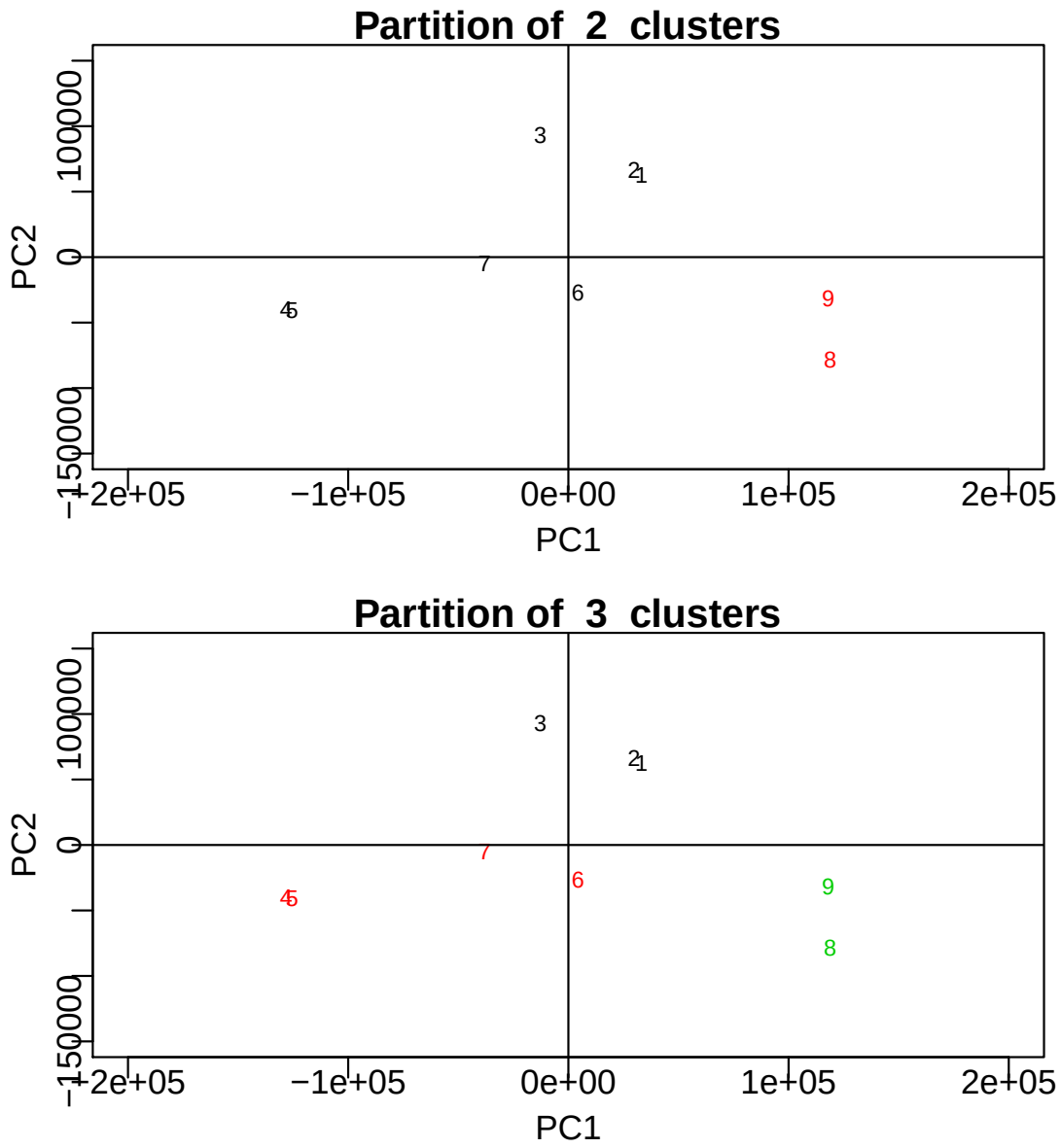


Figure 7.20: Scores plots of the first two PCs, superimposed with the cluster partitions for 2 and 3 clusters for Dunblane data, derived by the Canberra - Average clustering method. The labels of the points in the plots correspond to the sample numbers. The points are colour-coded by the cluster.

7.6.5 Application of HCA to the three Data Sets Combined (IV)

The three data sets (Aberdeenshire, Fort William and Dunblane) are again now combined, as in the analyses of Chapter 5. That is, the data set contains the selected 27,

14 and 9 samples (hives) with 921, 511 and 498 variables respectively. Here I investigate if each data set from the three data sets is separated from the rest depending on the location or chemical compounds. The distance matrix of the samples will be computed using four different distance measures, Euclidean, Manhattan, Maximum and Canberra, for comparison purposes. Five different agglomerative nesting methods will be used in order to perform HCA. These include Single linkage, Complete linkage, Average linkage, Ward's method and the McQuitty method. To facilitate identification of the best clustering method for data set IV, various statistics will be computed and plotting tools will be used to compare the results of the clustering analyses, including the agglomerative coefficient, the Cophenetic correlation, the Gower distance and the silhouette coefficient and plot. In addition, the optimal number of clusters will be identified with the help of plotting tools such as silhouette widths.

The agglomerative coefficient (AC), will be used to assess whether HCA finds natural

Table 7.6: Agglomerative coefficients for data set IV. The red coluor shows the clustering method with the largest agglomerative coefficient.

Metric	Euclidean	Manhattan	Maximum	Canberra
Single	0.753	0.818	0.716	0.745
Complete	0.863	0.898	0.888	0.998
Average	0.846	0.882	0.858	0.932
Ward	0.948	0.962	0.946	0.993
McQuitty	0.841	0.878	0.844	0.937

structure in data set IV or not. Table 7.6 gives the AC values obtained from the analyses of the 20 hierarchical clustering methods described above for data set IV. From Table 7.6, it is clear that the AC has the highest value, 0.998, for the method using the Canberra distance metric and Complete linkage. In general, the Canberra metric seems to give the best results for all the available agglomerative methods, and Ward's method gives the best results for all metrics. The second best method, obtained by the same distance metric

with Ward's linkage, has AC of 0.993.

To confirm the findings, two other statistics will be computed for all 20 methods, i.e. the Cophenetic correlation and the Gower distance (Borcard et al., 2011). Table 7.7 gives the Cophenetic correlation values for all 20 hierarchical clustering methods for data set IV. From Table 7.7 for data set IV, the model Manhattan - Average has the largest Cophenetic

Table 7.7: Pearson's r Cophenetic correlation for the 20 hierarchical clustering methods for data set IV. The blue and red colours show the two clustering methods selected using as criteria the agglomerative coefficient and, Cophenetic correlation respectively.

Metric	Euclidean	Manhattan	Maximum	Canberra
Single	0.897	0.947	0.630	0.109
Complete	0.954	0.957	0.821	0.215
Average	0.962	0.965	0.867	0.216
Ward	0.869	0.895	0.744	0.171
McQuitty	0.959	0.962	0.821	0.164

correlation of 0.965 among all 20 methods. The Cophenetic correlation for this method is shown in blue in Table 7.7. Canberra - Single has the lowest Cophenetic correlation of 0.109 among all 20 methods. Therefore according to this Cophenetic correlation statistic the appropriate method seems to be the model Manhattan - Average. More specifically, Euclidean - Average has the largest Cophenetic correlation among all five Euclidean methods, 0.962, and it is very close to the largest Cophenetic correlation. The Maximum - Average method has the largest Cophenetic correlation among all five Maximum methods, and the Canberra - Average method has the largest Cophenetic correlation among all five Canberra methods.

The Gower distance values for all the clustering methods can be seen in Table 7.8 for data set IV. From Table 7.8, the Gower distance value for the Canberra - Average method is the smallest in comparison to all other methods and in each and every metric. As the Manhattan - Average method had the highest Cophenetic correlation (as seen in Table

Table 7.8: Gower distance for the 20 hierarchical clustering methods for three data. In bold is shown the clustering method with the smallest Gower distance value. The values shown in the table have all been divided by 10^{12} for ease of displaying them.

Metric	Euclidean	Manhattan	Maximum	Canberra
Single	17.0068	3773.3710	4.5606	0.1773
Complete	8.4542	2257.7310	2.7119	40.1239
Average	1.4034	329.0425	0.8111	0.1497
Ward	892.0404	214062.9000	76.8373	3.0204
McQuitty	1.5663	366.0636	1.1385	0.1498

7.7), it seems so far that the best method for the agglomerative coefficient is given by the Canberra - Complete method, with the highest agglomerative coefficient of 0.998 for data set IV, However, further investigation is needed to confirm which of the three methods, Canberra - Complete, Manhattan - Average and Canberra - Average gives the best fit of data set IV.

Before confirming which method gives the best fit of data set IV, we look to identification of the optimal number of clusters using the silhouette widths method.

We use command *NbClust* in *R* to compare between several indices to determine the best number of clusters to use in the analysis for the three clustering methods (Canberra - Complete, Manhattan - Average and Canberra - Average). The results are shown for data set IV in Table 7.9. According to the majority rule in Table 7.9, the best number

data set IV		
Canberra-Complete	Manhattan-Average	Canberra-Average
2 proposed 2 as the best number of clusters	1 proposed 2 as the best number of clusters	2 proposed 2 as the best number of clusters
5 proposed 3 as the best number of clusters	2 proposed 3 as the best number of clusters	4 proposed 3 as the best number of clusters
12 proposed 4 as the best number of clusters	18 proposed 4 as the best number of clusters	11 proposed 4 as the best number of clusters
1 proposed 5 as the best number of clusters	1 proposed 6 as the best number of clusters	3 proposed 6 as the best number of clusters
3 proposed 7 as the best number of clusters	1 proposed 7 as the best number of clusters	1 proposed 7 as the best number of clusters

Table 7.9: The optimal number of clusters of data set IV.

of clusters is 4 for data set IV.

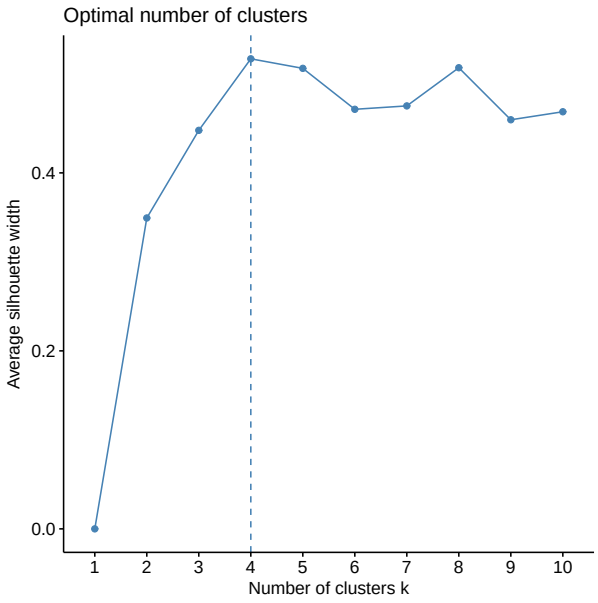
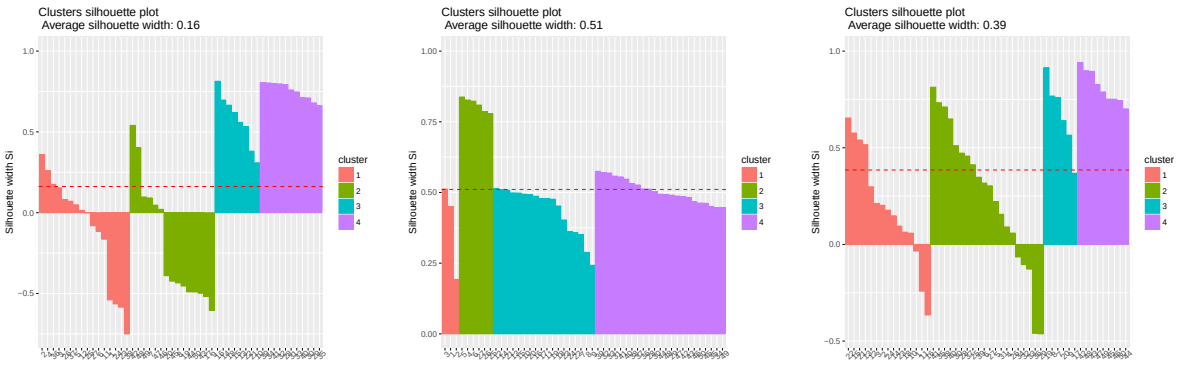


Figure 7.21: Silhouette plot for the 4-cluster partition derived by the Manhattan - Average clustering method for data set IV.

To confirm the result in Table 7.9, Silhouette widths can be used to define a criterion for the appropriate number of clusters. Figure 7.21 illustrates the average silhouette widths for ten partitions, from 2-10 clusters for data set IV, for the Manhattan - Average clustering method. It is clear that for this method the optimal number of clusters is 4 for data set IV. In fact, this is true for all 3 clustering methods.

We will now discuss which result obtained should be retained for further analyses of data IV as follows:

The silhouette plots for the three clustering methods for data set IV can be seen in Figure 7.22 for the Canberra - Complete, Manhattan - Average and Canberra - Average methods. Silhouette width values for all samples can be seen in the silhouette plot as bars. It is therefore clear which samples lie well within their cluster. In fact, the Average silhouette widths for the clusters differ considerably, as the numbers of samples in the four clusters are not the same. The clustering methods have Average Silhouette widths for the entire data set of 0.16, 0.51 and 0.39 respectively, which are also the Silhouette coefficients for



(i) Canberra-Complete method (ii) Manhattan-Average method (iii) Canberra-Average method

Figure 7.22: Silhouette plot for the 4-cluster partition derived by three clustering methods for data set IV.

these methods, therefore there is a difference between them. The method that has value 0.51 (Manhattan - Average) is better than the other methods (Canberra - Complete and Canberra - Average). In addition, the silhouette plot for method Canberra - Complete shows that there are 16 samples clearly misclassified (highly negative silhouette widths) as members of cluster 1 and 2 (7 samples in cluster 1 and 9 samples in cluster 2). Model Canberra - Average is definitely more balanced, with 8 samples misclassified in cluster 1 and 2 and having in general smaller negative silhouette widths than the misclassified samples in method Canberra - Complete. Model Manhattan - Average is definitely the best method, without any misclassified samples.

The findings and the information obtained by the silhouette plots practically mean that only the Manhattan - Average method should be retained for further analyses of data set IV, as it is deemed to be the best. A dendrogram for the clustering partition derived by the Manhattan - Average method clustering method can be seen in Figure 7.23.

To illustrate the clustering solution derived by the Manhattan - Average method for data set IV, a number of graphical tools will be used. A two-dimensional projection of the clustering solution can be seen in Figure 7.24. The first two principal component scores (according to the results from Chapter 5) can be seen superimposed with the par-

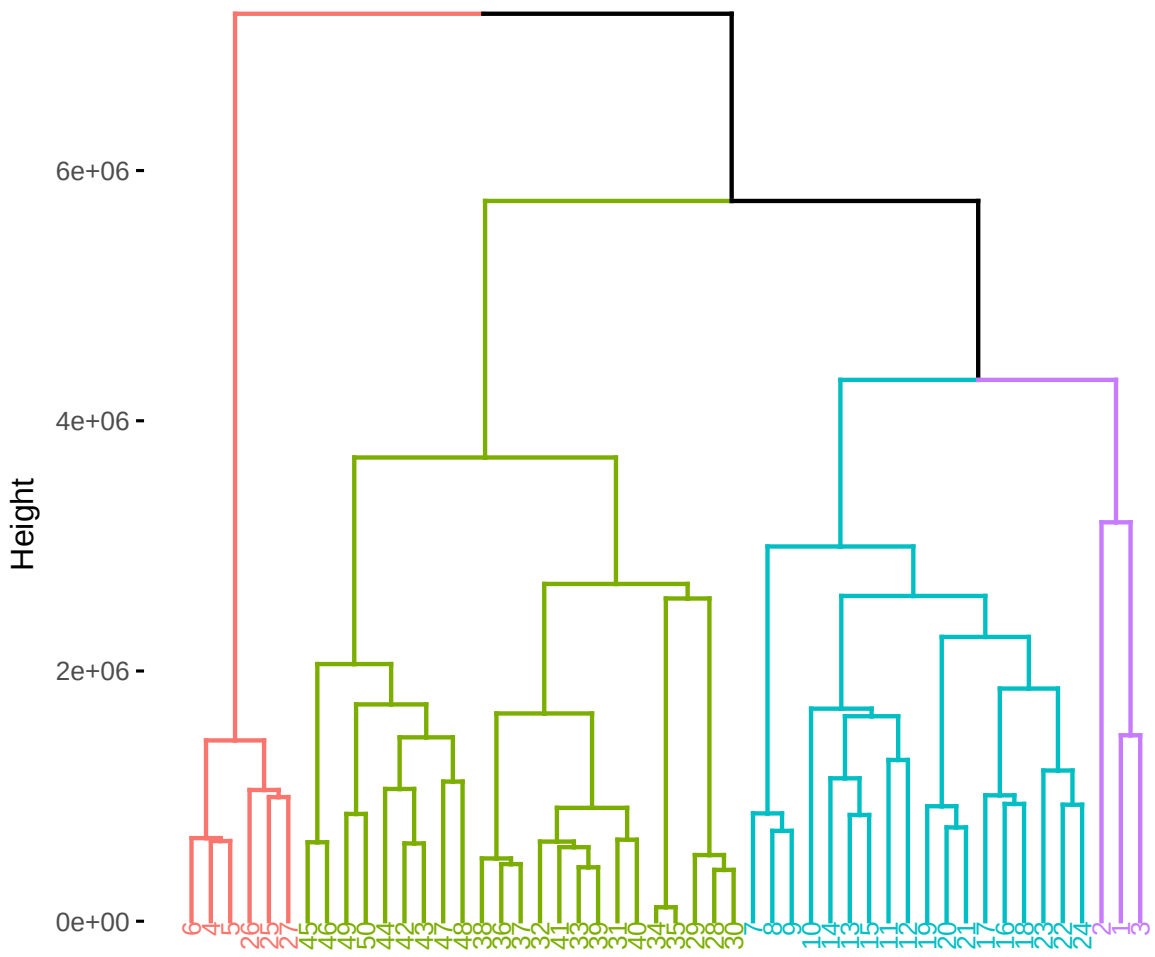


Figure 7.23: Dendrogram for the 4-cluster partition derived by the Manhattan - Average linkage clustering method. The labels at the end-leafs of the tree are the sample numbers in the three data sets. The colour coding shows the different clusters.

tition derived by the Manhattan - Average method. From Figure 7.24, the scores plots show that there is no clear discrimination between the samples of the three data sets based on location. Figure 7.24 shows the 50 samples of propolis, where samples 1 to 27 indicate Aberdeenshire, samples 28 to 41 indicate Fort William, and samples 42 to

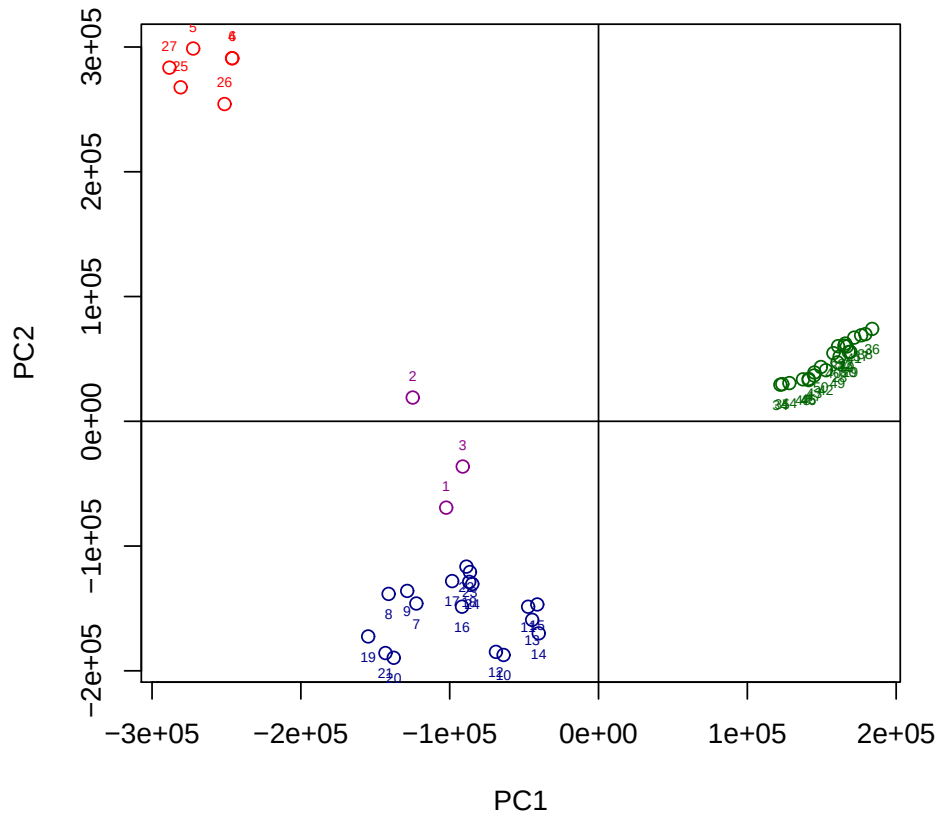


Figure 7.24: Scores plots of the first two PCs, superimposed with the 4-cluster partition derived by the Manhattan - Average clustering method for data set IV. Coloured points represent the samples in the clusters. The labels of the points in the plot correspond to the sample numbers.

50 indicate Dunblane. They could be classified into four groups by hierarchical cluster analysis without splitting the replicates. In this case the samples for Aberdeenshire split into three groups. The first group for Aberdeenshire includes samples 1, 2 and 3 and the second group includes samples 4, 5, 6, 25, 26 and 27, and the third group includes the remaining samples. Moreover, the samples from Fort William and Dunblane cluster together because the compounds of these two data sets have more similarity.

7.6.6 Application of HCA to the Libya Data set

The Libya data set that will now be used in the hierarchical clustering analyses is the same as in the analyses of Chapter 5. The Libya data set contains the selected 12 samples with 300 variables. The distance matrix of the samples will be computed using the same four different distance measures, Euclidean, Manhattan, Maximum and Canberra, as above. Five different agglomerative nesting methods will be used in order to perform HCA, namely Single linkage, Complete linkage, Average linkage, Ward's method and the McQuitty method. The best clustering method for the Libya data set among the 20 mentioned in the previous, will be examined using various statistics and plotting tools, including the agglomerative coefficient, the Cophenetic correlation, the Gower distance and the silhouette coefficient and plot. Also, the optimal number of clusters will be identified with the help of plotting tools such as silhouette widths.

The agglomerative coefficient (AC) will be used to assess whether HCA finds natural structure in the Libya data set or not. Table 7.10 gives the AC values from the analyses of the 20 hierarchical clustering methods above. From Table 7.10, it is clear that the AC

Table 7.10: Agglomerative coefficients for the Libya data. The red coluor shows the clustering method with the largest agglomerative coefficient.

Metric	Euclidean	Manhattan	Maximum	Canberra
Single	0.737	0.739	0.704	0.812
Complete	0.867	0.845	0.885	0.992
Average	0.819	0.809	0.824	0.951
Ward	0.888	0.883	0.895	0.986
McQuitty	0.824	0.808	0.833	0.946

of the Libya data set has the highest value, 0.992, for the result obtained by the Canberra distance metric using Complete linkage. In general, the Canberra metric seems to give the best results for all the tested agglomerative methods, and Ward's linkage gives the

best results for all metrics used. The second best method, obtained by the same distance metric for the Ward's agglomerative method, has an AC of 0.986.

To confirm the findings, the Cophenetic correlation and the Gower distance are computed for all 20 methods. Table 7.11 gives the Cophenetic correlation values for all 20 hierarchical clustering methods for the Libya data set. From Table 7.11, the model Euclidean

Table 7.11: Pearson's r Cophenetic correlation for the 20 hierarchical clustering methods for the Libya data set. The blue and red colours show the two clustering methods selected using as criteria the agglomerative coefficient and Cophenetic correlation respectively.

Metric	Euclidean	Manhattan	Maximum	Canberra
Single	0.769	0.732	0.798	0.341
Complete	0.825	0.772	0.831	0.377
Average	0.843	0.831	0.835	0.387
Ward	0.805	0.765	0.808	0.380
McQuitty	0.835	0.819	0.831	0.386

- Average has the largest Cophenetic correlation of 0.843 among all 20 methods. The Cophenetic correlation value of this method is shown in blue in Table 7.11. Canberra - Single has the lowest Cophenetic correlation of 0.341 among all 20 methods. The clustering method Canberra - Complete indicated by the largest AC value has Cophenetic correlation of 0.377 shown as red in Table 7.11. Therefore according to this Cophenetic correlation statistic the most appropriate method seems to be Euclidean - Average. More specifically, the Manhattan- Average method has the largest Cophenetic correlation among all five Manhattan methods, 0.831. Also, Maximum - Average has the largest Cophenetic correlation among all five Maximum methods, and Canberra - Average has the largest Cophenetic correlation among all five Canberra methods.

The Gower distance values for all the clustering methods can be seen in Table 7.12 for the Libya data set. From Table 7.12, the Gower distance value for the Canberra- Average method is the smallest compared to all other methods and in each and every metric. While

Table 7.12: Gower distance for the 20 hierarchical clustering methods for the Libya data set. In bold is shown the clustering method with the smallest Gower distance value. The values shown in the table have all been divided by 10^{12} for ease of displaying them.

Metric	Euclidean	Manhattan	Maximum	Canberra
Single	2.8378	4.9810	2.2821	9.6567e-09
Complete	4.0078	6.6597	3.7788	1.3545e-07
Average	0.7907	1.4237	0.6664	$7.0439e-09$
Ward	8.4379	17.2718	6.0409	4.2521e-08
McQuitty	0.9045	1.6324	0.7437	7.0798e-09

the Euclidean - Average method had the highest Cophenetic correlation (seen in Table 7.11), the best agglomerative coefficient is given by the Canberra - Complete method, which has the highest AC of 0.992 for the Libya data set, However, further investigation is needed to confirm which of the three methods, Canberra - Complete, Euclidean - Average and Canberra - Average gives the best fit of the Libya data set.

Before confirming which method gives the best fit of the Libya data set, we look to identification of the optimal number of clusters using the silhouette widths method.

We use the command *NbClust* in *R* to compare between several indices to determine the best number of clusters to use in the analysis for three clustering methods (Canberra - Complete, Euclidean - Average and Canberra - Average), with the results shown in Table 7.13. According to the majority rule in Table 7.13, the best number of clusters is 3 for

data set IV		
Canberra - Complete	Manhattan - Average	Canberra - Average
4 proposed 2 as the best number of clusters	2 proposed 2 as the best number of clusters	4 proposed 2 as the best number of clusters
14 proposed 3 as the best number of clusters	8 proposed 3 as the best number of clusters	8 proposed 3 as the best number of clusters
3 proposed 4 as the best number of clusters	5 proposed 4 as the best number of clusters	5 proposed 4 as the best number of clusters
2 proposed 6 as the best number of clusters	5 proposed 5 as the best number of clusters	1 proposed 5 as the best number of clusters
	3 proposed 7 as the best number of clusters	3 proposed 6 as the best number of clusters

Table 7.13: The optimal number of clusters for the Libya data set.

the Libya data set.

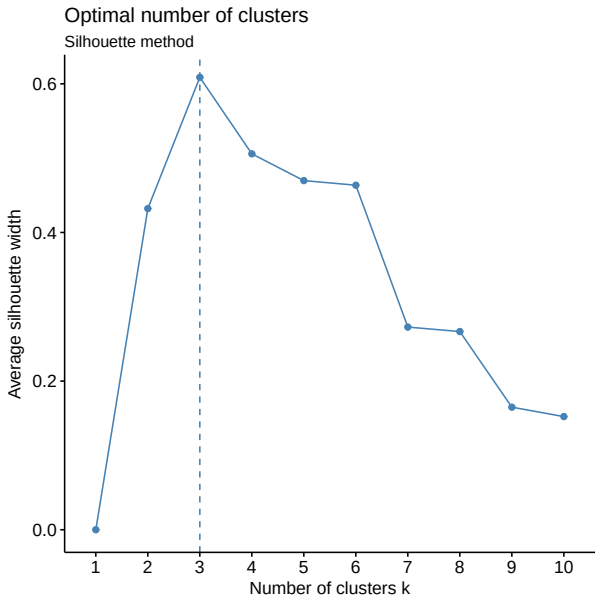
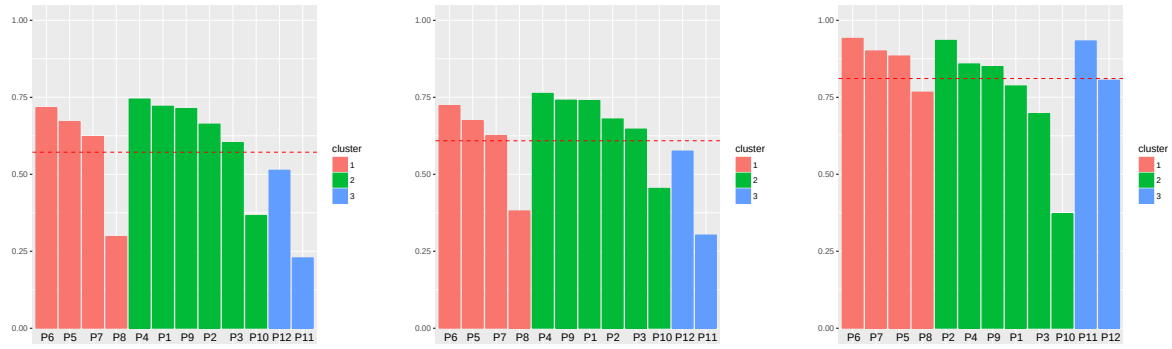


Figure 7.25: Average silhouette widths for partitions of 2-10 clusters for the Libya data set derived by the Euclidean - Average method. The optimal number of clusters is indicated by the dashed line.

To confirm the result in Table 7.13, Figure 7.25 illustrates the average silhouette widths for ten partitions, from 2-10 clusters for this data set, for the Euclidean - Average clustering method. It is clear that for this method the optimal number of clusters is 3 for the Libya data set. In fact, this is true for all 3 clustering methods.

We will now discuss which result obtained should be retained for further analyses of the Libya data as follows:

The silhouette plots for the three clustering methods for the Libya data set can be seen in Figure 7.26 for the Canberra - Complete, Euclidean - Average and Canberra - Average methods. All sample silhouette width values can be seen in the silhouette plot as bars. The Average Silhouette widths for the clusters differ considerably, as the numbers of samples in the three clusters are not the same. The three clustering methods have Average Silhouette width for the entire data set of 0.57, 0.61 and 0.81 respectively, which are also the Silhouette coefficients for the three methods, therefore there is a difference between



(i) Canberra-Complete method (ii) Euclidean-Average method (iii) Canberra-Average method

Figure 7.26: Silhouette plot for the 3-cluster partition derived by three clustering methods for the Libya data set.

them. The method that has a value of 0.81 (Canberra - Average) is better than the other methods (Canberra - Complete and Euclidean - Average). In addition, the silhouette plots for the three methods show that there are not any samples clearly misclassified (negative Silhouette widths) as members of a cluster.

The findings practically mean that only the Canberra - Average method should be retained for further analyses of the Libya data set. A dendrogram for the clustering partition derived by the Canberra - Average method clustering method can be seen in Figure 7.27. To illustrate the clustering solution derived by the Canberra - Average method for the Libya data set, a two-dimensional projection of the clustering solution can be seen in Figure 7.28. In order to get an overview of the differences in the chemical composition of the different propolis samples PCA, was used. This method reduces the 300 variables (chemical compounds) in the samples to a few principal components using the correlations within the Libya data set, essentially mapping the samples according to how close they are in composition. The first two principal component scores (according to the results from Chapter 5) can be seen superimposed with the partition derived by the Canberra - Average method. The data was first mean-centred and Pareto-scaled. Only samples P5, P6 and P7 from the South-east of the country and P8 from the South-west gave a distinct

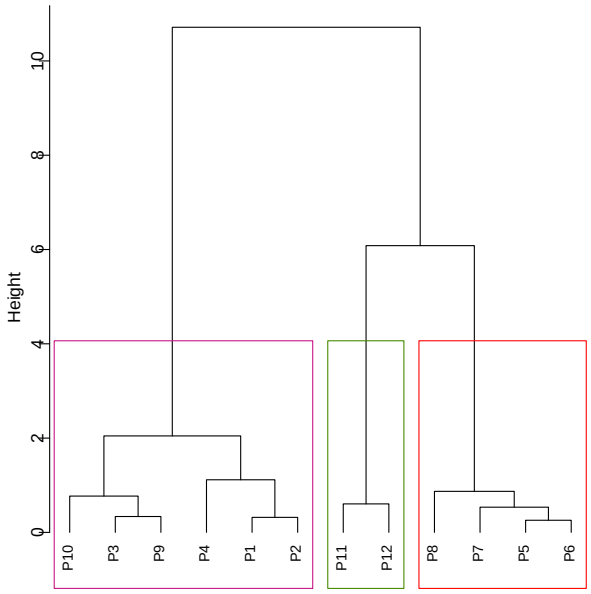


Figure 7.27: Dendrogram for the 3-cluster partition derived by the Canberra - Average linkage clustering method. The labels at the end-leafs of the tree are the names of the samples in Libya data set. The colours show the clusters found.

group and they were grouped. The samples from the coast did not divide according to longitude, and the two groups P1, P2, P3, P4, P9, P10, and the second group, P11, P12, are composed of samples from the East and West of the country, and although P10 was collected from a site close to P11 and P12 it seems to be quite different in composition.

7.7 Summary and Conclusions

This chapter has involved the application of hierarchical clustering algorithms to the propolis data sets. After extensive investigation of the literature, the algorithms deemed to be the most appropriate for metabolomics data included hierarchical clustering, hard clustering methods, and competitive learning algorithms. The last two methods will be applied in Chapters 8 and 9. The main aim was to assess the possible existence of any natural groupings in the data, and consequently identify any patterns of the samples, and

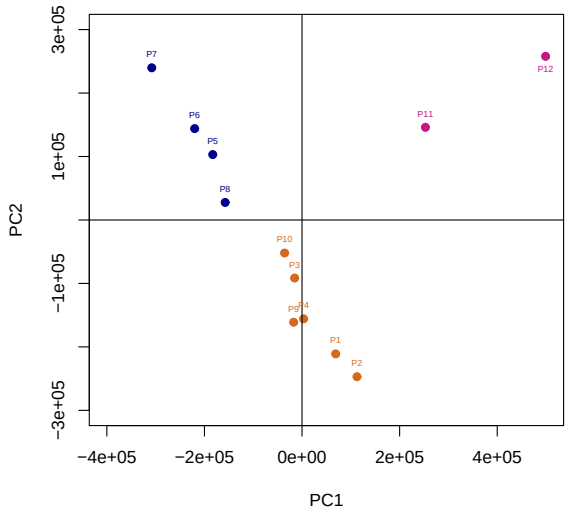


Figure 7.28: Scores plots of the first two PCs, superimposed with the 3-cluster partition derived by the Canberra - Average clustering method for the Libya data set. Coloured points represent the samples in the clusters. The labels of the points in the plot correspond to sample numbers.

in particular any discrimination of samples concerning their location.

Hierarchical methods (HCA) involved the clustering of the data with a range of agglomerative nesting algorithms, single linkage, complete linkage, average linkage, the McQuitty method and Ward's method. These algorithms cover most types of clusters from non-compact elongated (single linkage) to compact spherical clusters (Ward's method). Four different distance metrics were used in the construction of the agglomerative clustering models, namely the Euclidean, Manhattan, Maximum and Canberra distances. Therefore, to improve the chances of HCA identifying any natural groupings, 20 clustering models were constructed and their clustering results were compared.

After extensive experimentation with a range of statistics to assess the quality of fitting of the data by the clustering models (such as the Silhouette width, the agglomerative coefficient and the Cophenetic correlation), the overall best fitting results found were the

2-cluster partition derived by the Euclidean - Average model for data set I, the 4-cluster partition derived by the Euclidean - Average model for data set II, the 4-cluster partition derived by the Canberra - Average model for data set III, the 4-cluster partition derived by the Manhattan - Average model for data set IV, and the 3-cluster partition derived by the Canberra - Average model for the Libya data set. The Silhouette coefficients of 0.60, 0.66, 0.45, 0.51 and 0.81 were the highest among all models for data sets I, II, III, IV and Libya respectively.

Some of the clustering models, with many clusters found, were capable of discriminating the samples according to their hives for data sets I, II and III, also according to their geographical location in data sets IV and Libya.

The next chapter describes in detail the second method of the most important unsupervised classification techniques, for types of data such as metabolomics data, in the area of hard clustering, in a further attempt to devise suitable clustering models for these data. This k-means method is applied in Chapter 8 to identify groupings present in the data sets I, II, III, IV and Libya, to investigate whether it confirms the findings of HCA in Chapter 7.

Chapter 8

Partitioning Algorithms

We now consider a different approach to clustering. The most popular clustering techniques are hierarchical and partitioning methods (Xu and Wunsch, 2005). Clustering methods can be divided into hard clustering and fuzzy clustering. Hard clustering furnishes a partition in which each object of the data set is assigned to one and only one cluster (De Carvalho et al., 2012). HCA is an example of hard clustering. Fuzzy clustering generates a fuzzy partition that furnishes a degree of membership of each pattern in a given cluster. After an investigation of pre-treatment, PCA and MDS, the algorithms deemed to be the most appropriate to use on the propolis data sets included hierarchical clustering, partitioning methods and competitive learning algorithms. The partitioning approach will be applied in this chapter using the k-means method.

The main aim is to assess the possible existence of any natural groupings in the data and consequently to identify any patterns in the samples and in particular any discrimination of samples concerning their location. Moreover, the results will be compared between HCA and k-means. Section 8.2 gives a description of k-means, the most commonly used partitioning algorithm. Section 8.3 describes methods to determine the optimal number of clusters. For this study, k-means has been applied to the different propolis data sets. The outcomes are outlined in Sections 8.4, 8.5 and 8.6.

8.1 Overview of Hard Clustering

In hard clustering, every input vector x_i of a data set in a matrix X ends up matched with a single cluster. A hard m-clustering of data set of matrix X is defined by:

$$m_c : X \to A, c = 1, ..., K$$

where $A = \{0, 1\}$, and K is the number of clusters. A common form of a cost function for hard clustering algorithms is given by:

$$\sum_{i=1}^{N_s} \sum_{j=1}^{K} m_{ij} d(x_i, \vartheta_j) \tag{8.1}$$

subject to the constraints

$$m_{ij} \in \{0, 1\}, i = 1, ..., N_s, j = 1, ..., K$$

$$\sum_{j=1}^{K} m_{ij} = 1,$$

where N_s is the number of samples, m_c is a membership coefficient defining the degree of membership of vector x_i to a cluster c, d is a suitable distance measure between cluster centroids ϑ_j and input vector x_i, $M = \{m_{ij}\}$ is an $(N_s \times K)$ matrix, which signifies the dissimilarity between cluster centroids ϑ_j and input vector x_i using elements $(i, j) = (m_j(x_i), d(x_i, \vartheta_j))$, and

$$m_{ij} = \begin{cases} 1, & \text{if } d(x_i, \vartheta_j) = \min\limits_{1,...N_i} d(x_i, \vartheta_j) \\ 0, & \text{otherwise} \end{cases} . \tag{8.2}$$

According to Theodoridis and Koutroumbas (2003), the importance of calculation (8.1) is minimised when each input vector x_i has been matched with its closest cluster. Ultimately, for every input vector x_i only one m_{ij} is equivalent to 1 and the remaining associated coefficients are equivalent to zero. The most common hard clustering algorithm, k-means is described in detail in the next section.

8.2 The k-means Clustering Algorithm

The k-means hard clustering algorithm is one of the most common and widely applied (Ding and He, 2004; Hartigan and Wong, 1979; Lloyd, 1982). Point representatives (centroids, ϑ) are utilised, and squared Euclidean distance is employed as a way to calculate the distance between the ϑ_j and the input vectors x_i. All resulting clusters remain as compact as possible due to ϑ_j being the mean vector for cluster j. Theodoridis and Koutroumbas (2003) and Izenman (2008) describe k-means in the following manner:

1. Given a series of objects, x_i, $i = 1, 2, ..., N_s$ and if K is the number of clusters, k-means is implemented with the processes below:

   - Arrange the objects into K randomly selected clusters. For every cluster c, calculate the existing point centroid, $\bar{x_c}$.
   - Pre-prescribe K cluster centroids, $\bar{x_c}$, $c = 1, 2, ..., K$ randomly from among the data points or randomly in the data space.

2. Calculate the squared Euclidean distance between every object and its existing point centroid, giving the within cluster sum of squares

$$WSS = \sum_{i=1}^{K} \parallel x_i - \bar{x_c} \parallel^2$$

   where $\bar{x_c}$ denotes the centroid of the cluster containing x_i.

3. Re-allocate every object back to its nearest cluster centroid in order to decrease the size of WSS. Then re-compute the cluster centroids.

4. Repeat Stages 2 and 3 again until there are no more objects to allocate.

The benefits of employing the k-means clustering method are as follows:

- The technique tends to generate more compact clusters than HCA because it attempts to minimise the size of the sum of squares within clusters and maximise the

sum of squares between clusters.

- This technique is computationally faster than HCA if the number of variables is large, as in the study of metabolomics data.

As Myatt (2007) explains, k-means is faster and better at tackling large volumes of observations than hierarchical clustering. However, it also comes with some significant downsides, including:

- k-means cannot produce any kind of hierarchical organisation.
- The clusters cannot be generated before the number of clusters is specified.
- The quality of an optimal clustering may be affected by an outlier.
- Different last stage clusters may be created by different starting partitions. Therefore, using a number of different starting partitions or centroids and then choosing the best final solution is recommended. (The default in R is to use 10 random starts).
- It is not considered the most logical way to cluster data, especially if the cluster shapes are unlikely to be multivariate normal. The method is a better choice for finding compact spherical clusters.

As explained, k-means is a widely used algorithm, particularly among scientists. However, it is not yet broadly employed within the chemometrics field and HCA is much more common there. It is normally used in conjunction with additional visualisation and clustering techniques because there is a distinct lack of diagnostic and specific visualisation resources that are compatible with k-means.

The algorithm created by Hartigan and Wong (1979) will be applied in this research on the data sets I, II, III, IV and Libya. As Hartigan (1975) states, this is a highly efficient version of k-means, referred as algorithm AS 136. The modified algorithm commands a great deal of trust and almost all the authors of *R* functions implementing k-means

currently believe it is a superior choice to alternatives like those in the work of Lloyd (1982) and MacQueen (1967). AS 136 (Hartigan and Wong, 1979) is defined by two key phases. They are Quick Transfer (or QTRAN) and Optimal Transfer (or OPTRA). These make it possible to look for a k-cluster partition with a locally optimal (lowest) sum of squares within clusters. This is achieved by moving objects between clusters. Once all objects have been randomly allocated between the available clusters and a locally optimal solution has been discovered, the whole process may be performed for a pre-determined number of iterations (normally 100) (Legendre and Legendre, 1998). It is important to begin every run from a randomly selected configuration. If these steps are followed, it increases the probability of finding a global minimum. The ideal solution is any iteration with the lowest sum of squares within clusters across every run.

The work of Hartigan and Wong (1979) describes the phases of algorithm AS 136 as follows (Lithio and Maitra, 2018):

Initialising: The data matrix X of x_i, $i = 1, ..., N_s$ objects is given an initial centroid conformation in p-dimensional space, where p is the number of variables. This may refer to the set of initial centroids corresponding to the necessary number of clusters or just the number of clusters required, K. For the former, a random set of rows in X is selected to serve as the initial centroids, for example, $\bar{C}_j$, $j = 1, ..., K$. The number of points within a single cluster, say c, is denoted by NS_c. To find the points in each cluster, the Euclidean distance is used between the object x_i and the cluster centroid $\bar{C}_j$, denoted by $d(x_i, \bar{C}_j)$.

Phase 1: For every object x_i, $i = 1, ..., N_s$, find its nearest and second nearest centroids, which are $\bar{C1}_i$ and $\bar{C2}_i$ respectively. Match up object x_i with Cluster C1.

Phase 2: Update the centroids to be the averages of the points assigned to them above.

Phase 3: Initially, all clusters should be considered part of the live set.

Phase 4 - OPTRA: Take every object x_i in order and assign it to the live set only if cluster C is updated in the final QTRAN phase, then it belongs to live set in this phase. Otherwise, the object should not be changed in the final N_s optimal transfer phases or treated as part of the live set. Assign object x_i to cluster C_j. If C_j is not live set, go to Phase 4B. If it is part of the live set, go to Phase 4A.

Phase 4A: Calculate the minimum of the quantity

$$R2 = \frac{NS_c d(x_i, C)^2}{NS_c + 1}$$

for every Cluster C $(C = C_j, j = 1, ..., K)$. Cluster C_l denotes the one with the lowest value of R2. If the value is equal to or larger than

$$\frac{NS_{C_j} d(x_i, C_j)^2}{NS_{C_j} - 1},$$

no re-allocation is required and, C_l becomes the updated $\bar{C2}_i$. If this is not the case, object x_i must be allotted to cluster C_l and C_j becomes the updated $\bar{C2}_i$. The centroids are updated to become the means of the objects allotted to them if re-allocation has occurred. At this point, the two clusters linked to the transfer of object x_i are now in the live set.

Phase 4B: The same as Phase 4A, with the difference that the minimum R2 is calculated only for clusters in the live set.

Phase 5: If the live set is empty, do not proceed with the next phase. If it is not empty, start Phase 6 after one pass through the data set.

Phase 6-QTRAN: In order, let C_j=$\bar{C1}_i$ and C_l=$\bar{C2}_i$ for every object x_i. Calculate the values

$$R1 = \frac{NS_{C_j} d(x_i, C_j)^2}{NS_{C_j} - 1}$$

and

$$R2 = \frac{NS_{C_l} d(x_i, C_l)^2}{NS_{C_l} + 1}.$$

If R2 is greater than R1, cluster C_j still carries object x_i. If this is not the case, swap $\bar{C1}_i$ and $\bar{C2}_i$ around. Then, renew the central points for both clusters.

Phase 7: Return to phase 4 if no re-allocations were made within the last N_s phases. If this is not the case, return to phase 6.

8.3 Identifying the Optimal Number of Clusters

In hard clustering, identifying the optimal number of clusters is a vital process. Certainly, when using k-means clustering, it is necessary to state beforehand the number of clusters K that will be found. The problem is that there is no definite rule to use to choose K. The optimal number of clusters is, to some degree, subjective. It is affected by the constraints associated with partitioning, and the measure chosen for calculating similarities (through this is generally Euclidean distance).
In this section, we will describe different methods for determining the optimal number of clusters for k-means. These include direct techniques and statistical testing, as follows (Kassambara, 2017):

- Direct Techniques: The goal here is to choose K to give a suitable value of a criterion, for instance, the average silhouette width or the within cluster sums of squares. These are called the silhouette and elbow methods, respectively.
- Statistical Techniques: The aim here is to contrast data evidence for clustering with the null hypothesis of there being no clustering in the data. This is called the Gap statistic method.

For this study, the first of these kinds of techniques is applied, using for example the average silhouette width and the elbow methods. Using several criteria allows a more accurate estimation of the optimal number of clusters. Figure 8.1 shows a plot used to identify the number of clusters, by plotting the criterion value against K. There are

over twenty techniques and indices that can be used to determine the optimal number of clusters (as in the command *NbClust* in *R*, for example).

8.3.1 The Elbow Technique

As explained, the simple premise behind partitioning techniques such as k-means clustering is to form the clusters in a way that minimises intra-cluster variation or total within-cluster sum of squares (WSS). Crucially, the elbow technique treats the total WSS (or other suitable measure) as a function of the number of clusters. Therefore, it is important to select the number of clusters to give a low total WSS, yet not give a larger than necessary number of clusters. The optimal number of clusters is determined using the following process (Kassambara, 2017):

1. Decide on the clustering algorithm. In this case, it is k-means, for various values of K (such as 1-10 clusters). With every K, apply the algorithm and calculate the total within cluster sum of squares (WSS).

2. Using the number of clusters K, plot WSS against K, to form a curve.

3. If there is a bend or elbow within the plot, its position may be considered as a signifier of the most suitable number of clusters (see the plots in Figure 8.1).

8.3.2 The Average Silhouette Technique

This method is also used to calculate clustering quality. According to Kaufman and Rousseeuw (2009), it determines the typical silhouette for possible and varied values of K. Specifically, the technique identifies how successfully each object is positioned within its corresponding cluster. A high value of the silhouette width (see Section 7.5) implies a satisfactory degree of clustering (Kaufman and Rousseeuw, 2009). The optimal number of clusters K is the value that most increases the typical silhouette across a selection

of potential values of K. The Average Silhouette technique is comparable to the elbow technique. It is applied in the following way:

1. Choose the clustering algorithm, in this case, k-means, and apply it for various values of K (such as 1-10 clusters).

2. Determine the average silhouette width for each K value.

3. Using the number of clusters K, plot average silhouette width against K to form a curve.

4. The optimal number of clusters is equivalent to the position of the maximum value (see the plots in Figure 8.2).

8.4 Application of the k-means Algorithm for the Data Sets I, II and III

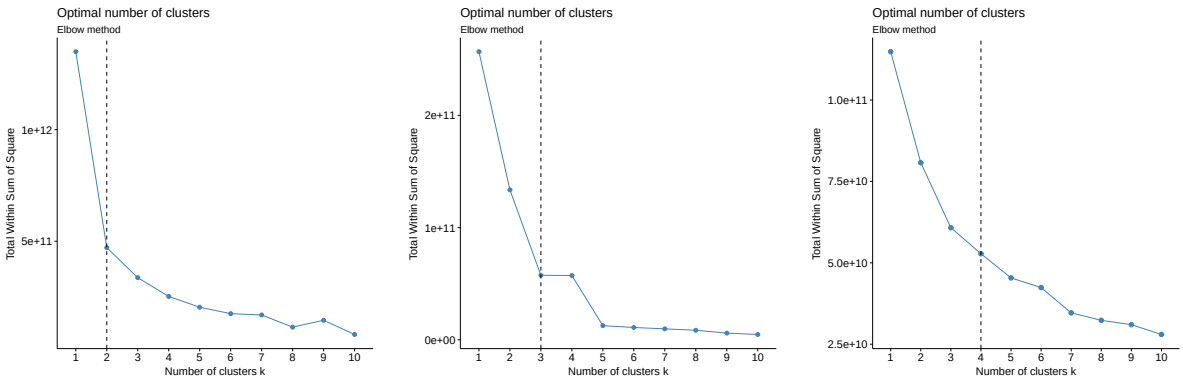
We now apply the methods above to the propolis data sets.

8.4.1 Overview

The data sets I, II and III (used in the HCA analyses) will now be analysed by k-means clustering. Data sets I, II and III include 27, 14 and 9 samples respectively (for Aberdeenshire, Fort William and Dunblane), and have been mean-centred and column-scaled by Pareto scaling before analysis.

8.4.2 Computing the Optimal Number of Clusters for Data Sets I, II and III

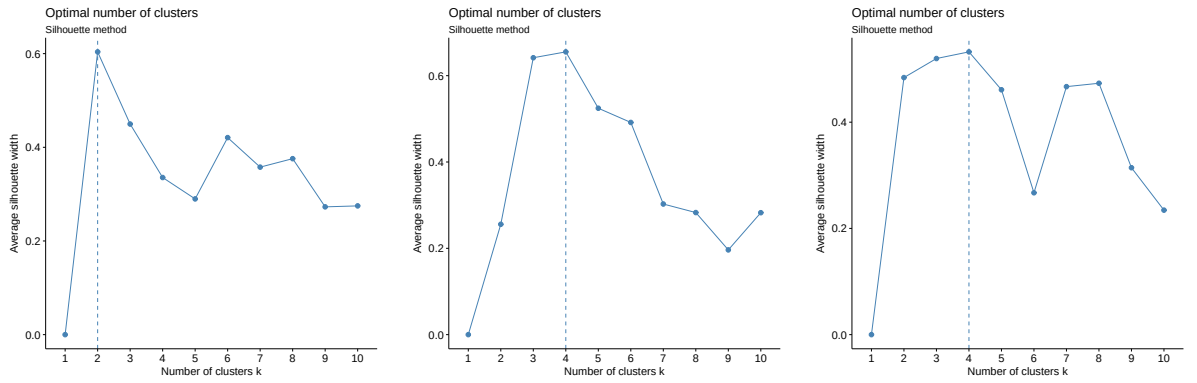
The algorithms described in Section 8.3 will be used to provide the number of clusters for k-means. To determine the optimal number of clusters, two different techniques will be used, to evaluate the derived k-means partitions of 2-10 clusters. First the elbow method is used. Figure 8.1 shows the results for this method for data sets I, II and III. The dashed line shows the point which corresponds to the optimum number of clusters. The optimum solutions marked by the *R* software are 2 clusters for the Aberdeenshire data, 3 clusters for the Fort William data and 4 clusters for the Dunblane data. To compare to the re-



(i) Aberdeenshire (ii) Fort William (iii) Dunblane

Figure 8.1: Values of the within cluster sum of squares for k-means partitions of 2-10 clusters for data sets I, II and III. The dashed line represents the optimum number of clusters.

sults of the elbow method, the silhouette values for the k-means partitions of clusters are also computed for data sets I, II and III. Figure 8.2 gives the overall average silhouette widths of the clusters. The dashed line shows the optimal number of clusters determined in *R*. From the silhouette information and the elbow method values, it can be concluded that the optimum numbers of clusters are 2 for Aberdeenshire, 4 for Fort William and 4 for Dunblane. There is a difference between the results for Fort William, where the opti-



(i) Silhouette of Aberdeenshire (ii) Silhouette of Fort William (iii) Silhouette of Dunblane

Figure 8.2: Average silhouette widths of clusters for k-means clustering method for data sets I, II and III. The optimal number of clusters is indicated by the dashed line.

mum numbers of clusters are 3 from the Elbow method and 4 from the silhouette method.

We will also now use the command *NbClust* in *R* to compare between many indices to determine which value of K to use in the analysis. The index to be calculated will be one of the following: "kl", "ch", "hartigan", "ccc", "scott", "marriot", "trcovw", "tracew", "friedman", "rubin", "cindex", "db", "silhouette", "duda", "pseudot2", "beale", "ratkowsky", "ball", "ptbiserial", "gap", "frey", "mcclain", "gamma", "gplus", "tau", "dunn", "hubert", "sdindex", "dindex", "sdbw", "all" (all indices except GAP, Gamma, Gplus and Tau), and "alllong" (all indices with Gap, Gamma, Gplus and Tau included) (Charrad et al., 2014). Applying this command gave results for data sets I, II and III as follows in Table 8.1. According to the majority rule, the best number of clusters are 2 for the Aberdeenshire data, 4 for the Fort William data and 4 for the Dunblane data.

It remains to be seen whether this clustering method can discriminate the samples depending on their location, where each three samples came from the same hive (or colony). The clusters are examined in more detail below.

Among all indices for Aberdeenshire
8 proposed 2 as the best number of clusters
7 proposed 3 as the best number of clusters
1 proposed 4 as the best number of clusters
2 proposed 5 as the best number of clusters
3 proposed 6 as the best number of clusters
2 proposed 14 as the best number of clusters
Among all indices for Fort William
3 proposed 2 as the best number of clusters
1 proposed 3 as the best number of clusters
12 proposed 4 as the best number of clusters
4 proposed 5 as the best number of clusters
2 proposed 6 as the best number of clusters
1 proposed 9 as the best number of clusters
Among all indices for Dunblane
3 proposed 2 as the best number of clusters
7 proposed 3 as the best number of clusters
8 proposed 4 as the best number of clusters
5 proposed 5 as the best number of clusters

Table 8.1: The optimal number of clusters for data sets I, II and III, using 23 different criteria.

8.4.3 Cluster Validation

As mentioned previously in Section 8.3.2, the silhouette coefficient measures how well an observation is clustered and it estimates the average distance between clusters. The silhouette plot displays a measure of how close each point in one cluster is to points in

the neighbouring clusters, and this method is now used to confirm the results of k-means for data sets I, II and III:

- **Aberdeenshire data (I)**

  A silhouette plot for the cluster partition can be seen in Figure 8.3 for the Aberdeenshire data. The average silhouette width values for clusters 1 and 2 are 0.75 and 0.56, respectively, and the average silhouette width for the entire data set is 0.60. Although these values are the same as from HCA clustering (see Figure 7.9 for the Euclidean-Average method, which was the best method for HCA), in the k-means clustering solution there are no misclassified samples (Figure 8.3). Samples 1, 2 and 3 have low silhouette values in cluster 2, whereas there are not any samples with similarly low values in cluster 1. The silhouette plot also shows that the first and second clusters contain 6 and 21 samples respectively, exactly as in the HCA clustering case. Also, k-means is as good as the HCA method is in fitting the data, as the HCA method's overall average silhouette width of 0.60 (for Euclidean-Average method) is the same as for the k-means clustering method (0.60). Comparing k-means with the HCA clustering method, the samples in the k-means partition fit as well, as cluster 2's silhouette widths are 0.60 and 0.60 for k-means and HCA respectively. In addition, the plot confirms that there is no discrimination of the samples 1, 2 and 3 with low silhouette values.

  The derived optimal 2-cluster k-means partition is the same as was obtained by the optimal HCA clustering partition, therefore the results of any extra analysis will be the same as in the HCA clustering case. Figure 8.4 also illustrates the clustering solution derived by the 2-cluster k-means method, that is, a two-dimensional projection of the clusters such that the first two principal component scores (according to the results from Chapter 5) can be seen according to the partition from the selected k-means clustering method. In the scores plot, blue and red represent the samples

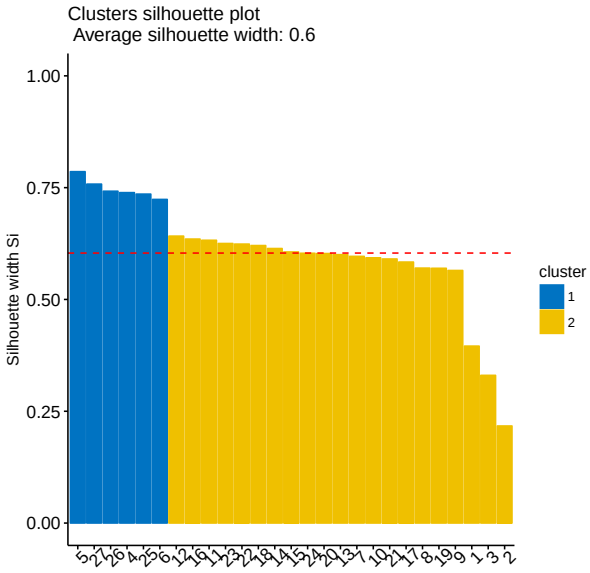


Figure 8.3: Silhouette plot for the 2-cluster partition from the k-means clustering method for the Aberdeenshire data. The x-axis shows the sample numbers. The average silhouette widths for clusters 1 and 2 are 0.75 and 0.56 respectively, and the average silhouette width for the entire data set is 0.60 (shown by the dashed red line).

assigned to the first and second cluster respectively. Similarly to HCA clustering, the cluster is compact for group 1 (Figure 8.4). Also, there is a distinction among samples in this clustering. Therefore, as in the HCA methods, this algorithm has classified samples according to their hives and samples from the same hive are not separated, but it has only separated samples 4, 5, 6, 25, 26 and 27 from the rest.

- **Fort William data (II)**

A silhouette plot for the cluster partition is shown in Figure 8.5 for the Fort William data. The average silhouette width values for clusters 1, 2, 3 and 4 are 0.52, 0.76, 0.91 and 0.64 respectively, and the average silhouette width for the entire data set is 0.66. Although these values are the same as from HCA clustering (Figure

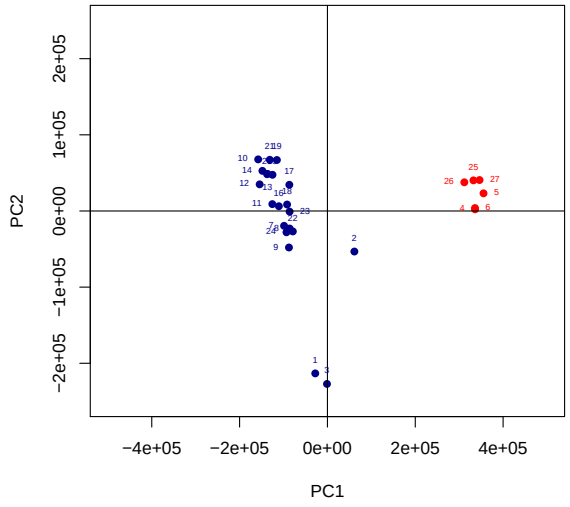


Figure 8.4: Scores plot of the first two PCs, superimposed with the 2-cluster partition from k-means clustering for the Aberdeenshire data. Blue and red points represent the samples in the first and second cluster. The labels of the points in the plot correspond to the numbers of the samples.

7.13 for the Euclidean-Average method, which was the best method for HCA), in the k-means clustering solution there are no misclassified samples (Figure 8.5). The silhouette plot also shows that the clusters contain 6, 3, 2 and 3 samples respectively, exactly the same points as in the HCA clustering case, and k-means is as good as the best HCA method is in fitting the data, as the HCA method's overall average silhouette width of 0.66 (for the Euclidean-Average method) is the same to that of the k-means clustering method (0.66). Moreover, k-means is better than the two methods Canberra-Complete and Canberra-Average for HCA, as these have silhouette widths of 0.33 and 0.43 respectively (see Figure 7.13). Comparing the results of k-means with the best HCA clustering method, the samples in each case are fitted as well, since the average silhouette widths are 0.66 for the both HCA and k-means. The derived optimal 4-cluster k-means partition is the same as was

obtained by the optimal HCA clustering, therefore the results of any extra analysis will be the same as in the HCA clustering case.

Figure 8.6 shows the clustering solution from the 4-cluster k-means, in terms of the

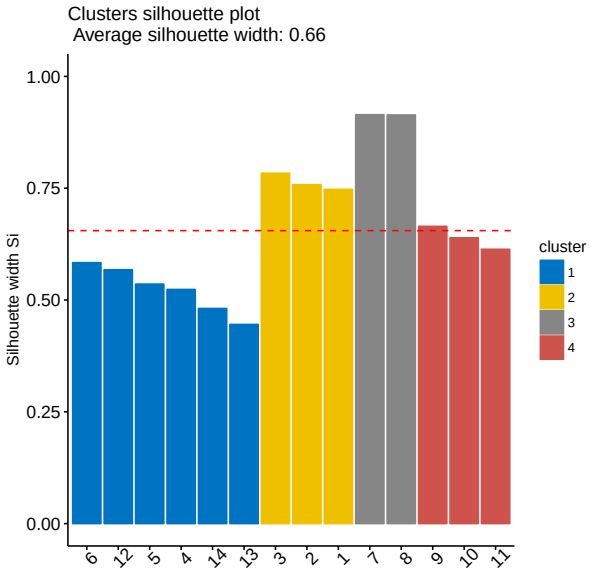


Figure 8.5: Silhouette plot for the 4-cluster partition from the k-means clustering of the Fort William data. The x-axis shows the sample numbers. The average silhouette widths for clusters 1, 2, 3 and 4 are 0.52, 0.76, 0.91 and 0.64 respectively, and the average silhouette width for the entire data set is 0.66 (shown by the dashed red line).

first two principal component scores. In the scores plot, blue, green, brown and red represent the samples assigned to the four clusters. Similarly to HCA clustering, all of the clusters are compact. Also, there is a distinction among locations in this clustering but there is no separation of samples from the same hive. (Note that samples 7 and 8 in this data set came from the same hive, but sample 9 cames a different hive). Therefore, as in HCA, this algorithm has been efficient in classifying samples according to their location (their colony).

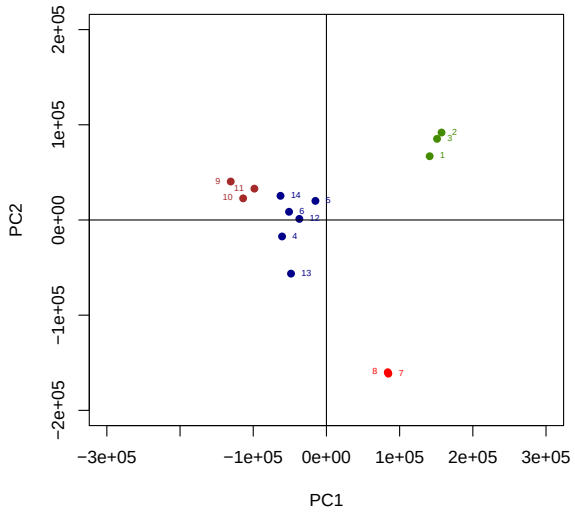


Figure 8.6: Scores plot of the first two PCs, superimposed with the 4-cluster partition from k-means clustering for the Fort William data. Coloured points represent the different clusters. The labels of the points in the plot correspond to the numbers of the samples.

- **Dunblane data (III)**

A silhouette plot for the cluster partition can be seen in Figure 8.7 for the Dunblane data. The average silhouette width values for clusters 1, 2, 3 and 4 are 0.27, 0.52, 0.48 and 0.16 respectively, and the average silhouette width for the entire data set is 0.35. The silhouette plot also shows that the four clusters contain 3, 2, 2 and 2 samples respectively. This solution is not quite the same as was obtained from HCA clustering (Figure 7.17). From Figure 8.7, there are no misclassified samples in the k-means clustering solution. Samples 6 and 7 have very low silhouette values in cluster 4. Comparing the k-means results with the HCA clustering, the samples in the HCA clustering are slightly better fitted as the average silhouette widths are 0.45 and 0.35 for HCA and k-means respectively. Also, there are different numbers of samples in each cluster for k-means and HCA, such as samples 6 and 7 belonging

to cluster number 4 in k-means clustering while in HCA only sample 7 belongs to cluster 4. Figure 8.8 illustrates the clustering solution derived by the 4-cluster k-

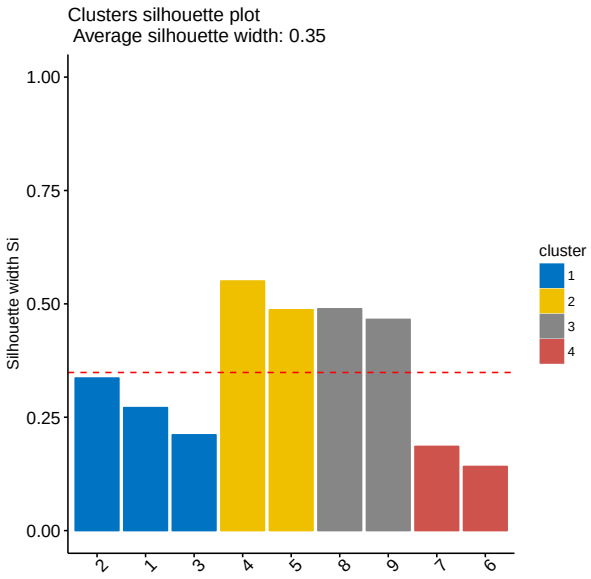


Figure 8.7: Silhouette plot for the 4-cluster partition derived by the k-means clustering method of Dunblane. The average silhouette width for clusters 1 to 4 are 0.27, 0.52, 0.48 and 0.16 respectively, and the average silhouette width for the entire data set is 0.35.

means method, in terms of the first two principal component scores. In the scores plot, the different colours represent the clusters. Samples 1, 2 and 3 are kept together but the rest are not. There is a separation of some of the samples that belong to the same hive. For instance, samples 4, 5 and 6 are from the same hive but they are separated in the clustering solution.

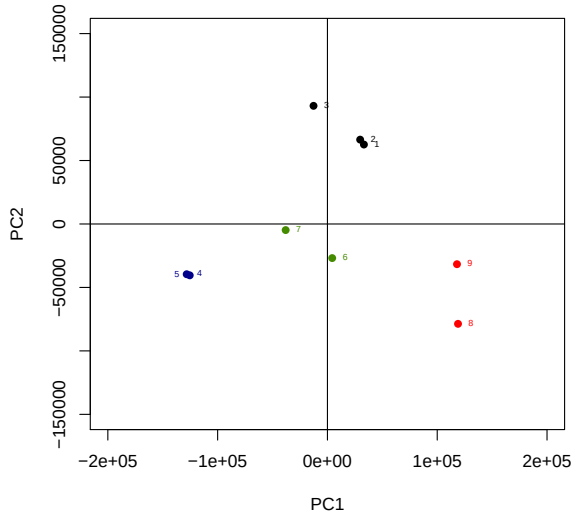
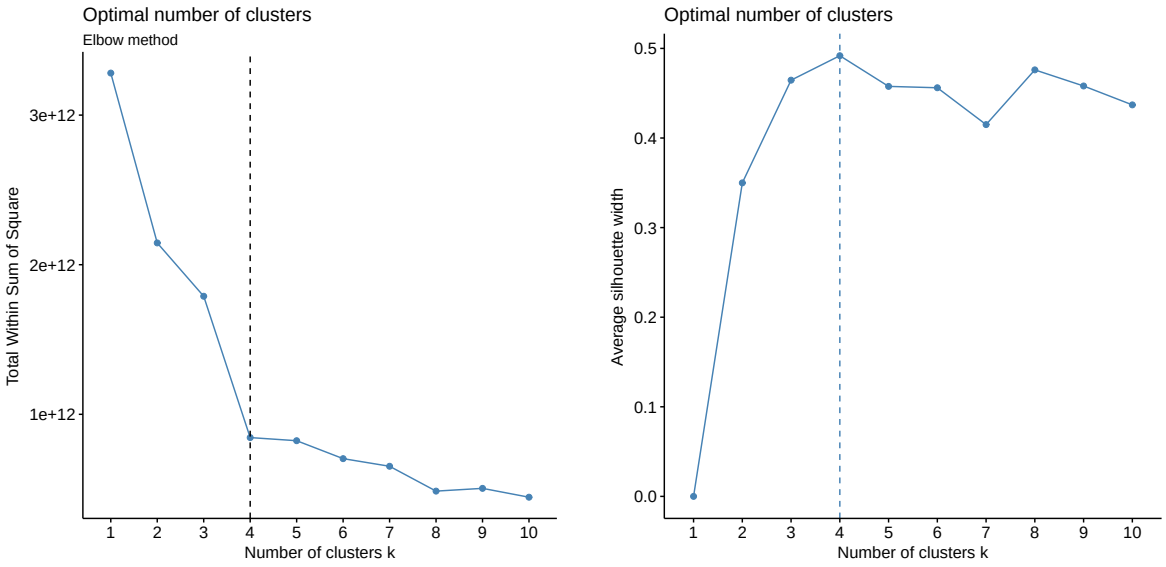


Figure 8.8: Scores plot of the first two PCs, superimposed with the 4-cluster partition from k-means for the Dunblane data. Colours represent the clusters in the four clusters. The labels of the points in the plot correspond to the numbers of the samples.

8.5 Application of the k-means Algorithm to the three Data Sets Combined (IV)

The three data sets (Aberdeenshire, Fort William and Dunblane) are now considered together as one data set, using k-means clustering, as was done in the HCA analyses. That is, data set IV contains the selected 27, 14 and 9 samples with 921, 511 and 498 variables respectively. The three data sets have been mean-centred and column-scaled by Pareto scaling before analysis. To determine the optimal number of clusters, Figure 8.9 (left plot) shows the results for the Elbow method on data set IV. The dashed line shows the point which corresponds to the optimum number of clusters. The optimum solution is 4 clusters for data set IV. To confirm the result of the Elbow method, the average silhouette value for the k-means partitions of clusters will be computed for data set IV. Figure 8.9

(right plot) gives the overall average silhouette widths of the clusters. The dashed line shows the optimal number of clusters. From the average silhouette information and the Elbow method, it can be concluded that the optimum number of clusters is 4 for data set IV. A silhouette plot for the cluster partition can be seen in Figure 8.10 for data set IV.



(i) The Elbow method (ii) Average silhouette method

Figure 8.9: K-means partitions of 2-10 clusters for data set IV. The dashed line represents the optimum number of clusters.

The average silhouette width values for clusters 1, 2, 3 and 4 are 0.70, 0.39, 0.45 and 0.42 respectively, and the average silhouette width for the entire data set is 0.49. From Figure 8.10 for k-means and Figure 7.22 for the Manhattan-Average method for HCA, there are no misclassified samples in both methods, but the silhouette plot also shows that the four clusters from k-means contain 6, 14, 21 and 9 samples respectively, which is completely different from the HCA clustering. However, the k-means clustering method does fit the data IV almost as well as HCA. It is approximately as good as the HCA method is in fitting data set IV, as the HCA method's overall average silhouette width of 0.51 is quite close to that of the k-means clustering method (0.49).

A two-dimensional projection of the clustering solution can be seen in Figure 8.11 in

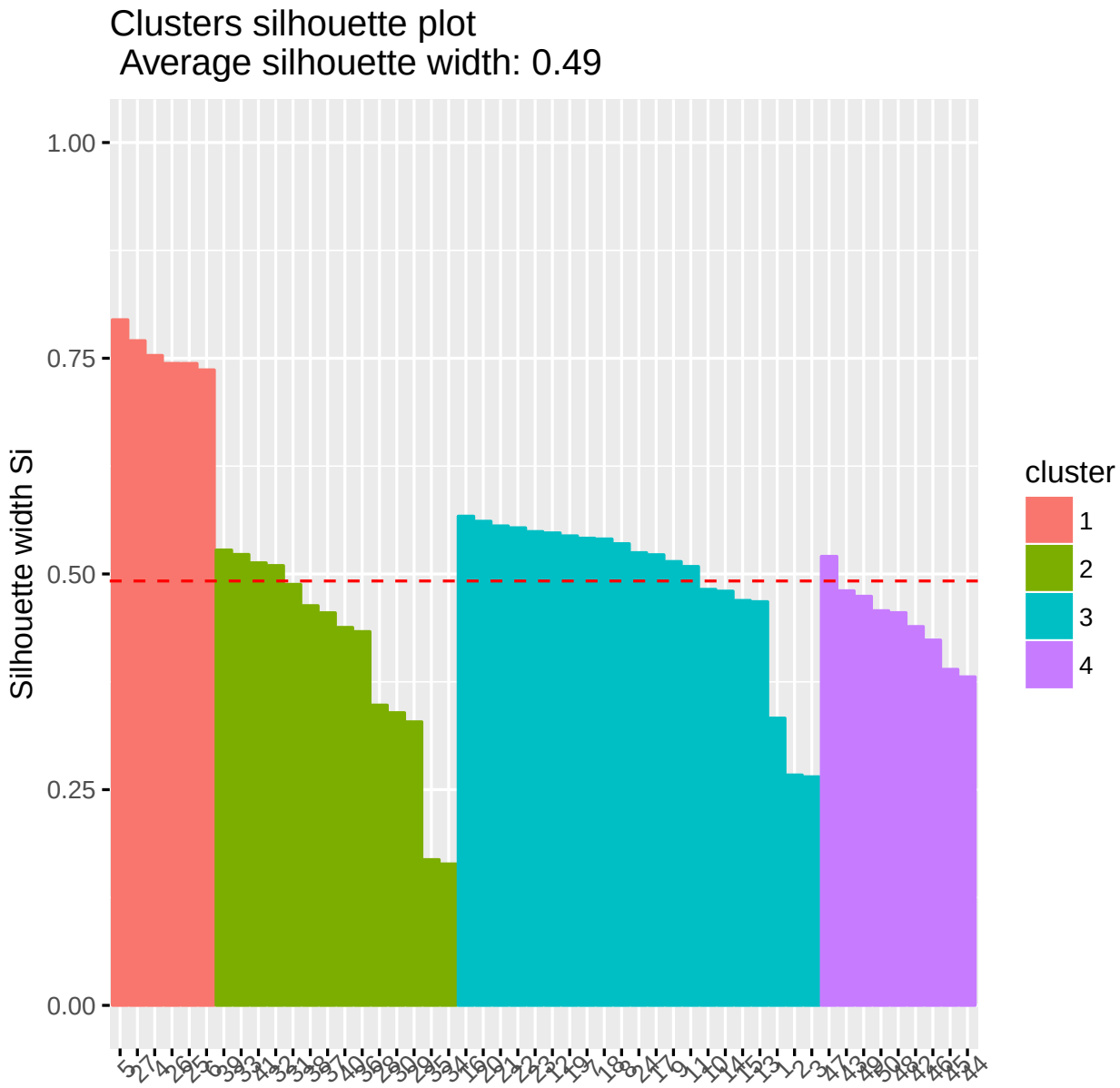


Figure 8.10: Silhouette plot for the 4-cluster partition derived by the k-means clustering method on data set IV. The average silhouette width for the four clusters are 0.70, 0.39, 0.45 and 0.42 respectively, and the average silhouette width for the entire data set is 0.49.

principal component space. In the scores plot, the colours represent the different clusters of points and the points are labelled with the sample numbers. Figure 8.11 shows the three data sets I, II and III divided from each other. Also, it separates the Aberdeenshire

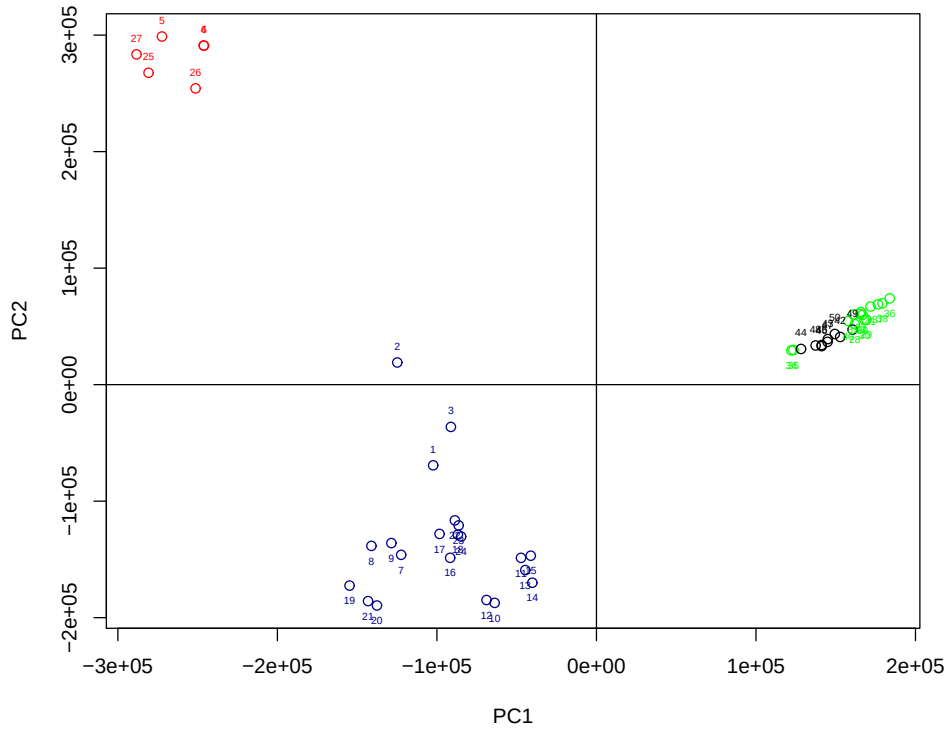


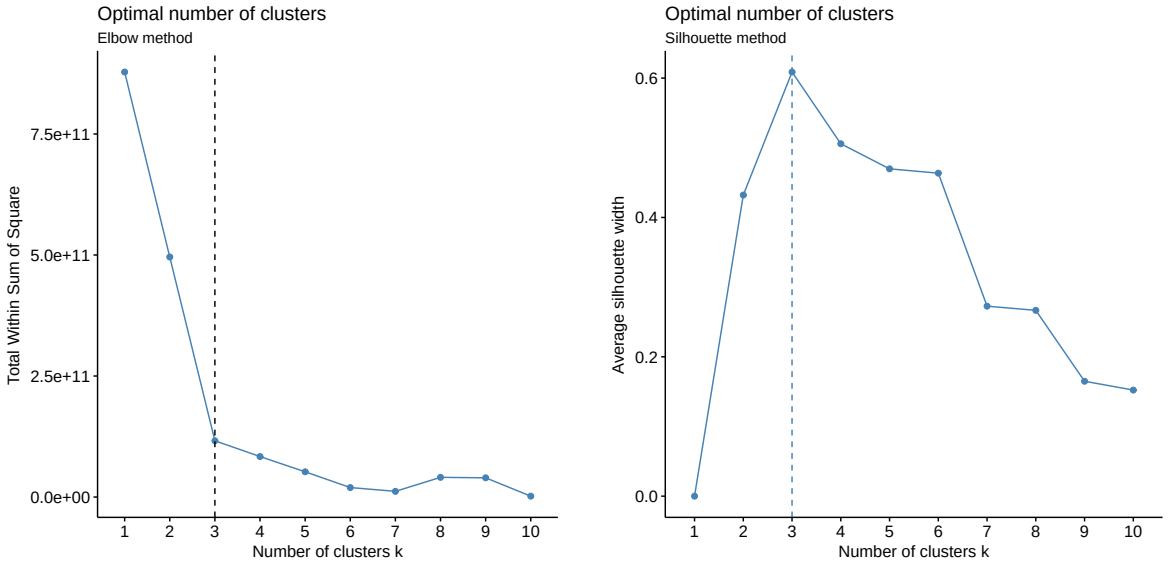
Figure 8.11: Scores plots of the first two PCs, superimposed with the 4-cluster partition derived by the k-means clustering method for data set IV. Coloured points represent the samples in the clusters. The labels of the points in the plot correspond to sample numbers.

data (samples 1 to 27) into two clusters, keeps the Fort William data (samples 28 to 41) as one cluster and also the Dunblane data (samples 42 to 50) as one cluster. These samples can be clustered into four groups by k-means and hierarchical cluster analysis without splitting the replicates from each hive. The first group for the Aberdeenshire data set includes samples 4, 5, 6, 25, 26 and 27, and the second group includes the remaining samples. Although, k-means provides a completely different result from HCA (Figure 7.24), the k-means method seems to be better than HCA here because it is able to separate samples of data set IV depending on location, even though 4 clusters were optimal for both methods.

8.6 Application of the k-means Algorithm to the Libya Data

To determine the optimal number of clusters, the Elbow technique will be used in the derived k-means partitions of 2-10 clusters. Figure 8.12 (left plot), shows the results for the Elbow method on the Libya data set. The dashed line shows the point which corresponds to the optimum number of clusters. The optimum solution is 3 clusters for the Libya data.

To confirm the result of the Elbow method, the average silhouette value for the k-means



(i) Elbow method (ii) Average silhouette widths

Figure 8.12: Elbow method and Average silhouette widths for k-means partitions of 2-10 clusters for the Libya data, where the dashed line represents the optimum number of clusters.

partitions of clusters will be computed for the Libya data set. Figure 8.12 (right plot) gives the overall average silhouette widths of clusters. The dashed line shows the optimal number of clusters. From the average silhouette information and the Elbow method, it

can be concluded that the optimum number of clusters is 3 for the Libya data set.

The silhouette coefficient measures how well an observation is clustered and it estimates the average distance between clusters. The silhouette plot displays a measure of how close each point in one cluster is to points in the neighbouring clusters. A silhouette plot for the cluster partition can be seen in Figure 8.13 for the Libya data. The average

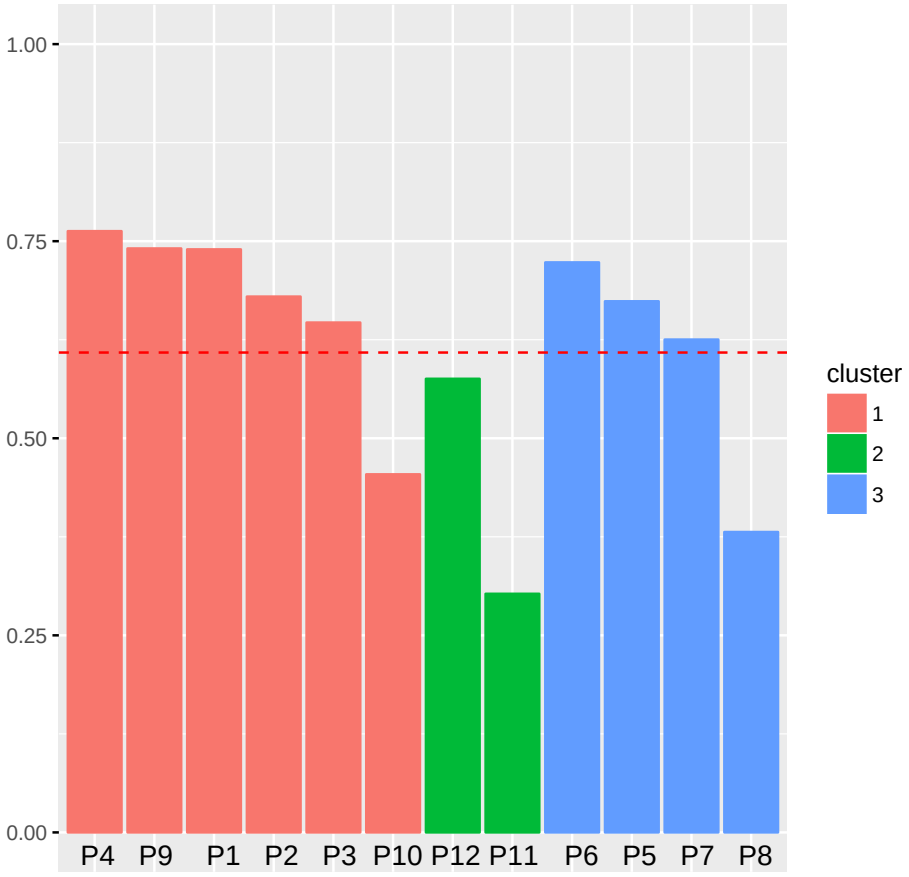


Figure 8.13: Silhouette plot for the 3-cluster partition derived by the k-means clustering method from the Libya data. The average silhouette widths for the 3 clusters are 0.67, 0.44 and 0.60 respectively, and the average silhouette width for the entire data set is 0.61.

silhouette width values for clusters 1, 2 and 3 are 0.67, 0.44 and 0.60 respectively, and the average silhouette width for the entire data set is 0.61. From Figure 8.13 for k-means

and Figure 7.26 for the Canberra-Average method for HCA, there are no misclassified samples (negative silhouette widths) in both methods. The silhouette plot also shows that the three clusters contain 6, 2 and 4 samples respectively, the same as for the HCA clustering case. However, the HCA method is better than k-means, as the HCA method's overall average silhouette width of 0.81 is high compared to that of the k-means clustering method (0.61).

A two-dimensional projection of the clustering solution can be seen in Figure 8.14 in principal component space. In the scores plot, the colours represent the different clusters of points and the points are labelled with the sample numbers. There is no clear distinction among locations in this clustering. The principal component analysis (PCA) based on the 300 features was constructed, from the mean-centred and Pareto scaled data. Samples P5, P6 and P7 from the South-east of the country and also P8 from the South-west were grouped fairly close together (see Figure 3.5 for the location of samples). The samples from the coast did not divide according to longitude. Samples P1, P2, P3, P4, P9 and P10 are in the first group, the second group is P11 and P12, and the third group is P5, P6, P7, and P8, composed of samples from the East and West of the country. Although P10 was collected from a site close to P11 and P12, it seems to be quite different in composition as it is separated from them in the plot.

8.7 Conclusions from the k-means Method

In the case of hard partitioning, the k-means method was the obvious choice, as it is the most popular in metabolomics data analyses. The k-means algorithm described in Section 8.2 was used on the propolis data sets and the optimum number of clusters was determined with the aid of a range of stopping rules, such as the Elbow technique and the average Silhouette coefficient. The 2, 4, 4, 4 and 3 cluster partitions derived by the k-means clustering were the best hard partition of data sets I, II, III, IV and Libya. The

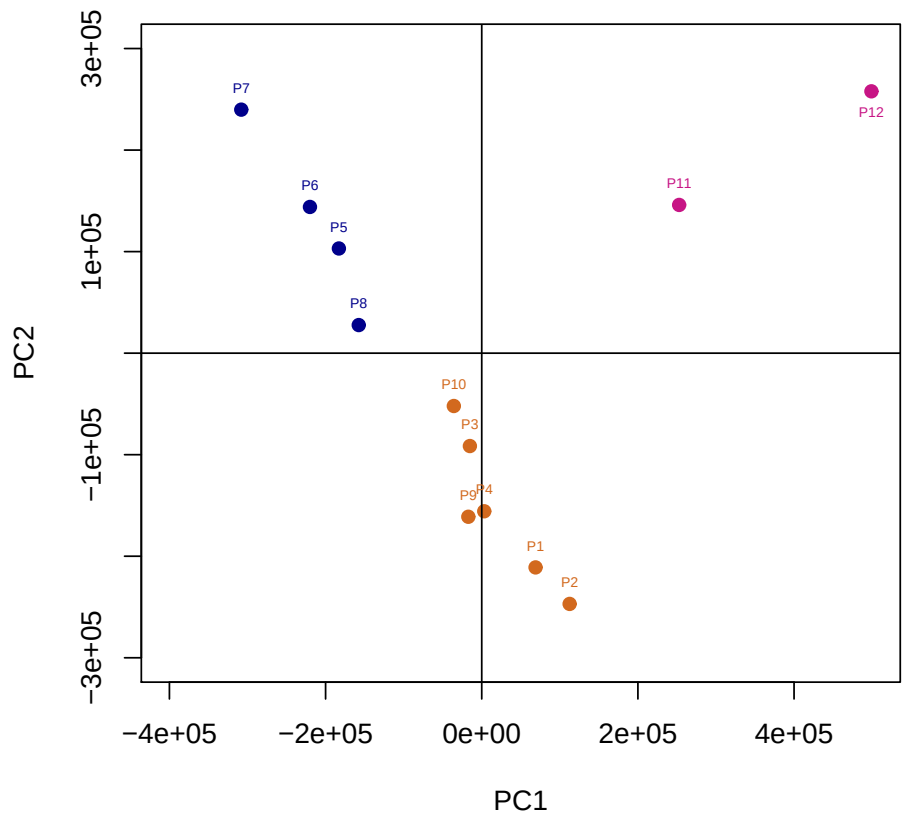


Figure 8.14: Scores plots of the first two PCs, superimposed with the 3-cluster partition derived by the k-means clustering method for the Libya data. Coloured points represent the samples in the clusters. The labels of the points in the plot correspond to the names of samples.

2-cluster k-means partition of the Aberdeenshire data is exactly the same as in the 2-cluster HCA clustering, and the silhouette widths in the two methods are the same. In addition, the 4-cluster k-means partition for the Fort William data is exactly the same as the 4-cluster HCA clustering. On the other hand, the 4-cluster k-means partition for the Dunblane data is different from the 4-cluster HCA clustering with regard to number of samples contained in the four clusters of the partitions, as there are different numbers of

samples in each cluster for the two methods, and the silhouette widths of the samples and the average silhouette widths differ between the two methods. For data set IV, k-means provides a completely different result from HCA, and k-means seems to be better than HCA because it is able to separate samples of data depending on their location. Finally the HCA method is better than k-means for the Libya data set, as the HCA method's overall average silhouette width of 0.81 is much higher than that of the k-means clustering method (0.61) and both methods cannot clearly discriminate samples depending on their location.

The next chapter describes in detail a relatively new method for types of data such as metabolomics data, i.e. Self Organising Maps (SOM), in another attempt to devise suitable clustering models. It is also applied there to identify groupings present in the data sets I, II, III, IV and Libya, and the results are compared to those of HCA and k-means.

Chapter 9

Competitive Learning Algorithms

As a particular type of artificial neural networks, self-organising maps (SOMs) are a machine learning method which are trained using an unsupervised, competitive learning algorithm to produce a mapping from a multidimensional input space onto a lattice of clusters (or neurons). We will compare the results here with the previously used methods. This chapter describes Self Organising Maps in Section 9.1, application of SOMs to data sets I, II, III in Section 9.2, and also application of SOMs to data sets IV and Libya in Sections 9.3 and 9.4 respectively.

9.1 Self Organising Maps (SOMs)

9.1.1 Overview

The Self- Organising Map (or SOM) is one of the most prominent competitive learning algorithms. It was outlined by Kohonen, who describes it as a numerical method of data visualisation and cluster analysis (Honkela et al., 1995; Kohonen, 1990). He first showcased the concept of SOMs in the 1970s. Over the decades, its primary algorithms have

been extensively updated and enhanced. The work of Wang et al. (2002), Wang et al. (2005), Dittenbach et al. (2002), Salas et al. (2007) and Jin et al. (2004) describes some of these evolutions. One of the biggest changes is their increased suitability and tolerability of diverse input types and fields of application.

For example, SOMs can be used for the ordering of representative species observed as part of multivariate ecological research (Park et al., 2006). As Kalteh et al. (2008) proves, SOMs may be employed as part of attempts to investigate water source complications. They can even be used to assess the metabolic indicators of patients with chronic conditions like diabetes (Makinen et al., 2008) and cardiovascular dysfunction (Suna et al., 2007). This method has much in common with Artificial Neural Networks (ANN) and corresponding learning processes (Taner, 1997). As Izenman (2008) points out, SOMs are utilised as a tool for generating basic facsimiles of the human brain and its complex neural connections.

In the field of chemistry, SOMs are valuable and highly regarded. While there is a steadily expanding collection of studies about its application (Miljkovic, 2017), its methods and processes have not been fully investigated. Certainly, SOMs are not as well known or as explored as partial least squares (PLS) and PCA. This may be due to a preference for mainstream plugin packages among contemporary chemists, which do not offer SOM analysis. Despite this, it is clear that SOMs offer remarkable levels of tractability. With this method, many of the constraints typically associated with analytical chemistry (like technological capabilities) become redundant or significantly lessened. Within this field, SOMs may be shaped to fit a variety of scenarios. For instance, this method can be used for quality control or by using supervised SOMs. Consequently, it has a much broader scope than simple visualisation of clusters (Li and Pan, 2013).

The basic premise underpinning SOMs is very straightforward. Given a series of representatives $w_i, i = 1, ..., m$, if input vector x is presented to the algorithm, every representative w_i must then compete with each other to be nearest to x (using the relevant distance measure). The one that is nearest to x is considered to be the 'winning' representative

(Theodoridis and Koutroumbas, 2003). It gets to stay there, in close proximity to x, while the other representatives advance at a delayed pace or remain unaltered.

The SOM technique relies on unsupervised learning to generate a mapping of high dimensional input space within much simpler 2D or 3D planes. According to Brereton (2009), it seeks to locate clusters such that any two clusters sharing the same output area end up with representatives very near to one another on the input plane (Dittenbach et al., 2002). When visualising a suitable input area for SOM, there needs to be a substantial network containing many linked nodes (see Figure 9.1). The two dimensional demonstration of this algorithm would be as nodes linked to form a rectangle, oblong, hexagon, or square. At all times, the topological correlations between the input data components must remain as static and authentic as possible. Makinen et al. (2008) explains that, for researchers of chemometrics, the technique is useful for gathering information about relationship between samples. It is also an efficient way to visualise characteristic variables, specific samples of interest or groups of samples (Suna et al., 2007).

As Silva and Marques (2007) explain, there are two common forms of SOMs. They are batch and on-line (Izenman, 2008). The batch technique uses all input vectors simultaneously. On-line SOM, on the other hand, sees input vectors used in a carefully controlled manner. They are still selected arbitrarily, but they get added one by one. Before an investigation can begin, the researcher must decide on the dimensions of the SOM diagram. Crucially, this decision tends to be inaccurate because it is common to pick a node (map unit) volume that is much greater than the predicted number of clusters within the data set. This is because, at first, the researcher simply does not know the requisite volume of nodes. It takes a little guessing and a lot of testing to get to a place where the SOM diagram has a more feasible shape. The diagram is repeatedly altered and this decreases the volume of nodes which it contains until they offer a more realistic prototype. Every node is linked to a representative (or prototype or codebook vector) within the input plane, for instance w_c. When a weight vector is applied (see below), the net weight of all nodes aligns with the number of factors considered as part of the initial

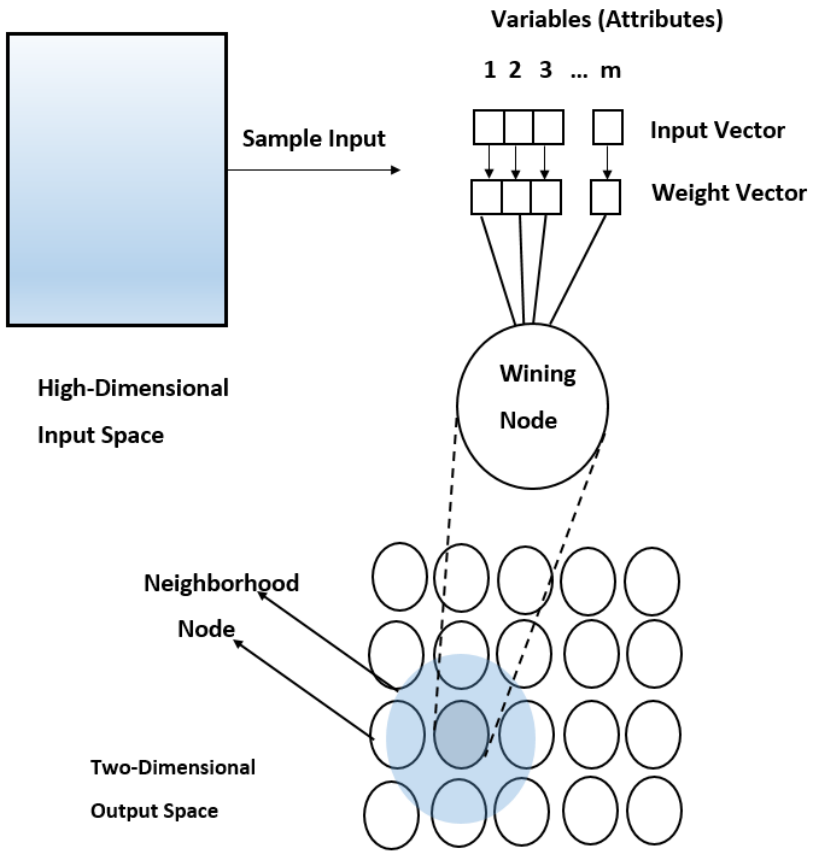


Figure 9.1: Graphical illustration of a self-organising map

data area. Consequently, this stratum of weight linked to every significant factor may be treated as the third dimension of the SOM diagram. At the start, all w_c vectors and their assigned elements are configured as arbitrary numbers. They are selected and ordered using a random number tool that picks any number, as long as it is covered by the scope of the available data.

9.1.2 Classic On-line SOM Algorithm

If a researcher wishes to operate an on-line SOM investigation, they must adhere to the following process:

Select the Best Matching Unit

- First, an initial representative is chosen. Some of the choices that can be made include random samples from the data, random vectors with elements from a N(0,1) distribution, and vectors from the direction of the first two principal components of the data. Usually a sample vector, x_k, is selected randomly from the data set, with or without replacement.

- Following this, there is a calculation of a dissimilarity a distance $d(x_k, w_{ij})$ between the arbitrarily picked sample x_k and every one of the unit weights for the map w_{ij}. Much the most common distance metric used for that purpose is the Euclidean distance measure:

$$d_{(x_k,w_{ij})} = \|x_k - w_{ij}\| = \sqrt{\sum_{j=1}^{n}(x_{kj} - w_{ij})^2}$$

where x_{kj} is the value of variable j for sample x_k, w_{ij} stands for the weight of variable j for map unit i, and n represents the total number of samples. The map unit with the lowest $d(x_k, w_{ij})$ for vector x_k is highlighted and renamed the Best Matching Unit (or BMU) of sample x_k for this particular investigation. That is:

$$d_{(x_k,w_{BMU})} = \min_k\{d_{(x_k,w_{ij})}\}$$

where

$$BMU = \arg\min_k\{d_{(x_k,w_{ij})}\}.$$

Determine and Adjust the Nearest Map Unit

- The best way to determine which map unit is nearest to the BMU is to consider the concept of neighbourhood and corresponding neighbourhood units. Assuming the Euclidean distance of the codebook vectors w_U and w_{BMU} for map unit $U \in M$ is lower than a pre-established limit β known as the 'neighbourhood width,' it may be

considered a neighbour of the BMU unit (U_{BMU}). However, this newly configured neighbourhood set (or N_{BMU}) must carry units that meet the conditions below:

$$N_{BMU} = \{u : d_{(w_{BMU}, w_u)} < \beta\}.$$

- According to Izenman (2008), learning rate frequency represents the degree to which the targeted map is due to be changed during every new repetition or run through. It does this to approximate a sample as well as it can. It is only necessary to adapt and adjust the weight vectors for units that are deemed part of a BMU neighbourhood group. This is achieved by applying a distance weighted equation in which ϕ is a learning frequency signifier and β_f is a neighbourhood factor:

$$w_{u+1} = w_u + \phi\beta_f(x_k - w_u)$$

There are other similar processes, and some of the most common are as follows:

$$Linear : \phi_j = \phi_0 \left(1 - \frac{j}{J}\right)$$

$$Exponential : \phi_j = \phi_0 \left(e^{\frac{-j \ln(\beta_0)}{J}}\right)$$

$$Power : \phi_j = \phi_0 \left(\frac{0.005}{\phi_0}\right)^{\frac{j}{J}}$$

$$Inverse : \phi_j = \frac{\phi_0}{\left(1 + \frac{100j}{J}\right)}$$

where j represents the latest algorithm repetition, J is the total number of repetitions or versions, ϕ_0 is the preliminary learning frequency and β_0 signifies the earliest neighbourhood width. If any of the above mentioned are applied as functions, learning frequency is seen to decline monotonically right through to the close of the investigation. It is necessary to adjust or calculate a new neighbourhood width every time the algorithm is repeated. Normally, β has a large value initially.

However, as the investigation advances, it is expected that the function will slowly drop in size. Eventually, just the BMU and its contiguous neighbours are left to be adjusted and renewed. Some neighbourhood functions that are suitable for SOM testing are as follows:

$$Exponential : \beta_j = \beta_0 e^{\frac{-j \ln(\beta_0)}{J}}$$

$$Gaussian : \beta_j = e^{\frac{-|w_u - w_{BMU}|^2}{2(r_\beta)^2}}$$

and

$$Square(or\ \ bubble) : \beta_j = \begin{cases} 1, & \text{if} \quad |w_u - w_{BMU}| \leq r_\beta \\ 0, & \text{if} \quad |w_u - w_{BMU}| > r_\beta \end{cases} . \tag{9.1}$$

According to Brereton (2009), J is often used to signify a number 500 times the number of units on the map for β_0 equal to half the width of the map and for ϕ_0 of 0.1. In some instances, a k-means algorithm is applied as an adjustment phase. When r_β represents the neighbourhood radius, the algorithm is repeated until the predetermined number of tests or versions (J) have been completed.

9.1.3 Classic Batch SOM Algorithm

Batch SOM works by adjusting and renewing the weight vectors only after each learning phase is complete. In this case, a phase is the application of the entire training data set to the algorithm. The adjusted weights and optimal BMU are calculated with the formula below:

$$w_{\acute{u}} = \frac{\sum(\beta_{\acute{f}} x_k)}{\sum \beta_{\acute{f}}}$$

and the winner unit (BMU) can then be found using equations

$$d_{(x_k, w_{\acute{u}})} = \|x_k - w_{\acute{u}}\|^2 = \sqrt{\sum_{j=1}^{n} (x_{kj} - w_{\acute{u}})^2}$$

and

$$d_{(x_k,w_{BMU})} = \min_k \{d_{(x_k,w_{ú})}\}.$$

The batch algorithm process is much faster than the on-line one. For one thing, learning frequency is no longer a necessary consideration, as it is for batch techniques. Therefore, the likelihood of low convergence and related problems is much smaller, although, it does need the entire series of input vectors throughout the training phase, not just a partial selection. Lastly, the sequence of the input vectors and their application is not as significant here because each weight vector is renewed at the end of the iteration. It means that the final input vectors have no real effect on the eventual outcome. This process is replicated, from the start, over and over until it reaches the predetermined convergence requirements. Here, the neighbourhood function is the same as it is for the on-line process.

9.1.4 Quality of Mapping

As Villmann et al. (1997) states, self-organising maps (or SOMs) seek to maintain the topological characteristics of the input plane. The quality of a map may be influenced by the fact that this process involves vector quantisation algorithms. It is defined by the application of the multidimensional input area to a 2D or 3D output plane, as normally occurs when working with biological data spaces (Vesanto, 1999). It is very important that the algorithm is seen to protect input space topology. If mapping continues uninterrupted, samples that are near to the input plane get positioned close together within its topology and the output space. If the map is of a high quality, this process is conducted with a similarly high degree of accuracy. No samples that are not close together on the input are placed together in the output (Bauer et al., 1999; Pölzlbauer, 2004). Brereton (2009) explains that, even though the probability variation of the input plane is unlikely to be presented poorly by a SOM algorithm, it is still necessary to account for other factors

that may have an impact on mapping quality. They include mapping resolution (Brereton, 2009; Polani, 1999) and continuity (Kiviluoto, 1996; Neme and Miramontes, 2005). Other indicators of metabolomics and chemometric accuracy are as follows:

The Mean Quantisation Error (E_{MQ})

Mean Quantisation Error (or E_{MQ}) is a quality signifier that is used to test the efficacy and accuracy of any type of clustering algorithm or vector quantisation. It considers the mean distance between the sample vectors and their assigned cluster centroids. According to Pölzlbauer (2004), for SOM, this means the typical distance between every sample and its corresponding BMU, but it must be calculated after the final SOM algorithm has been performed, as E_{MQ}, defined as:

$$E_{MQ} = \frac{1}{m} \sum_{k=1}^{m} d_{(x_k, w_{BMU})}$$

where m is the number of units (centroids) in the SOM. Brereton (2009) points out that E_{MQ} relies on the training data and initialisation processes. Therefore, it is most useful when the number of map units is equivalent to or bigger than the number of training samples. It should be noted that E_{MQ} is not a suitable way to test SOMs with varying grid or diagram dimensions. The bigger the map, the smaller the value of E_{MQ}, as it reduces monotonically as the dimensions increase.

Topographic Error (E_{TE})

Topographic Error (or E_{TE}) is an indicator of topological conservation (Pölzlbauer, 2004). To be precise, it determines the degree of stability within SOMs. It is fair to consider a map locally preserved and unbroken if, for a sample x, the closest and second closest representatives are contiguous units. If this is not the case, the topology has been altered and is, therefore, not fully preserved. As with E_{MQ}, this measure of quality becomes more reliable as the size of the map increases. The work of Kiviluoto (1996) calculates topological accuracy by counting and regulating the amount of local inaccuracies across all samples. This generates a measure of

topological continuity. E_{TE} is represented by the following equation:

$$E_{TE} = \frac{1}{m} \sum_{k=1}^{m} \sigma_{(x_k, w_u)}$$

where m is the number of units (centroids) in the SOM. $u = 1, ..., U$, U is the last unit in the map and u_i stands for the vector of unit i (Neme and Miramontes, 2005; Villmann et al., 1997). If the neuron winner w_u of vector x_k is near the neuron, the distance from x_k to it is the smallest one, regardless of the neuron winner, then $\sigma_{(x_k, w_u)} = 0$, otherwise, $\sigma_{(x_k, w_u)} = 1$. While this formula can reveal the degree of accuracy across all local neighbourhoods, it is not a suitable way to measure the nature of these inaccuracies (Kiviluoto, 1996). On the other hand, it is reasonable to assume that a great many errors indicate a breakdown of topology across the input plane. In this case, the SOM should be considered carefully, as many of its features may be imprecise.

9.1.5 Visualisation

Before talking about visualisation, we should determine the shape and size of a grid for the SOM, as follows:

Identify the grid of the SOM

The shape of the SOM grid may be chosen to be hexagonal, to avoid a preference of the SOM algorithm towards horizontal or vertical directions (Park et al., 2006). In addition, the size of the map must not be such that it has more units than the number of samples in the data set, to ensure a better response from the map quality criteria. In this work, the classic online SOM algorithm was used. The map size is important to detect any deviation in the data. If the map size is too small, it might not explain some important differences that should be detected. Conversely, if the map size is too big, the differences are too small (Wilppu, 1997). The map size depends on the number of samples to be trained. Although no strict rules exist to

define the optimal map size, there are several possible methods. First, setting the number of output neurons approximately equal to the number of the input samples seems to be a useful rule-of-thumb for many applications where the data sets are relatively small (Kaski, 1997). However, attention should be paid to over-fitting problems when a large map size is used. This may happen when the number of output units is as large as or larger than the number of samples. A second method to determine the map size is by using Vesanto's heuristic formula (Vesanto et al., 2000) which states that the total number of map units, N_U, is given by:

$$N_U = 5\sqrt{\text{number of samples}} \ .$$

In this case, the lengths of the grid sides can be calculated by setting the ratio of the lengths of the sides similar to that of the two largest eigenvalues of the training data such that the product of the lengths is as close as possible to N_U.

There is more than one way to demonstrate the outcomes of SOM investigations. Some of the most common are outlined below and illustrated using the Scottish and Libya data sets used previously in this thesis, before a formal analysis of those data sets.

Unified Distance Matrix (U-matrix)

The Unified Distance Matrix is a particular kind of visualisation process. It helps the researcher to accurately locate any clustering on the map. This is achieved with a highlighting technique, like assigning a colour to the Euclidean distance between adjacent representatives, for example. When map nodes are positioned near to one another on both the output and input planes, a lower value is generated. As Brereton (2009) points out, only units with a substantial distance produce the bigger values. Figure 9.2 shows an example of a U-matrix for the Aberdeenshire, Fort William, and Dunblane data. The configuration sizes were chosen as described in Section 9.2.2. For the Fort William results (the middle plot) the upper right cell in the plot

is highlighted in white and represents units that are far away from one another. The generated values are vast ($> 1.8e^{+10}$), particularly when compared with the dark red markers and smaller values ($< 5e^{+9}$) in the upper and bottom left side areas. Looking at the data set for Aberdeenshire, the largest plot is highlighted with one white and three yellow markers. It suggests that their points are positioned far from one another on the input plane and that the values generated are substantial ($> 4e^{+10}$). They are certainly bigger than for the points highlighted with red markers right in the plot ($< 2e^{+10}$). Finally, the Dunblane data set shows yellow and white markers at the upper left side and the middle right of the map (Figure 9.2). The colours indicate distant points (samples), with substantial values ($> 1.4e^{+10}$). As with the other two data sets, the red marker indicates units or samples with lower values ($< 1e^{+10}$).

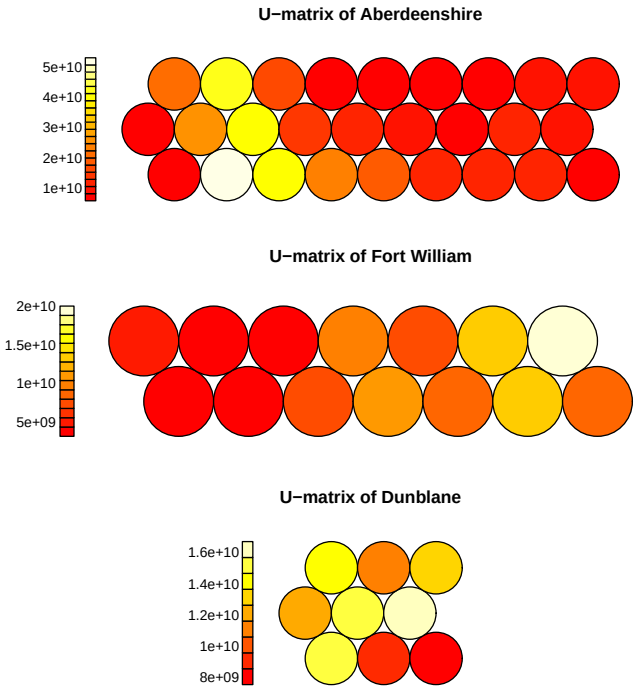


Figure 9.2: Examples of Unified Distance matrices (U-matrices) for the Scottish propolis data sets.

Hit Histogram

The hit histogram is applied when a researcher needs to visualise the BMU for every one of the samples upon completion of testing. It assigns every unit a value. This value represents the frequency with which each unit has been selected as Best-matching unit at the end of training (Brereton, 2009). The process may involve a 2D or 3D plot. Both are a valid way to calculate this measure. When using a 3D histogram, the height of each hexagonal bar is proportional to the number of hits. When using a 2D plot (shown below), the number of BMU hits is shown by the size of the shaded map units. Preferably, when several classes are contained in a data set, only one or at least a small number of map units should match the BMU of all samples from the same class and therefore correlate with a large number of hits. When there is a high number of hits on a map, it tends to give a concentration of samples around its units and these units are mainly on the periphery of a SOM grid. Furthermore, regions of the map that show high numbers of hits tend to link with areas of similarity within a U-matrix plot.

Figure 9.3 shows an example of a 2D colour-coded hit histogram for a grid of dimensions 9×3 for Aberdeenshire, 7×2 for Fort William and 3×3 for Dunblane. For example, for the Aberdeenshire data (the top plot in Figure 9.3), the six yellow units in the hit histogram have a value of 3, meaning that these special units were the best-matching unit of any sample three times when the training finished. The grey units contain no samples. The three orange units had two hits each. Concerning the Fort William data (the middle plot in Figure 9.3), the plot shows a top left unit marked in yellow. It carries a value of three, which means that this special unit was the BMU of a sample three times when the training finished. The four orange units had two hits each. Finally, the Dunblane data set shows three units marked in yellow, with a BMU value of two. The three units marked in red carry a value of one. The grey units are not considered valuable for this test.

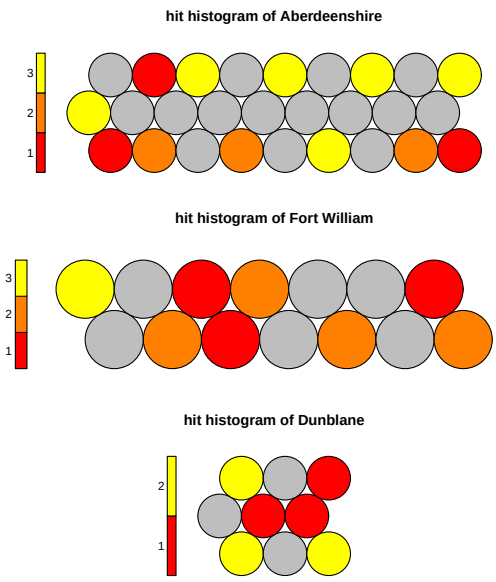


Figure 9.3: Examples of hit histograms for Aberdeenshire, Fort William and Dunblane, respectively from top to bottom.

Component Planes

Unfortunately, the U-matrix and Hit Histogram are not complete tools. They have a key weakness, that neither is able to demonstrate the significance of variables within the input plane. Nevertheless, it is important to investigate the influence these variables have on the shape and implications of the map (in a similar way to PCA). Component planes are employed as a way to demonstrate the effects which each input variable has on the map, and correlations between samples and variables (Brereton, 2009). The biggest advantage of employing component planes is the capacity to determine whether a variable can be used to understand a class and, if this is not the case, whether it can make distinctions between classes. If the data set contains just two classes, any variable that can be used to define a class may also be considered an effective distinguisher and identifier. If there are more than two classes, this may not be the case (Makinen et al., 2008).

It works by building a single element plane for every one of the variables. Then, their

respective weight or significance is represented with a coloured marker in relation to their usefulness for expanding a particular part of the map. Consequently, correlations between variables and samples are demonstrated with colour-coded highlights. Each map unit, k, is given a specific colour signifying its proportionality to the significance of w_{kj} of unit k for the selected variable j.

Figures 9.4, 9.5 and 9.6 show component planes for three variables contained in the Aberdeenshire, Fort William and Dunblane data respectively. Each plane depicts the size of the respective map unit influences or weights, in this case, phenylacetic acid, hydroxybenzoate, and benzoic acid. This is the case for both the Dunblane and Fort William data sets. Where the map is coloured in a darker tone, the variables carry bigger values and match areas of similarly higher value in the U-matrix plot. To be specific, the higher variables in the U-matrix are more strongly linked to those samples with a dark colour on the component plane. It means that, for the examples provided, the three variables have a dark colour code but quite small values (< -5000 for the first variable, < 0 for the second variable and < -2000 for the third variable) and contain samples which are strongly associated with these variables. However, considering the white colours with values $> 1e^{+4}$ for the first variable, $> 3e^{+4}$ for the second variable and > 4000 for the third variable of the Aberdeenshire data, the samples contained in these units are related to these specific variables.

From Figures 9.4, 9.5 and 9.6, it can be seen that their respective planes show areas in which particular variables carry high values, but they also show the opposite - areas with rather small values. This is another reason why component planes are such a valued tool. They make it easy to recognise the areas of higher values and lower values in these figures.

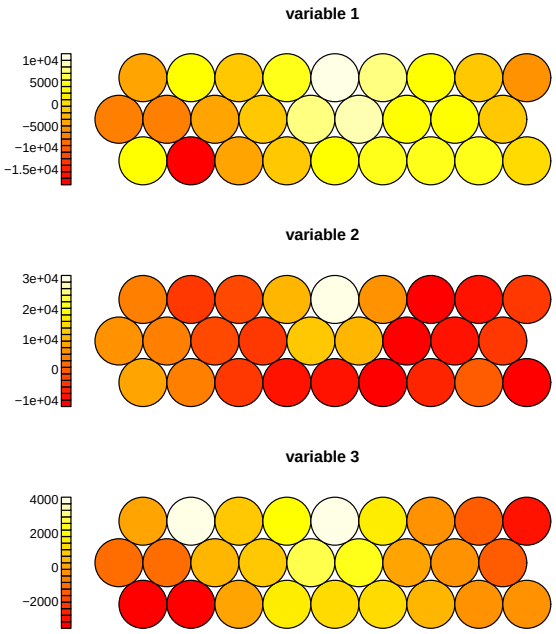


Figure 9.4: Examples of component planes of the Aberdeenshire data for the first three variables.

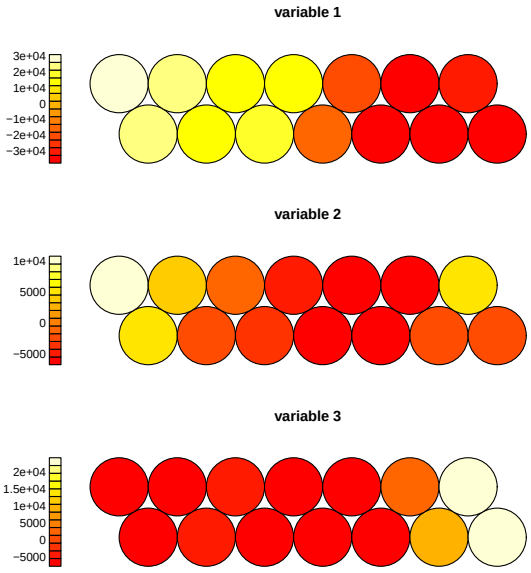


Figure 9.5: Examples of component planes of the Fort William data for the first three variables.

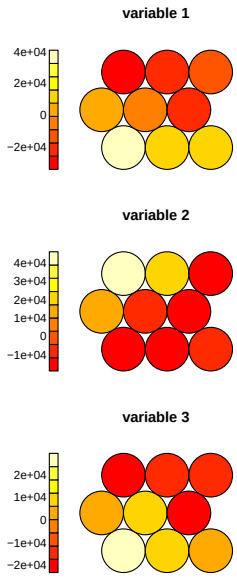


Figure 9.6: Examples of component planes of the Dunblane data for the first three variables.

9.2 Application of SOMs to Data Sets I, II and III

9.2.1 Overview

As the SOM maps that are produced are completely dependent on the input data used for the analyses and the various learning parameters, there are no standard rules established to produce a good quality map, i.e. a highly accurate and well-ordered map in every case. This does make the method hard to use. The SOM seeks to produce a low-dimensional representation of samples. For example, in the cases of the three propolis data sets (Aberdeenshire, Fort William and Dunblane), it is probably more important to obtain a highly accurate map (as good a representation as possible of the input space in the output space) than to preserve the topological order of the input space, whereas in a data-mining application the order would logically be more important than the accuracy, as it commonly involves documents. Therefore, in the cases of the propolis data sets I,

II and III, the mean quantisation error is probably more important than the topographic error (Brereton, 2009). With these considerations in mind, the following SOM analyses were chosen to investigate whether the SOM algorithms can be used to identify any classification of samples depending on the chemical properties or location, and not to find the best possible mapping of the data.

9.2.2 Initialisation

Before the analyses, the samples were normalised by mean-centring and Pareto scaling, to eliminate the possibility of any influence on the SOM results by any of the metabolites. In our case, the total number of map units N_U is approximately 26 ($\backsimeq 5\sqrt{27}$) for Aberdeenshire (27 samples), and 19 ($\backsimeq 5\sqrt{14}$) > 14 (number of samples) for Fort William, therefore the size of the map is taken as 14 for Fort William, and 15 ($\backsimeq 5\sqrt{9}$) > 9 (number of samples) for Dunblane, therefore the size of the map is taken as 9 for Dunblane, to ensure better response from the map quality criteria for Fort William and Dunblane. The ratio of the two largest eigenvalues of the covariance matrix is approximately $\frac{7}{1}$ for Aberdeenshire, $\frac{4}{1}$ for Fort William and $\frac{2}{1}$ for Dunblane. Thus the grid dimension can be 13×2 for Aberdeenshire data, 7×2 for Fort William and 3×2 for Dunblane (3×3 gives poor quality of the grid), to approximate as closely as possible the map size without violating the ratio rule for these three data sets (Park et al., 2006; Vesanto et al., 2000).

9.2.3 Training

The term training for SOMs refers to be the process of iteratively updating the map unit weights to become more similar to vectors representing the original samples. The training parameters used here are taken as the following: the learning rate initially has a value of 0.05 and decreases monotonically until it reaches the value 0.01 at the end of iterations for the three data sets. The initial radius of the neighbourhood function is approximately

$\frac{2}{3}$ of the estimated map width, to allow for a large part of the map to be updated initially. A sufficient initial radius value will usually cover $\frac{2}{3}$ of all unit-to-unit distances (Brereton, 2009). Thus, the initial radius for the 13×2 grid is approximately 9 for Aberdeenshire, the initial radius for the 7×2 grid is approximately 5 for Fort William, and the initial radius for the 3×2 grid is 2 for Dunblane. The final value of the radius of the data sets I, II and III is 0 at the end of the algorithm.

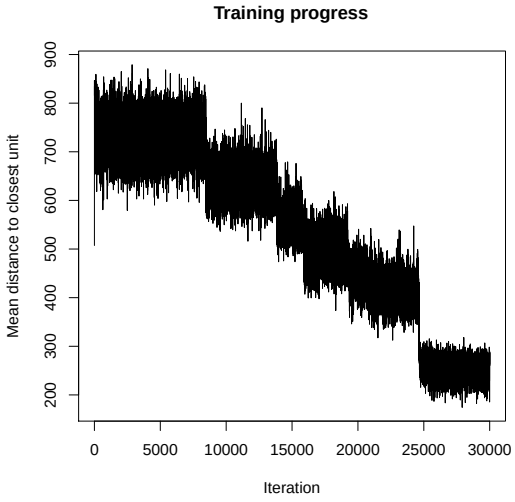
The initial representatives were chosen randomly without replacement from data sets I, II and III. A series of runs was performed using the recommended values for the training parameters (Brereton, 2009; Tan and George, 2004). Thus, the total number of iterations was chosen in each case to be 500 times the map size of the three data sets. One map size was used for each data set, specifically the 13×2, 7×2 and 3×3 maps were chosen for further investigation for Aberdeenshire, Fort William and Dunblane respectively. The neighbourhood width function converges to a small number close to 0 after 25000, 20000 and 26000 iterations respectively for the estimated maps of data sets I, II and III (Figure 9.7 shows the convergence for the grids for Aberdeenshire, Fort William and Dunblane).

9.2.4 Quality of mapping

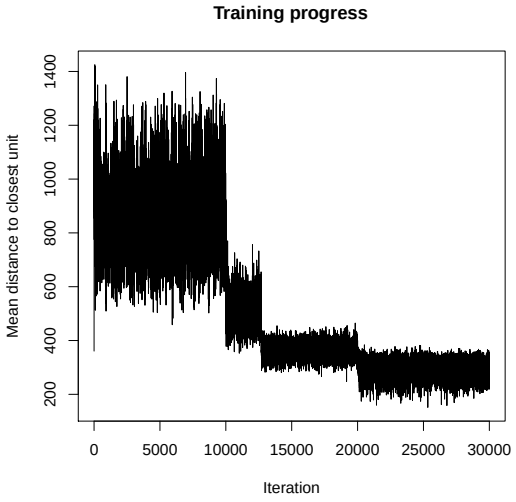
The quality of the mapping can be examined by using specific plots to illustrate how closely to the prototype codebook vectors in each unit the samples in the unit have been mapped. The mean distance of samples mapped in each unit to the codebook vector of that unit can be illustrated using colour-coding such that the smaller the distances (darker colouring), the better the samples in the unit are represented by this unit's codebook vector. We will investigate the analysis of data sets I, II and III by SOMs as follows:

Aberdeenshire (I)

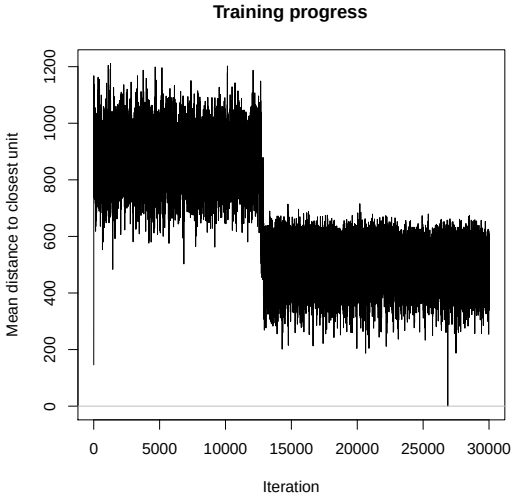
In Figure 9.8 the quality map is illustrated for the Aberdeenshire samples. It can be seen that in general, the quality of the mapping is quite good, as in most of the



(i) Aberdeenshire data (data set I)



(ii) Fort William data (data set II)



(iii) Dunblane data (data set III)

Figure 9.7: Convergence of the neighbourhood width function for the selected maps (13×2 grid) of the Aberdeenshire data, (7×2 grid) of the Fort William data and (3×3 grid) of the Dunblane data.

map units the samples are quite close to the respective codebook vectors (the colour is dark). However, clearly, two of the units have not been mapped accurately, with the worst approximation being in the top right unit (white). The bottom right map

unit (in light yellow) has also been mapped badly.

In addition, the Unified Distance matrix can be used to illustrate the average

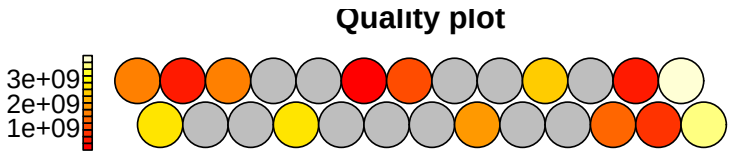


Figure 9.8: Illustration of the quality of mapping for the Aberdeenshire samples. A grey unit in the quality map means that there is no sample mapped to this unit.

distance of each map unit to all immediate neighbour units. Units near a cluster boundary are expected to have higher average distances to their neighbour units (Figure 9.9). The black lines indicate a two-cluster solution using hierarchical clustering, to allow comparison of the SOM clustering to the best HCA solution. The units on the right side of the U-matrix are closer to each other than those on the left side of the matrix, however there is no indication from this matrix that there is a big number of clusters in the data according to the 13×2 SOM analysis. A

(i) Unified Distance matrix (ii) Hit histogram

Figure 9.9: Unified Distance matrix and Hit histogram for Aberdeenshire for the 13×2 grid.

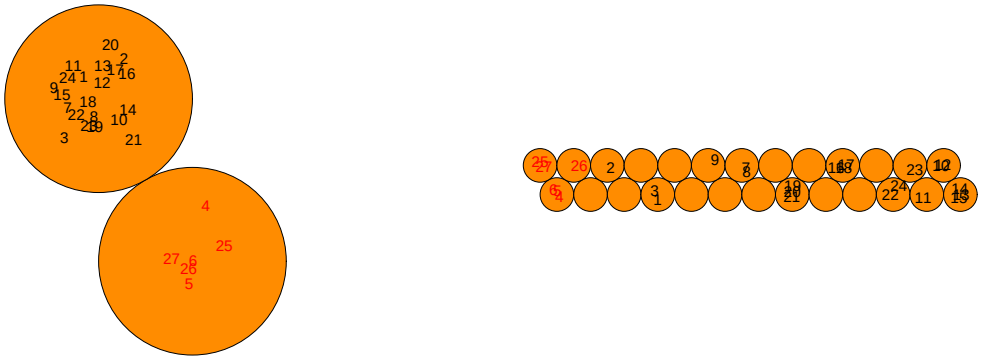
two-dimensional colour-coded hit histogram for the SOM solution can be seen in Figure 9.9. The units with three hits indicate the existence of clusters in the areas surrounding these units, and the concentration of samples around these units is expected to be far larger than elsewhere in the map. Four such units are (2,10), (1,1),

(1,8) and (1,13), where the first number corresponds to the row and the second to the column in which the unit is located in the map. (with (1,1) being the bottom left-most unit and (2,13) the top right-most unit in the map). The results of the hit histogram for these four units correspond to those regions in the U-matrix with high similarity, e.g. the single unit (1,1) at the bottom-left of the map and the 3 units (1,8), (1,13) and (2,10) at the right part of the U-matrix plot. As they are separated by the black lines of the best two-cluster HCA partition, these indicate the existence of clusters in these areas.

The mapping of the samples assigned to each of the two groups provided by the 2×1 map, as well as the colour-coded samples map (using the corresponding label colours of the groups obtained by the 2×1 map) for the 13×2 grid can be seen in Figure 9.10. The cluster sizes of the map on the left are 6 and 21 for groups 1-2 respectively. In the 2×1 grid, group 1 corresponds to the bottom unit, while group 2 corresponds to the top unit. In the 13×2 grid, the corresponding groups to the 2×1 grid are 1 and 2 from left to the right in the map. Comparing the clustering results of the 13×2 map to those of the hierarchical clustering in the U-matrix plot (Figure 9.9), it is clear that there is identical clustering in both solutions.

To examine how the variables influence the map and what is the relationship between each variable and the samples in the data, component planes have been created for a selected number of variables. The maps for the ten variables with the largest mean values can be seen in Figure 9.11. Most of these variables are also among the ten variables with the largest variance. The darker a unit in a component plane for a variable is, the closer the relation of this variable to the unit is. Upon investigation of the component planes, the following can be deduced:

- The most commonly appearing variables are 117, 147 and 177, i.e. these metabolites appear to be very closely related to almost all units in the map.



(i) Samples - 2×1 map (ii) Samples - 13×2 map

Figure 9.10: Illustration of clustering of the Aberdeenshire data to two groups using SOM.

- The least common variables are 57 and 569. Very few units appear to be associated with these metabolites, with the first variable being the least associated to the map, of the two.

- The samples with consistently high intensity values in all component planes are samples 4, 5, 6, 25, 26 and 27 in cluster 2 and samples 1, 2 and 3 in cluster 1.

- Samples 7, 8, 9, 19, 20 and 21 in cluster 1, are the samples least associated with the variables.

- None of the component planes is capable of describing the two clusters.

A two-dimensional projection of the Aberdeenshire data superimposed with the clustering solution derived by the 2-cluster SOM partition can be seen in Figure 9.12. We can compare between the first two principal component scores according to the results from HCA and k-means, which both indicate the same result, and SOM with the partition derived by the selected SOM clustering model. We can see

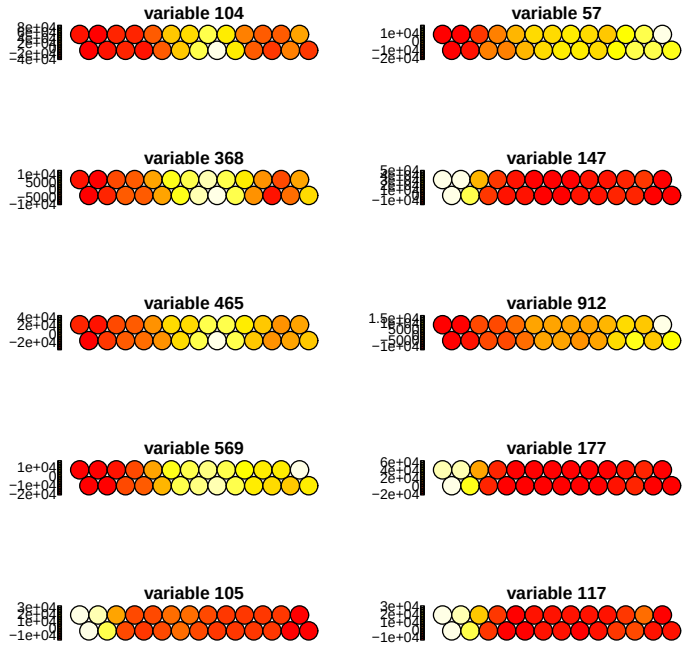
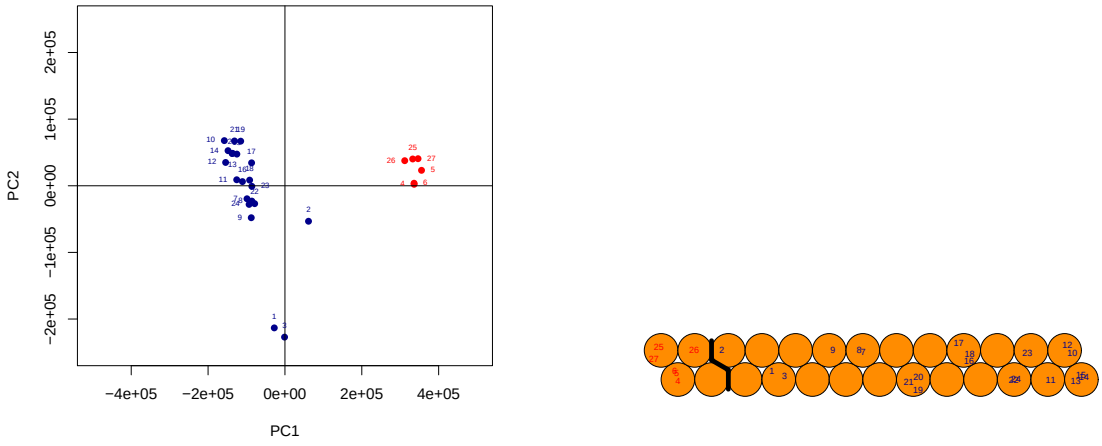


Figure 9.11: Component planes for 10 selected variables among the chemical components of the Aberdeenshire samples, labelled by number of the chemical components for Aberdeenshire.

roughly the same proximities and clusters. But there is a linear constraint in the PCA (the components are linear combinations of the initial variables) that does not exist in SOM. This constraint, as well as the orthogonality between the PCs, can be a drawback for the handling of nonlinear problems. Unlike PCA, the output of SOM is very often in 2D space.

Fort William (II)

In Figure 9.13 the quality map for the Fort William samples is illustrated. It can be seen that, in general, the quality of the mapping is quite good, as in most of the map units the samples are quite close to the respective codebook vectors (the colour is dark). However, clearly, two of the units have not been mapped accurately, with the worst approximation being in units (1,2) and (2,4). The top left map unit



(i) PCA with HCA and k-means (ii) SOM

Figure 9.12: HCA, k-means and SOM clustering of the Aberdeenshire data to two groups.

(in yellow) has also been mapped badly. In addition, the Unified Distance matrix

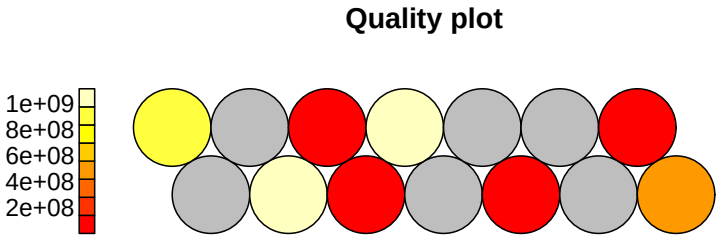


Figure 9.13: Illustration of the quality of mapping with regard to the samples of the Fort William data. The grey units in the quality map mean that there is no sample mapped to those units.

can be used to illustrate the average distance of each map unit to all immediate neighbour units. Units near a cluster boundary are expected to have higher average distances to their neighbouring units (Figure 9.14). The black lines indicate the best four-cluster solution using hierarchical clustering, to allow comparison of the SOM clustering to the HCA solution, however there is no indication from this matrix

that there is a smaller number of clusters in the data according to the 7×2 SOM analysis. A two-dimensional colour-coded hit histogram for the SOM solution can

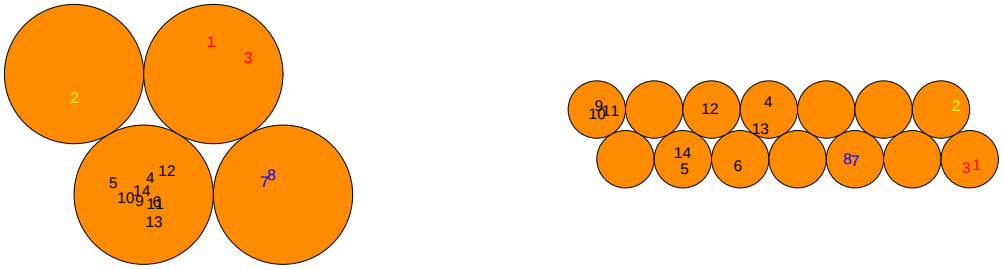
(i) Unified Distance matrix (ii) Hit histogram

Figure 9.14: Unified Distance matrix and Hit histogram for Fort William for the 7×2 grid.

be seen in Figure 9.14. The units with three hits indicate the existence of clusters in the areas surrounding these units, and the concentration of samples around these units is expected to be far larger than elsewhere in the map. The only such unit is (2,1), where the first number corresponds to the row and the second to the column in which the unit is located in the map (with (1,1) being the bottom left-most unit and (2,7) the top right-most unit in the map). The result of the hit histogram for this unit corresponds to this region in the U-matrix (Figure 9.14) with high similarity. The mapping of the samples assigned to each of the four groups provided by the 2×2 map, as well as the colour-coded samples map using the corresponding label colours of the groups obtained by the 2×2 map for the 7×2 grid, can be seen in Figure 9.15. The cluster sizes of the left map are 9, 2, 1 and 2 for groups 1-4 respectively. In the 2×2 grid, group 1 corresponds to the bottom-left unit, while group 4 corresponds to the top-right unit, counting from left to right and then from the bottom to the top row of the map. In the 7×2 grid, the groups corresponding to the 2×2 grid are 1, 2, 4 and 3 from left to right in the map. Comparing the clustering results of the 7×2 map to those of the hierarchical clustering in the U-matrix plot (Figure 9.14), it is clear that cluster 1 at the left and middle-most side of the map, is split into two groups where the first group contains samples 9, 10 and 11, and the second group contains samples 4, 5, 6, 12, 13 and 14. Also, cluster

2 is identical in both solutions, with the other cluster having large differences e.g. clusters 3 and 4 at the right part of the map have been merged to one cluster in the hierarchical clustering solution (see Figure 9.14 and 9.15).

To examine how the variables influence the map and what is the relationship



(i) Samples - 2×2 map (ii) Samples - 7×2 map

Figure 9.15: Illustration of clustering the Fort William data to four groups using SOM.

between each variable and the samples in the data, component planes have been created for a selected number of variables. Again the maps for the ten variables with the largest mean values can be seen in Figure 9.16. Most of these variables are also in the group of ten variables with the largest variance. Upon investigation of the component planes, the following can be deduced:

- The most commonly appearing variables are 265, 491, 190 and 79, i.e. these metabolites appear to be very closely related to almost all units in the map.
- The least common variable is 358. Very few units appear to be associated with this metabolite.
- The samples with consistently high intensity values in all component planes are samples 7 and 8 in cluster 2.

– Samples 5 and 14 in cluster 1, are the samples least associated with the variables.

– None of the component planes is capable of describing the four clusters.

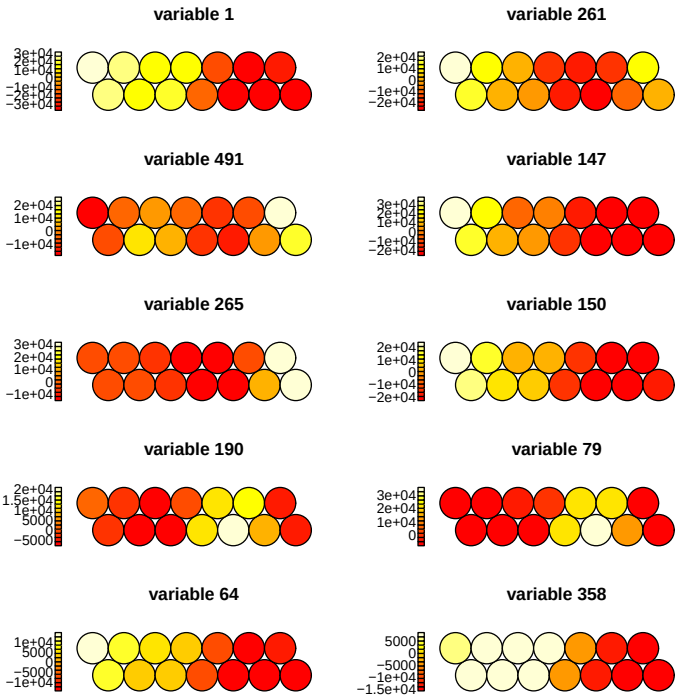
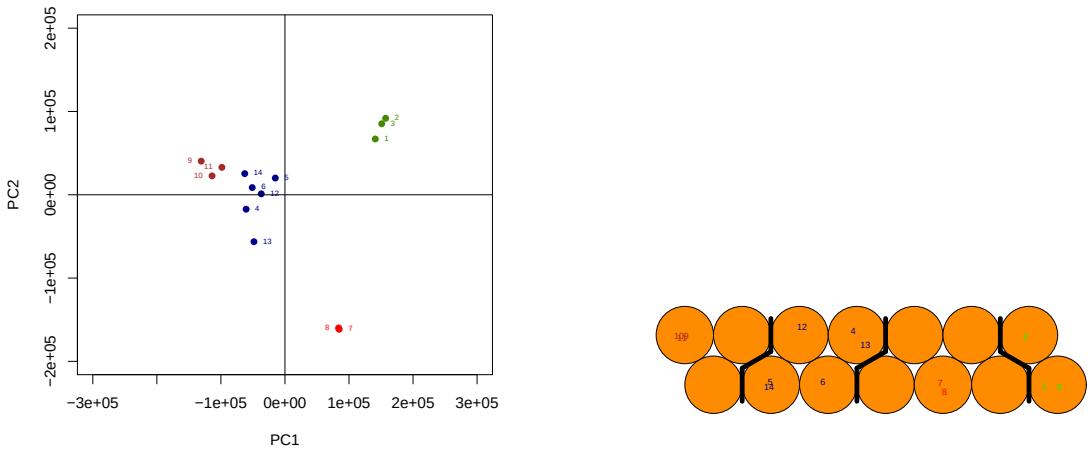


Figure 9.16: Component planes for 10 selected variables among the chemical components of the Fort William samples, labelled by number of the chemical components of Fort William.

A two-dimensional projection of the Fort William data superimposed with the clustering solution derived by the 4-cluster SOM partition can be seen in Figure 9.17. We can compare between the first two principal component scores according to the results from HCA and k-means, which both indicate the same result, and SOM with the partition derived by the selected SOM clustering model. We can see roughly the same proximities and clusters.



(i) PCA with HCA and k-means (ii) SOM

Figure 9.17: HCA, k-means and SOM clustering of the Fort William data to your groups.

Dunblane (III)

In Figure 9.18 the quality map is shown for the 3×3 and 3×2 solutions for the Dunblane samples. It can be seen that, in general, the quality of the 3×3 grid is bad. However, clearly from the 3×2 grid, two of the units have not been mapped accurately, with the worst approximation being in the middle and right bottom map units (in light yellow). The top left map unit (in orange) has also been mapped badly. On the other hand, the quality of the 3×2 grid is quite good. Therefore, the 3×2 map is chosen for further investigation for Dunblane. The Unified Distance matrix for Dunblane can be seen in Figure 9.19. The black lines indicate the four-cluster solution using hierarchical clustering to allow comparison of the SOM clustering to the HCA solution, however there is no indication from this matrix that there is a smaller number of clusters in the data according to the 3×2 SOM analysis. A two-dimensional colour-coded hit histogram for the SOM solution for Dunblane can be seen in Figure 9.19. The units with two hits indicate the existence of clusters in the areas surrounding these units, and the concentration of samples around these

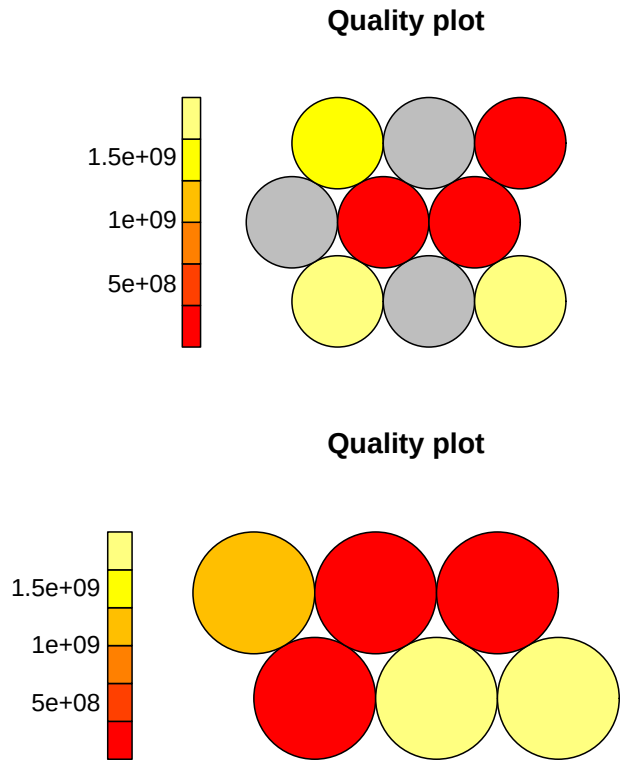


Figure 9.18: Illustration of the quality of two mappings for the samples of the Dunblane data. A grey unit in the quality map means that there is no sample mapped to this unit.

(i) Unified Distance matrix (ii) Hit histogram

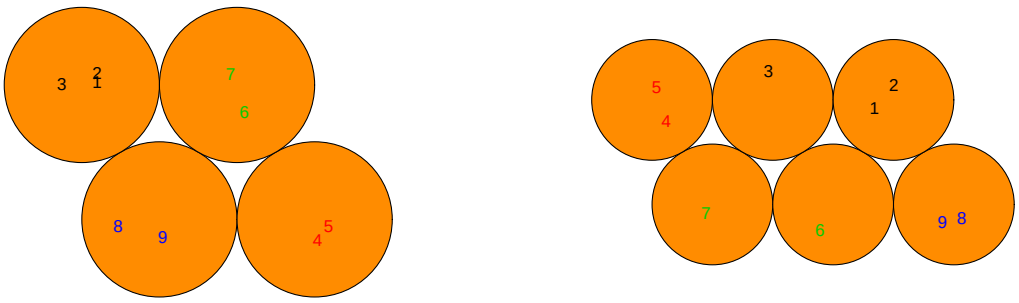
Figure 9.19: Unified Distance matrix and Hit histogram for Dunblane for the 3×2 grid.

units is expected to be far larger than elsewhere in the map. Three such units are (1,2), (1,3) and (2,1), where the first number corresponds to the row and the second to the column in which the unit is located in the map. The results of the hit histogram for units (1,2) and (1,3) correspond to those regions in the U-matrix

with high similarity. The unit (1,2), and units (1,3) and (2,1) of the U-matrix plot are separated by the black lines of the four-cluster HCA partition, indicating the existence of clusters in these areas.

The mapping of the samples assigned to each of the four groups provided by the 2×2 map, as well as the colour-coded samples map (using the corresponding label colours of the groups obtained by the 2×2 map) for the 3×2 grid can be seen in Figure 9.20. The cluster sizes of the left map are 2, 2, 3 and 2 for groups 1-4 respectively. In the 2×2 grid, group 1 corresponds to the bottom-left unit, while group 4 corresponds to the top-right unit counting from left to right and then from the bottom to the top row of the map. In the 3×2 grid, the corresponding groups to the 2×2 grid are 4, 1, 2 and 3 from left to right in the map. Comparing the clustering results of the 3×2 map to those of the hierarchical clustering in the U-matrix plot (Figure 9.19), it is clear that clusters 1, 2, 3 and 4 are identical in both solutions.

To examine how the variables influence the map and what is the relationship



(i) Samples - 2×2 map (ii) Samples - 3×2 map

Figure 9.20: Illustration of clustering the Dunblane data to four groups using SOM.

between each variable and the samples in the data, component planes have been created for a selected number of variables. The maps for the ten variables with the largest mean values can be seen in Figure 9.21. Again most of these variables are also in the group of ten variables with the largest variance. Upon investigation of the component planes, the following can be deduced:

- The most common variables are 19, 33 and 105, i.e. these metabolites appear to be very closely related to almost all units in the map.

- The least common variables are 9 and 21. Very few units appear to be associated with these metabolites, with 9 being the least associated to the map of the two.

- The samples with consistently high intensity values in all component planes are samples 3 in cluster 3, sample 7 in cluster 4 and sample 4 and 5 in cluster 2.

- Samples 8 and 9 in cluster 1 are the samples least associated with the variables.

- None of the component planes is capable of describing the four clusters.

A two-dimensional projection of the Dunblane data superimposed with the clustering solution derived by the 4-cluster SOM partition can be seen in Figure 9.22, to compare between the results from HCA and k-means and SOM with the partition derived by the selected SOM clustering model. We can see that the SOM clustering model has been capable of discriminating the samples in the same way as k-means clustering. Also, there is a small difference between the results from HCA and SOM, where samples 6 and 7 have been merged to one cluster in the SOM solution.

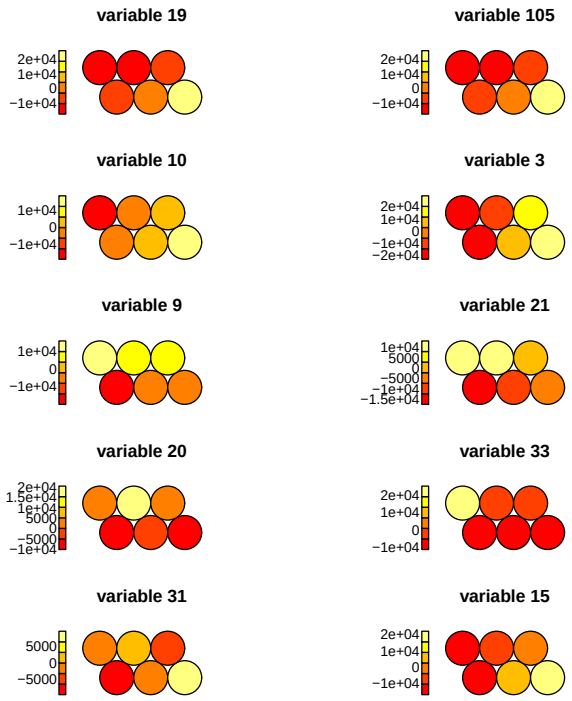
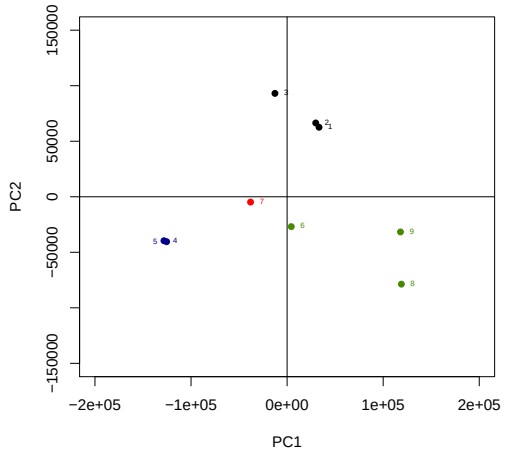


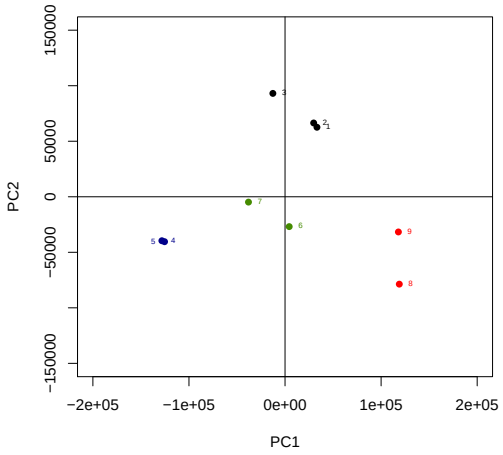
Figure 9.21: Component planes for 10 selected variables in the chemical components of the Dunblane samples, labelled by number of the chemical components of Dunblane.

9.3 Application of SOM to the three data sets combined (IV)

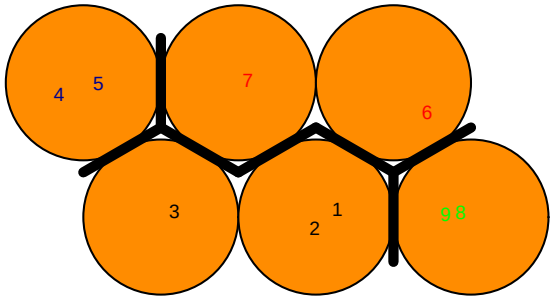
The data used here is the combined data set (Aberdeenshire samples 1:27, Fort William samples 28:41 and Dunblane samples 42:50). Before the analyses, the samples were normalised to eliminate the possibility of any influence on the SOM results by any of the variables due to a variable's large variance. The shape of the SOM grid was chosen to be hexagonal, as before. The map size is important to detect any deviation in the data. If the map size is too small, it might not explain some important differences that should be detected. Conversely, if the map size is too big, the differences are too small (Wilppu, 1997). The lengths of the grid sides can be calculated by setting the ratio of the lengths of the sides similar to that of the two largest eigenvalues of the training data, such that the



(i) PCA with HCA



(ii) PCA with k-means clustering



(iii) SOM

Figure 9.22: HCA, k-means and SOM clustering of the Dunblane data to groups.

product of the lengths is as close as possible to the total number of map units (N_U). In our case, N_U is approximately 35 ($5\sqrt{50}$), whereas the ratio of the two largest eigenvalues of the covariance matrix is approximately $\frac{3}{2}$, thus the grid dimension can be 7×5, to approximate as closely as possible the map size without violating the ratio rule (Park et al., 2006; Vesanto et al., 2000). The training parameters used here are the following: the learning rate initially has a value of 0.05 and decreases monotonically until it reaches the value 0.01 at the end of the number of epochs. The initial radius of the neighbourhood

function is approximately $\frac{2}{3}$ of the estimated map width, to allow for a large part of the map to be updated initially. A sufficient initial radius value is usually to cover $\frac{2}{3}$ of all unit-to-unit distances. Thus, the initial radius for the 7×5 grid is approximately 5 for data set IV. The final value of the radius of data set IV is 0 at the end of the algorithm. The initial representatives were chosen randomly without replacement from the data set IV. A series of runs was performed using the recommended values for the training parameters (Brereton, 2009; Tan and George, 2004). The total number of iterations was chosen in each case to be 500 times the map size of data set IV. One map size was used, specifically the 7×5 map was chosen for further investigation for these combined data. The neighbourhood width function converges to a small number close to 0 after 20000 iterations respectively for the estimated maps (Figure 9.23 shows the convergence for the grid of these data).

In Figure 9.24 the quality map is illustrated where samples 1 to 27 indicate Aberdeenshire, samples 28 to 41 indicate Fort William and samples 42 to 50 indicate Dunblane. It can be seen that, in general, the quality of the mapping is quite good, as in most of the map units the samples are quite close to the respective codebook vectors (the colour is dark). However, clearly, two of the units have not been mapped accurately, with the worst approximation being in the bottom right unit (1,7) (in white) (where the first number corresponds to the row and the second to the column in which the unit is located in the map. For example (1,1) is the bottom left-most unit and (5,7) the top right-most unit in the map). The unit (5,4) (in yellow) has also been mapped badly.

In addition, the Unified Distance matrix is shown in Figure 9.25. The black lines indicate a four-cluster solution using hierarchical clustering, to allow comparison of the SOM clustering to the HCA solution. The units on the top left side of the U-matrix are closer to each other than those on the bottom left side of the matrix, however there is no indication from this matrix that there is a big number of clusters in the data according to the 7×5 SOM analysis. A two-dimensional colour-coded hit histogram for the SOM solution can be seen in Figure 9.25. The units with four or five hits indicate the concentration of sam-

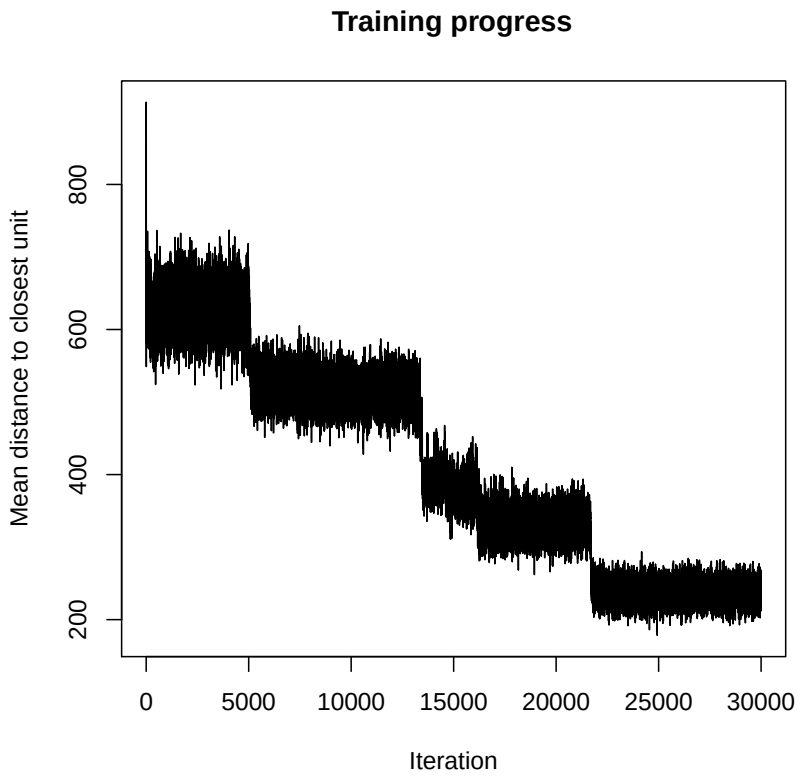


Figure 9.23: Convergence of the neighbourhood width function for the selected map (13×2 grid) of data set IV.

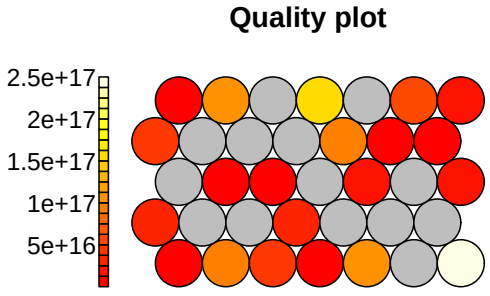
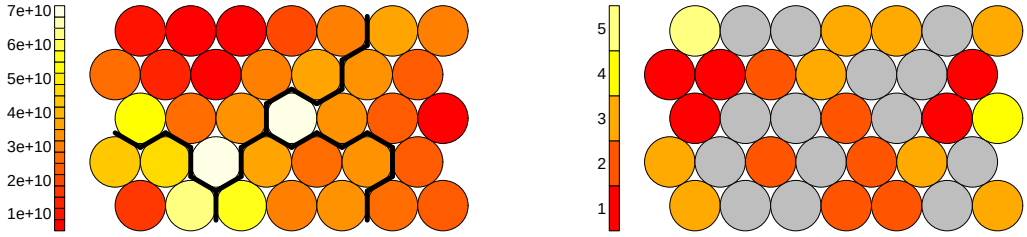


Figure 9.24: Illustration of the quality of mapping for the samples of data set IV. A grey unit in the quality map means that there is no sample mapped to this unit.

ples around these units is expected to be far larger than elsewhere in the map. Two such units are (3,7) and (5,1). The results of the hit histogram for these two units correspond



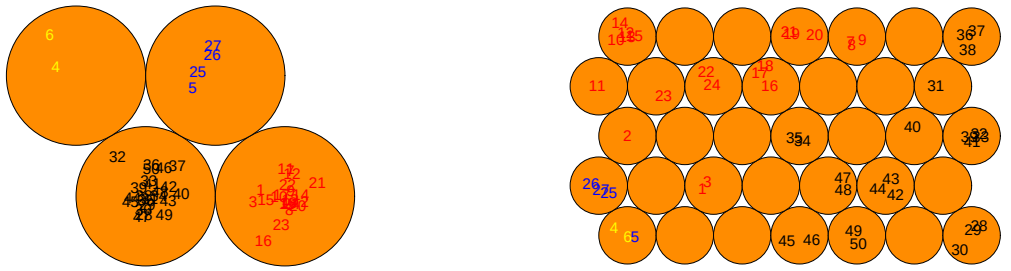
(i) Unified Distance matrix (ii) Hit histogram

Figure 9.25: Unified Distance matrix and Hit histogram for data set IV for the 7×5 grid.

to those regions in the U-matrix with high similarity. In the U-matrix plot (Figure 9.25), as they are separated by the black lines of the four-cluster HCA partition, this indicates the existence of clusters in these areas.

The mapping of the samples assigned to each of the four groups provided by the 2×2 map, as well as the colour-coded samples map (using the corresponding label colours of the groups obtained by the 2×2 map) for the 7×5 grid, can be seen in Figure 9.26. The cluster sizes of the left map are 20, 21, 2 and 4 for groups 1-4 respectively. In the 2×2 grid, group 1 corresponds to the bottom-left unit, while group 4 corresponds to the top-right unit counting from left to right, and then from the bottom to the top row of the map. In the 7×5 grid, the corresponding groups to the 2×2 grid are 3, 4, 1 and 2 from left to right in the map, then from the bottom to the top of the map. Comparing the clustering results of the 7×5 map to those of the hierarchical clustering in the U-matrix plot (Figure 9.25), it is clear that cluster 1 at the right side of the map is split into two clusters that separate Fort William (samples 28:41) and Dunblane (samples 42:50). Also, cluster 2 at the left side of the map is identical in both solutions. The other clusters have large differences, e.g. clusters 3 and 4 at the bottom - left part of the map have been merged to one cluster in the hierarchical clustering solution (in colours yellow and blue).

To examine how the variables influence the map and what is the relationship between each variable and the samples in the data, component planes have been created for se-



(i) Samples - 2×2 map (ii) Samples - 7×5 map

Figure 9.26: Illustration of clustering of data set IV to four groups using SOM.

lected variables. The maps for the ten variables with the largest mean values can be seen in Figure 9.27. Most of these variables are also in the group of ten variables with the largest variance. Upon investigation of the component planes, the following can be deduced:

- The most commonly appearing variables are 435, 106 and 436, i.e. these metabolites appear to be very closely related to almost all units in the map.
- The least common variables are 155, 128 and 569. Few units appear to be associated with these metabolites.
- The samples with consistently high intensity values in all component planes are Fort William samples 28:41.
- Samples 4, 5, 6, 25, 26 and 27 from Aberdeenshire are the samples least associated with the variables.
- None of the component planes is capable of describing the four clusters.

A two-dimensional projection of data set IV superimposed with the clustering solution

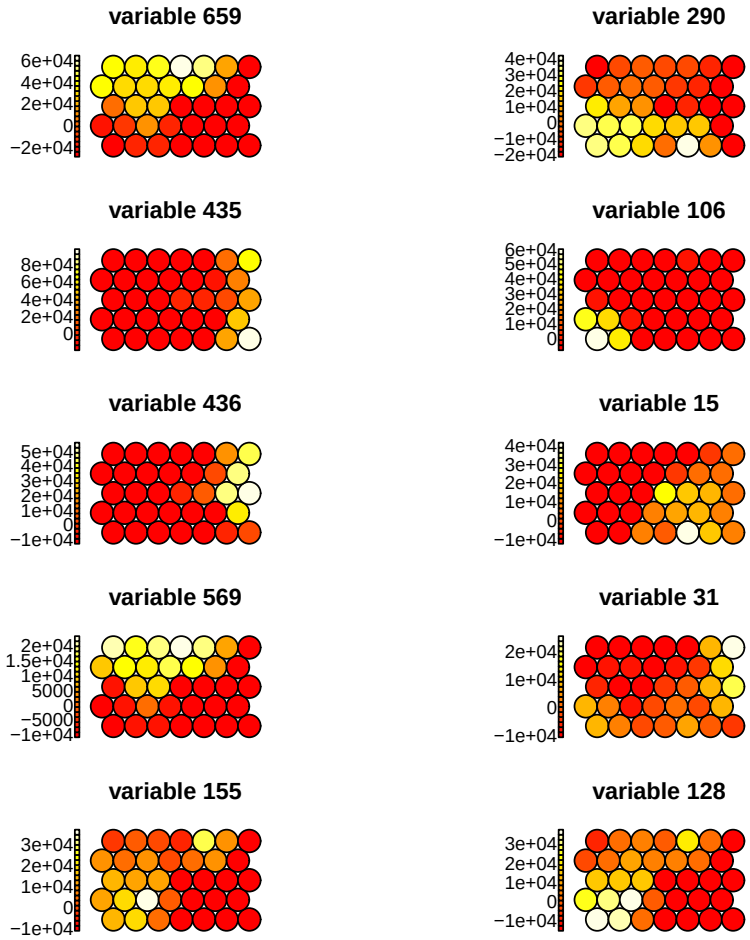
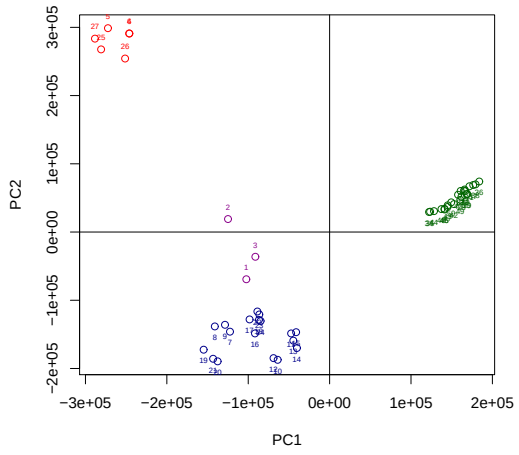
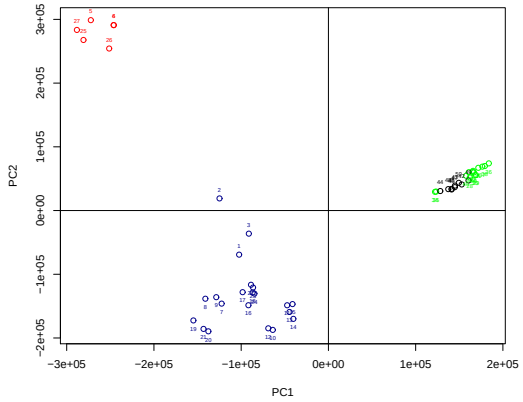


Figure 9.27: Component planes for 10 selected variables among the chemical components of data set IV samples, labelled by numbers of the chemical components for data set IV.

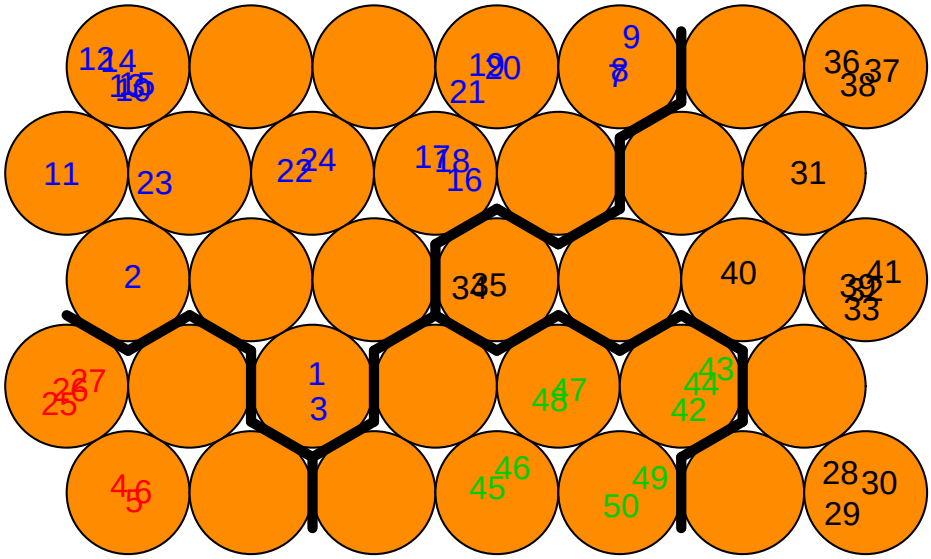
derived by the 4-cluster SOM partition can be seen in Figure 9.28, to compare between the results from HCA and k-means and SOM with the partition derived by the selected SOM clustering model. We can see that the SOM clustering model has been capable of discriminating the samples as in the result of k-means clustering. Also, there is a big difference between the results from HCA and SOM, where samples from Fort William and Dunblane have been merged to one cluster in the HCA solution. Also, the Aberdeenshire data have been separated into three groups in the HCA solution.



(i) PCA with HCA cluster



(ii) PCA with k-means cluster



(iii) SOM

Figure 9.28: HCA, k-means and SOM clustering of data set IV to groups.

9.4 Application of SOM to the Libya data set

The lengths of the grid sides are again calculated by setting the ratio of the lengths of the sides similar to that of the two largest eigenvalues of the training data such that the product of the lengths is as close as possible to the total number of map units (N_U). In this case, N_U is approximately 17 ($5\sqrt{12}$) $>$ 12 (number of samples), therefore, the size of the map is taken as 12 for the Libya data, whereas the ratio of the two largest eigenvalues of the covariance matrix is approximately $\frac{3}{2}$, thus the grid dimension can be 4×3, to approximate as closely as possible the map size without violating the ratio rule (Park et al., 2006; Vesanto et al., 2000). The training parameters are taken as the following: the learning rate initially has a value of 0.05 and decreases monotonically until it reaches the value 0.01 at the end of the number of epochs. The initial radius of the neighbourhood function is approximately $\frac{2}{3}$ of the estimated map width, to allow for a large part of the map to be updated initially. A sufficient initial radius value is usually taken to cover $\frac{2}{3}$ of all unit-to-unit distances. Thus, the initial radius for the 4×3 grid is approximately 3 for the Libya data. The final value of the radius of the Libya data is 0 at the end of the algorithm. The initial representatives were chosen randomly without replacement from the Libya data. A series of runs was performed using the recommended values for the training parameters (Brereton, 2009; Tan and George, 2004). The total number of iterations was chosen in each case to be 500 times the map size of the Libya data. Specifically, the 4×3 map were chosen for further investigation for the Libya data. The neighbourhood width function converges to 0 after 25000 iterations for the estimated maps (Figure 9.29 shows the convergence for the grid of the Libya data).

Figure 9.30 illustrates the quality map. It can be seen that in general the quality of the mapping is quite good, as in most of the map units the samples are quite close (see the red colour). However, clearly, two of the units have not been mapped accurately, with the worst approximation being in the top left unit (1,3), and unit (3,1) (in white) has also

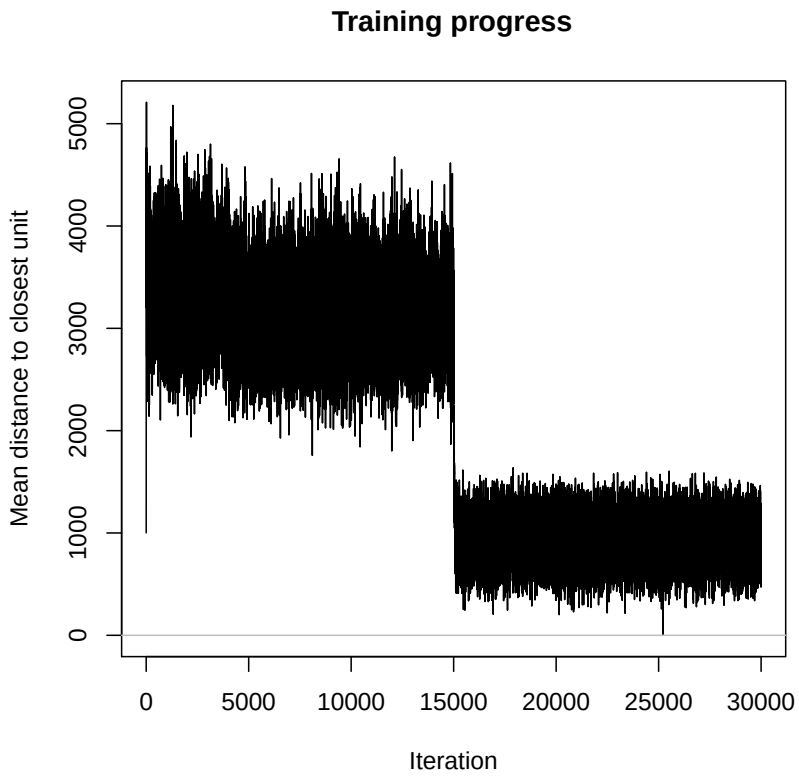


Figure 9.29: Convergence of the neighbourhood width function for the selected map (4×3 grid) of the Libya data.

been mapped badly. In addition, the Unified Distance matrix is shown in Figure 9.31. The black lines indicate the best three-cluster solution using hierarchical clustering, to allow comparison of the SOM clustering to the HCA solution. A two-dimensional colour-coded hit histogram for the SOM solution can also be seen in Figure 9.31. The units with two hits indicate the existence of clusters in the areas surrounding these units, and the concentration of samples around these units is expected to be far larger than elsewhere in the map. Several such units are (1,3), (3,1) and (3,2), which have 2 samples in each unit. In the U-matrix plot (Figure 9.31), as they are separated by the black lines of the three-cluster HCA partition, this indicates the existence of clusters in these areas.

The mapping of the samples assigned to each of the three groups provided by the 3×1

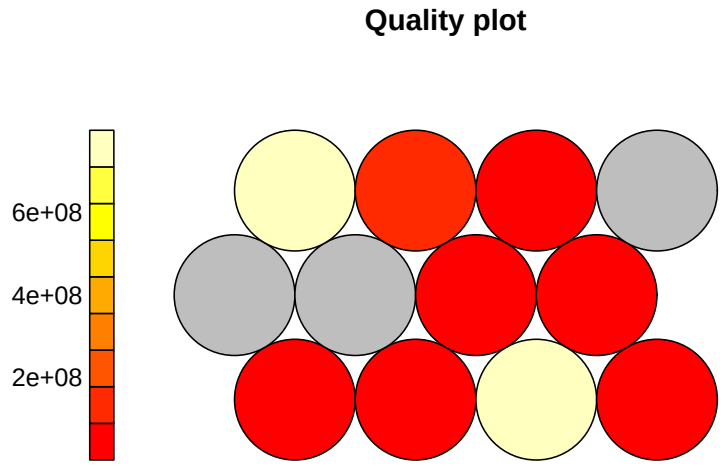
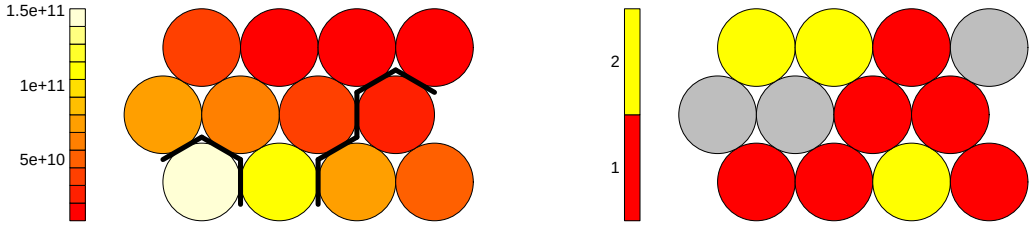


Figure 9.30: Illustration of the quality of mapping for the samples of the Libya data set. A grey unit in the quality map means that there is no sample mapped to this unit.



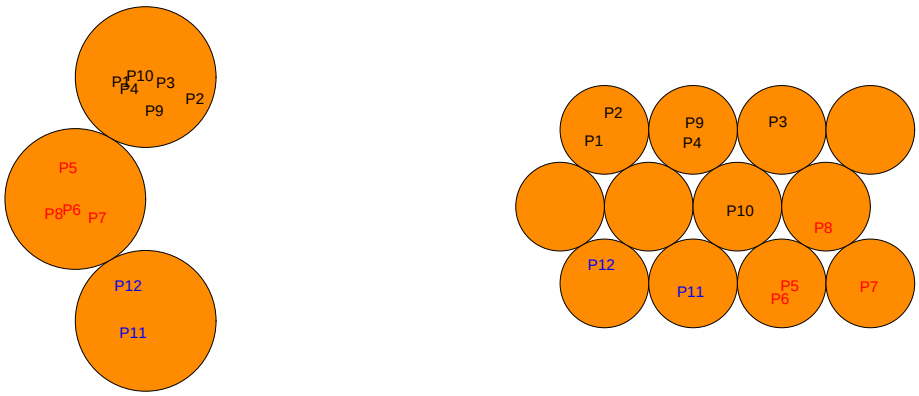
(i) Unified Distance matrix (ii) Hit histogram

Figure 9.31: Unified Distance matrix and Hit histogram for the Libya data set for the 4×3 grid.

map, as well as the colour-coded samples map (using the corresponding label colours of the groups obtained by the 3×1 map) for the 4×3 grid can be seen in Figure 9.32. The cluster sizes of the bottom map are 2, 4 and 6 for groups 1-3 respectively. In the 3×1 grid, group 1 corresponds to the bottom unit, while group 3 corresponds to the upper unit, counting from the bottom to the top of the map. In the 4×3 grid, the corresponding

groups to the 3×1 grid are 1, 2 and 3 from left to right in the map, then from the bottom to the top of the map. Comparing the clustering results of the 4×3 map to those of the hierarchical clustering in the U-matrix plot (Figure 9.31), it is clear that samples P12 and P11 at the left and lower-most side of the map in Figure 9.32 are split into two groups, as P12 is in one group and samples P1, P2, P3, P4, P9, P10 and P11 combine in another group. Also, cluster 2 is identical in both solutions (points marked in red in Figure 9.32).

To examine how the variables influence the map and what is the relationship between



(i) Samples - 1×3 map

(ii) Samples - 4×3 map

Figure 9.32: Illustration of clustering of the Libya data to three groups using SOM.

each variable and the samples in the data, component planes have been created for selected variables. The maps for the ten variables with the largest mean values can be seen in Figure 9.33. Most of these variables are also in the group of ten variables with the largest variance. Upon investigation of the component planes, the following can be deduced:

- The most commonly appearing variables are 105, 122 and 183, i.e. these metabolites appear to be very closely related to almost all units in the map.
- The least common variables are 50 and 166. Very few units appear to be associated with these metabolites.

– The samples with consistently high intensity values in all component planes are samples P11 and P12 in cluster 1, and P1 and P2 in cluster 3.

– Samples P5 and P6 in cluster 2 are the samples least associated with the variables.

– None of the component planes is capable of describing the three clusters.

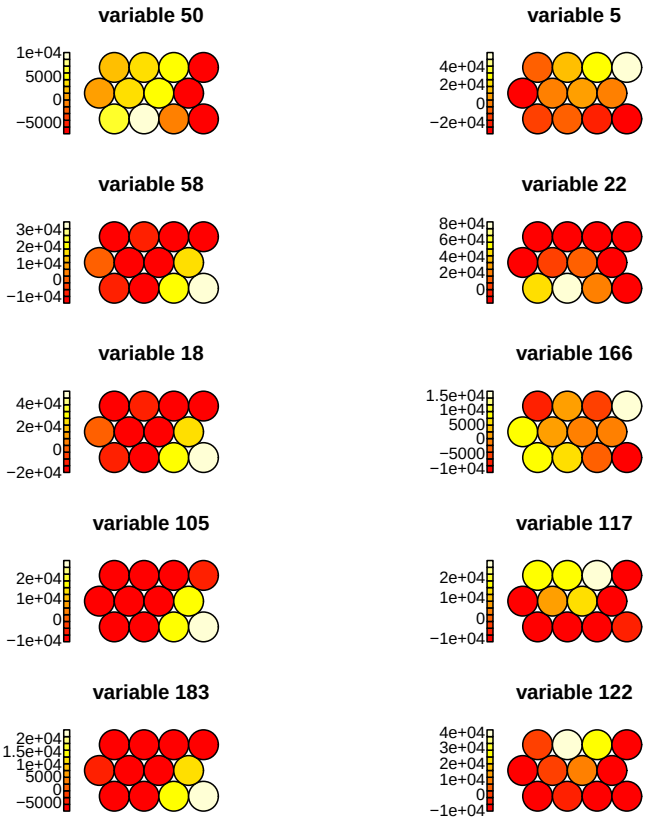
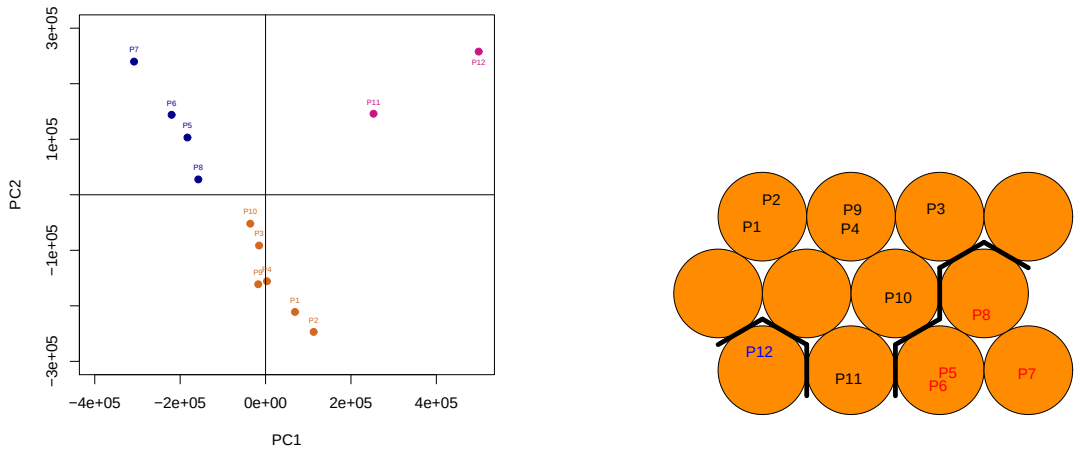


Figure 9.33: Component planes for 10 selected variables among the chemical components of the Libya data samples, labelled by numbers of the chemical components for the Libya data.

A two-dimensional projection of the Libya data superimposed with the clustering solution derived by the 3-cluster SOM partition can be seen in Figure 9.34, to compare between the results from HCA, k-means and SOM with the partition derived by the selected SOM clustering model. We can see that the SOM clustering has a small difference from the results from HCA and k-means, as sample 11 has been merged with the third cluster



(i) PCA with HCA and k-means (ii) SOM

Figure 9.34: HCA, k-means and SOM clustering of the Libya data to groups.

(black colour) in the SOM solution.

9.5 Summary and Conclusions from the SOM Method

A category of clustering algorithms which have not been used widely in metabolomics is that of the competitive learning algorithms. The classic online Self organising maps (SOM) algorithm was chosen, as it has some innovative advantages compared to the other clustering methods. Apart from allowing visualisation of the data in a map-like graph, it provides a range of visualisation tools for assessing the quality of the derived map, such as unified distance matrix plots, hit histograms, quality maps and component planes. One map size was chosen for the data sets for comparison and analysis purposes. The map size was 13×2 for Aberdeenshire, 7×2 for Fort William, 3×2 for Dunblane, 7×5 for data set IV and 4×3 for the Libya data set. The available visualisation tools showed that the quality of the mappings of the data sets I, II, III, IV and Libya were quite good. The sizes of the two clusters for the Aberdeenshire maps were 6 and 21 for clusters 1-2,

respectively, the sizes of the four clusters for the Fort William maps were 9, 2, 1 and 2 for clusters 1-4 respectively, and the sizes of the four clusters for the Dunblane maps were 2, 2, 3 and 2 for clusters 1-4 respectively, the sizes of the four clusters for the data set IV maps were 6, 9, 14 and 21 for clusters 1-4 respectively, and the sizes of the three clusters for the Libya data maps were 2, 8 and 2 for clusters 1-3 respectively. Component planes showed the variables that were very closely related to almost all map units, and also the least commonly associated variables, as well as, the samples that were the most closely associated with the previously mentioned variables.

None of these clustering methods were able to completely discriminate the samples depending on location for data data sets III and Libya. In general, the SOM solution was more like the k-means solution than that of HCA, where these were different, as for data sets III and IV. For the Libya data, the SOM result was slightly different from those of HCA and k-means.

Chapter 10

Case study on data from Europe

After studying two methods (PCA and MDS) to reduce the dimensionality of the data set, cluster analysis (HCA, k-means and SOM) was used to find a natural grouping of samples and compare between methods. Briefly, we will now try to apply the best identified methods in this study to data from the UK (England and Northern Ireland) and Eastern Europe and see the results to find out about this data set. The main objective here is not so much to study the methods but to use them for data analysis and to interpret the results. The data will be explained in Section 10.1, followed by a discussion of the data analysis in Section 10.2, and a conclusion in Section 10.3.

10.1 The European data

Samples of propolis were collected by beekeepers from several of their honey bee colonies, located in many different areas of the UK (England and Northern Ireland), but also a few from elsewhere (Eastern Europe), for comparison (see Figure 10.1). These samples were profiled using liquid chromatography high resolution mass spectrometry, as for the previous data sets used, in Dr David Watson's lab in SIPBS, at the University of Strath-

clyde. The propolis samples contain several hundred compounds, many of which are still

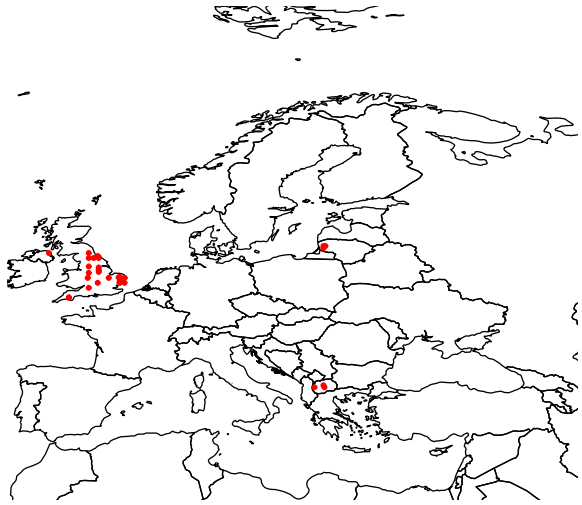


Figure 10.1: Map of Europe, including the locations of the colonies supplying the analysed propolis samples (Map created in R).

unknown structures. The data consists of 35 samples, 30 from various locations in England and Northern Ireland, and 5 from eastern Europe. In the data set, as before, the 285 rows are chromatographic peak areas for putatively identified compounds, while the column headings represent the label for each sample (hive or colony). The code and origin of each sample is shown in Table 10.1. The data were transposed for analysis.

10.2 The analysis of the European data by PCA

Each metabolite in the European data set was column-scaled by mean-centering and Pareto scaling, as was identified in Chapter 4 as being the best approach to pre-treatment, to enable more comparability in the samples. Column-scaling was done by dividing each

Sample Code	Sample Origin from UK
1	Suffolk 4, UK
3	Suffolk 2, UK
4	North Yorkshire 1, UK
5	Northamptonshire 1, UK
6	Essex 1, UK
7	Essex 2, UK
8	Norfolk 1, UK
9	Devon 1, UK
10	Leicstershire 1, UK
11	Leicstershire 2, UK
12	Derbyshire, UK
15	Suffolk 1, UK
16	Suffolk 3, UK
19	Cambridgeshire 1, UK
20	Norfolk 2, UK
21	Northamptonshire 2, UK
22	Cambridgeshire 2, UK
23	North Yorkshire 2, UK
24	Northern Ireland, UK
25	North Yorkshire 3, UK
26	North Yorkshire 4, UK
27	North Yorkshire 5, UK
28	North Yorkshire 6, UK
29	Essex 3, UK
30	Berkshire, UK
31	Midlands, UK
32	Devon 2, UK
33	Buckinghamshire, UK
34	Norfolk 3, UK
35	Norfolk 4, UK
Sample Code	Sample Origin form Europe
2	Bulgaria 1
13	Lithuania 1
14	Lithuania 2
17	Bulgaria 2
18	Bulgaria 3

Table 10.1: The sample origin for the European data set, as it was provided.

element by the square root of the standard deviation of the variable, thus transforming the variables into the same unit of measurement. Since PCA gives the best results for metabolomic data, and is useful for determining the important compounds in the data, it is selected here rather than MDS (see the conclusion of Chapter 6). The Gleason-Staelin statistic and the normalised entropy are calculated using equations (5.6) and (5.9) respectively. The values of the Gleason-Staelin statistic and the normalised entropy are 0.56 and 0.40 respectively. Both statistics confirm that the European data set is suitable for PCA analysis.

After confirming the suitability of the data set for PCA, the next step is to identify the number of principal components to retain. Table 10.2 and Figure 10.2 show the results

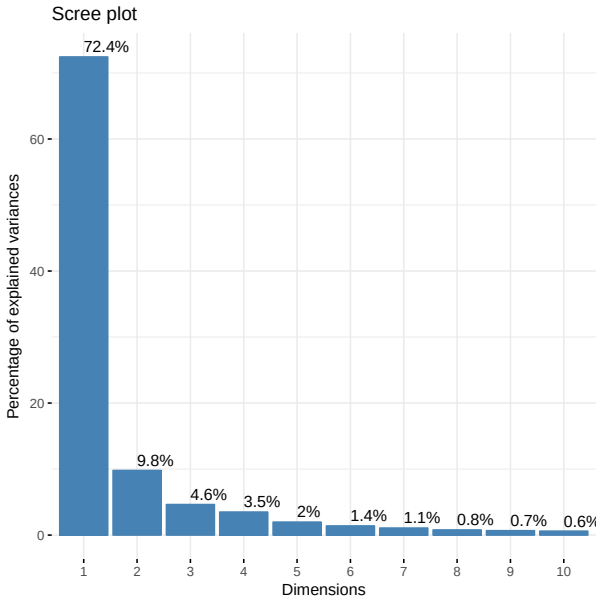


Figure 10.2: Percentages of the total variation in the European data explained by the first ten PCs.

of PCA. These detailed results for the variance of the PCs indicate that no more than two or three components need to be retained for further analysis, as they explain most of the variation in the data set; approximately 72.43%, 82.23% and 86.86% in total for one, two and three PCs respectively. Having identified that the first two PCs are sufficient for

PCs	% of Variance	Cumulative %
PC1	72.43	72.43
PC2	9.79	82.23
PC3	4.63	86.86
PC4	3.49	90.35
PC5	1.97	92.32
PC6	1.39	93.72
PC7	1.07	94.79
PC8	0.79	95.57
PC9	0.67	96.24
PC10	0.60	96.84

Table 10.2: Percentage of total variance explained by, and cumulative percentages of variance for, the first ten PCs of the European data.

further analysis, creating a graphical representation of the data set structure is the next step in the PCA process (see Figure 10.3). Figure 10.3 shows the spread of the propolis samples in PCA space. The samples are broadly spread in terms of PC1 and PC2. The most interesting plot is the score plot for the first two PCs. We can see from Figure 10.3 that the samples from Bulgaria (samples 2, 17 and 18) are close together and also the samples from Lithuania (samples 13 and 14) are close together and at opposite ends of the PC1 axis. There is a spread of samples from the UK on the PC1 axis, likely to be because of the sources of different propolis from many locations in the UK. Also, the samples from Lithuania overlap with samples 4, 10 and 16 from different places in the UK.

As PCA is affected by outliers, it is important to determine whether or not any of these samples are outliers, and this is itself of interest to identify any unusual points in the data. Diagnostic plots using the score distance and the orthogonal distance for the European samples in the data can be seen in Figure 10.4. As Figure 10.4 shows, there are no bad leverage outliers (high score distance and high orthogonal distance). However,

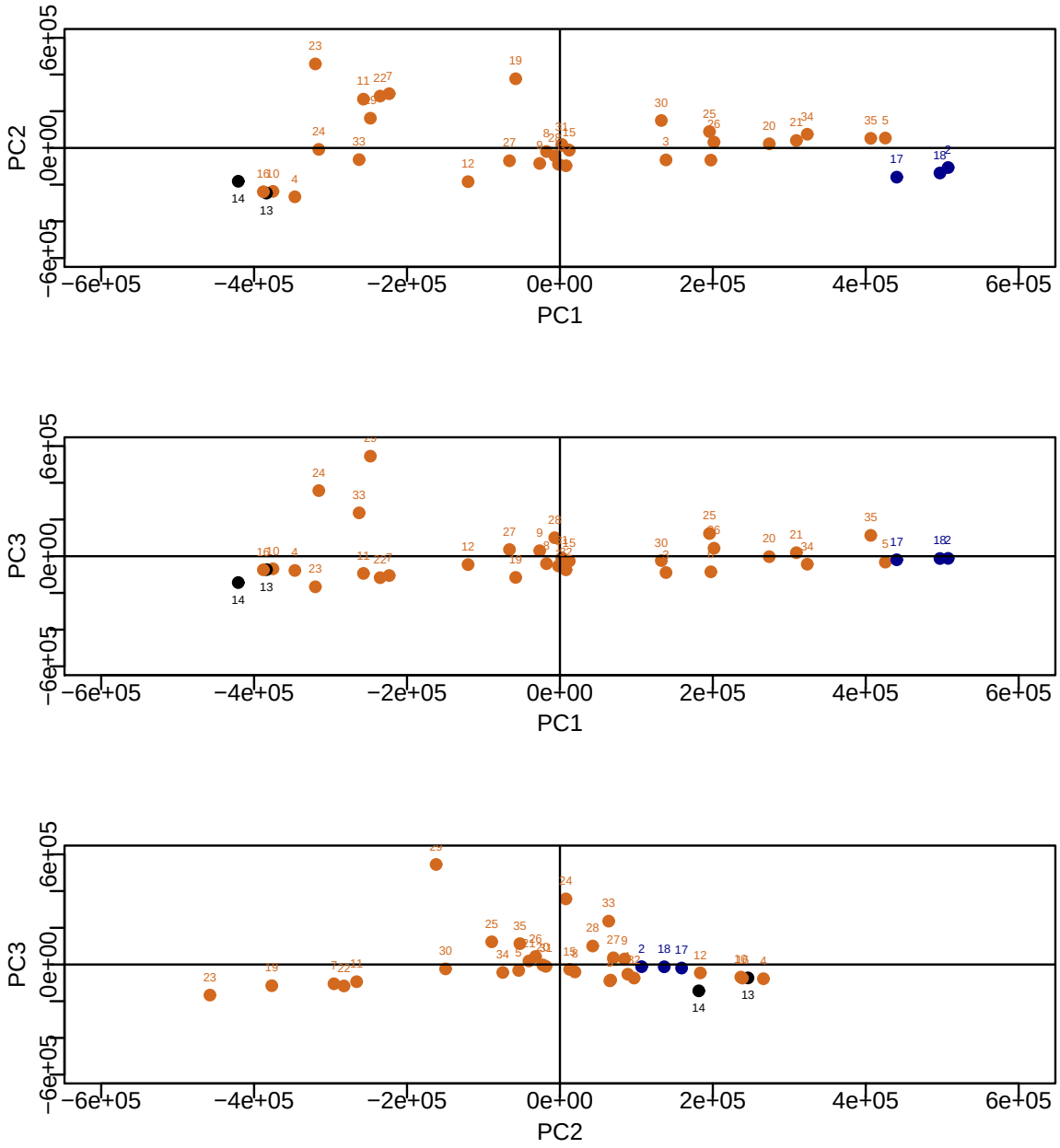


Figure 10.3: Scores plots of the European data for the first three PCs, superimposed with the sample numbers (hives) and the brown colour indicates the samples from the UK, the blue shows the samples from Bulgaria and the black shows samples from Lithuania.

removing samples 2, 10, 13, 14, 16, 17, 18, 23 (detected as outliers) from the data set and re-running the analyses showed that there was no effect from the excluding or inclusion

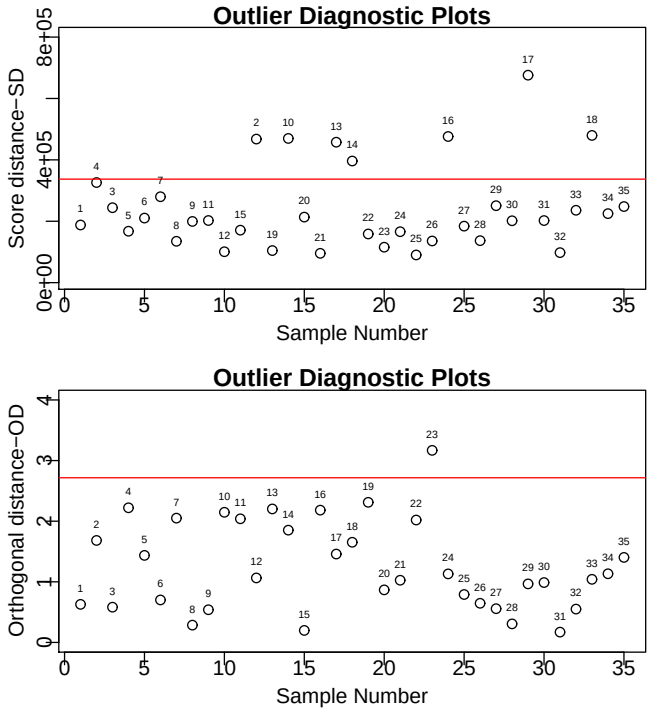


Figure 10.4: Outlier diagnostic plots using the score distance (SD) and the orthogonal distance (OD) for the European data. The numbers in the plots are the numbers of the 35 samples. The horizontal lines in the two plots represent the cutoff values, such that any point above these lines is a leverage point (top plot) or an orthogonal outlier (bottom plot). Table 10.1 shows the location of each sample.

of these samples in the PCA, as the results obtained were similar. Therefore, the original data set of the selected 35 samples can be used for further analysis.

Regarding Figure 10.5 for the European data set there are positive and negative loadings on PC1 and PC2. However, variables 6, 27, 10, 25, 7 and 51 respectively have the highest positive loadings in PC1 (see Appendix, Table G), thus samples from Bulgaria (2, 17 and 18) and samples from the UK numbers 5 and 35 tend to have larger values on these variables, as well as variables 15, 18, 13, 14 and 29 being observed to have the most negative loadings in PC1, thus, the samples from Lithuania (13 and 14) and the samples from the UK numbered 4, 10 and 16 tend to have larger values on these variables.

In the second step, hierarchical agglomerative cluster analysis will be employed to create

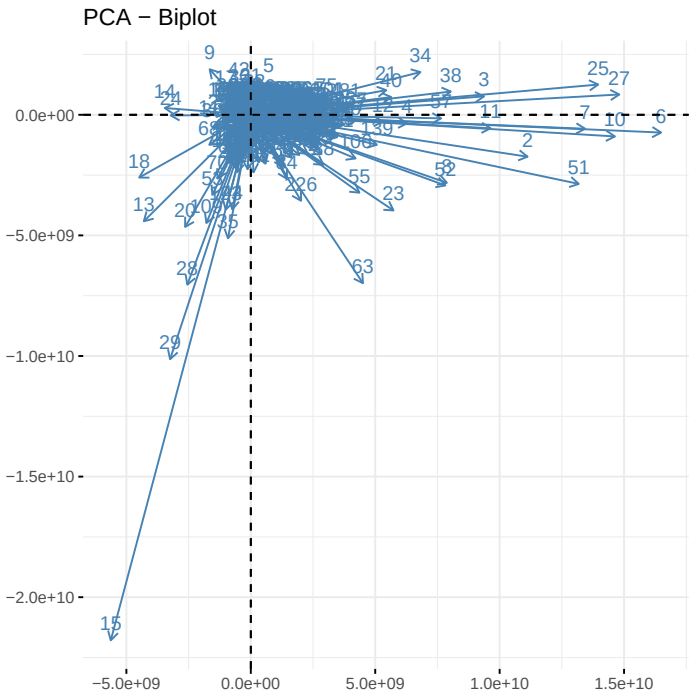


Figure 10.5: Biplot of variables for the first two principal components of the European data set.

a group of samples. From the results of the investigation in Chapter 7, the best combinations of distance and linkage method to use with the propolis data set were identified as Euclidean-Average (for Aberdeenshire and Fort William), Canberra-Average (Dunblane data and Libya) and Manhattan-Average (data set IV). These methods will be compared in order to determine which is most suitable for the European data set.

Before confirming which method gives the best fit of the European data set, the optimal number of clusters should be identified. The command *NbClust* in *R* was used to compare between the number of clusters chosen using 30 indices. From Table 10.3 it is clear that, for all three clustering methods, the optimal number of clusters is 5 for the European data set.

Among all indices for Euclidean-Average
2 proposed 2 as the best number of clusters
3 proposed 3 as the best number of clusters
4 proposed 4 as the best number of clusters
8 proposed 5 as the best number of clusters
2 proposed 8 as the best number of clusters
2 proposed 9 as the best number of clusters
2 proposed 10 as the best number of clusters

Among all indices for Canberra-Average
4 proposed 2 as the best number of clusters
5 proposed 3 as the best number of clusters
1 proposed 4 as the best number of clusters
6 proposed 5 as the best number of clusters
1 proposed 6 as the best number of clusters
3 proposed 7 as the best number of clusters
1 proposed 8 as the best number of clusters
1 proposed 9 as the best number of clusters
1 proposed 10 as the best number of clusters

Among all indices for Manhattan-Average
2 proposed 2 as the best number of clusters
6 proposed 4 as the best number of clusters
7 proposed 5 as the best number of clusters
1 proposed 12 as the best number of clusters
1 proposed 13 as the best number of clusters
3 proposed 14 as the best number of clusters
3 proposed 15 as the best number of clusters

Table 10.3: Comparison between several methods of determining the optimal number of clusters, using hierarchical clustering.

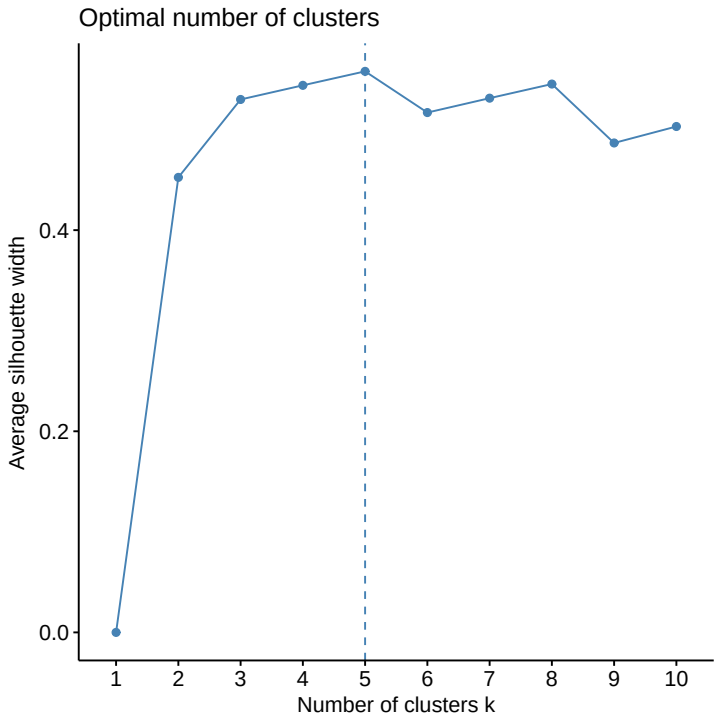
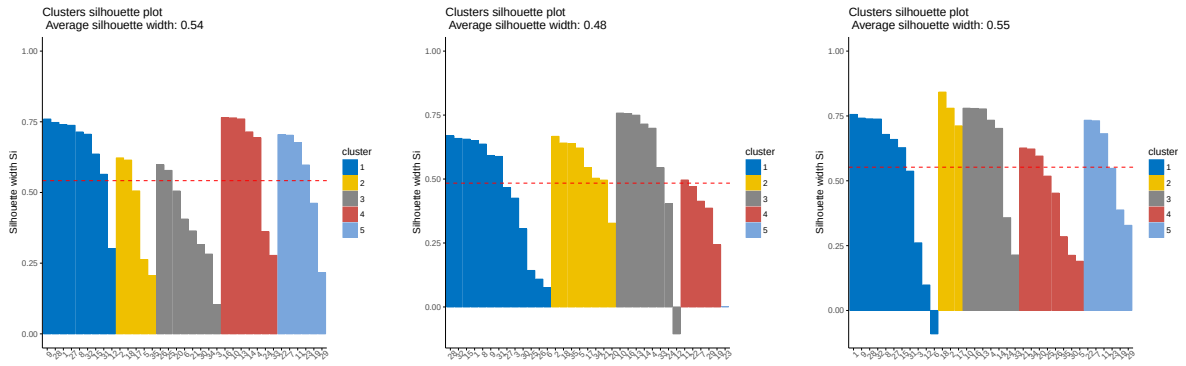


Figure 10.6: Average silhouette widths for partitions of 2-10 clusters for the European data set using the Euclidean-Average method. The optimal number of clusters is indicated by the dashed line.

Figure 10.6 also illustrates the average silhouette widths for ten partitions, from 2-10 clusters for the Euclidean-Average method. It is clear that the optimal number of clusters is 5 for the European data set. In fact, this is true for all 3 clustering methods.
Which results should be retained for further analysis of the European data set will be discussed and determined as follows:

The silhouette plots for a 5 cluster solution from the three clustering methods for the European data set can be seen in Figures 10.7 for the Euclidean-Average, Canberra-Average, and Manhattan-Average methods. The silhouette width values of all samples are depicted in the silhouette plot as bars. The average silhouette widths for clusters 1 to 5 are 0.66, 0.62, 0.44, 0.56 and 0.40 respectively of Euclidean-Average, 0.43, 0.58, 0.61, 0.76 and 0.00 respectively of Canberra-Average method and 0.48, 0.75, 0.55, 0.40 and 0.56 respectively



(i) Euclidean-Average method. (ii) Canberra-Average method. (iii) Manhattan-Average.

Figure 10.7: Silhouette plots for the 5-cluster partition derived by the three clustering methods for the European data set.

of Manhattan-Average. The three clustering methods have an average silhouette width for the entire data of 0.54, 0.48 and 0.55 respectively, which are also the respective silhouette coefficients for the three methods. Therefore, there is a slight difference between them (the best ones are Manhattan-Average and Euclidean-Average with 0.55 and 0.54). In the case of Manhattan-Average, with the highest average silhouette width, there is one misclassified member of a cluster (with negative silhouette width) which is sample 6 in cluster 1. Also, Canberra-Average has misclassified sample 12 in cluster 3. The findings and the information obtained by the silhouette plots indicate that the Euclidean-Average method is best for the European data set, because it does not misclassify any points and also as it has an average silhouette width of 0.54. A dendrogram for the clustering partition derived by the Euclidean-Average method can be seen in Figure 10.8, showing the 5 clusters.

Following HCA, k-means will be used, to compare results of these methods, and the main question is whether k-means gives the same groups of samples as HCA or not. According to the majority rule in Table 10.4, the best number of clusters found by the *R* software is again 5 for k-means. A silhouette plot for the 5 cluster partition from k-means can be seen in Figure 10.9 for the European data. Although the average silhouette width

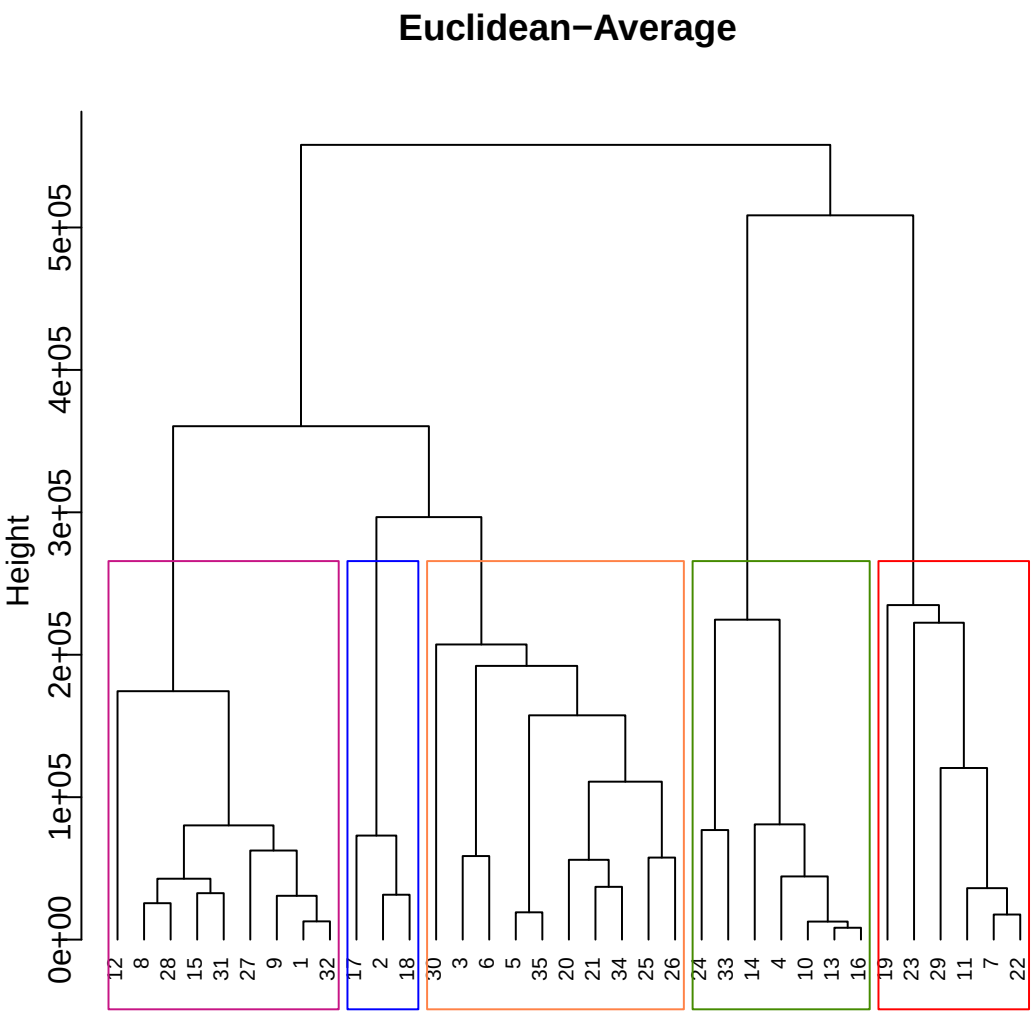


Figure 10.8: Dendrogram for the 5-cluster partition derived by the Euclidean - Average linkage clustering method. The labels at the end-leafs of the tree are the numbers of the samples in the European data set. The 5 clusters are indicated by the coloured rectangles.

for k-means is very slightly smaller than was obtained from HCA clustering (Figure 10.7 for Euclidean-Average, which was the best method for HCA), in the k-means clustering solution there are no misclassified samples (Figure 10.9). The derived optimal 5-cluster k-means partition and average silhouette width are the same as were obtained by the optimal HCA clustering partition. From the above results and Figure 10.10, it is clear that HCA and k-means give a slight difference where samples 5 and 35 in HCA belong

Among all indices for the European data set
4 proposed 2 as the best number of clusters
2 proposed 3 as the best number of clusters
10 proposed 5 as the best number of clusters
3 proposed 7 as the best number of clusters
1 proposed 8 as the best number of clusters
3 proposed 10 as the best number of clusters

Table 10.4: Comparison between several methods of determining the optimal number of k-means clusters.

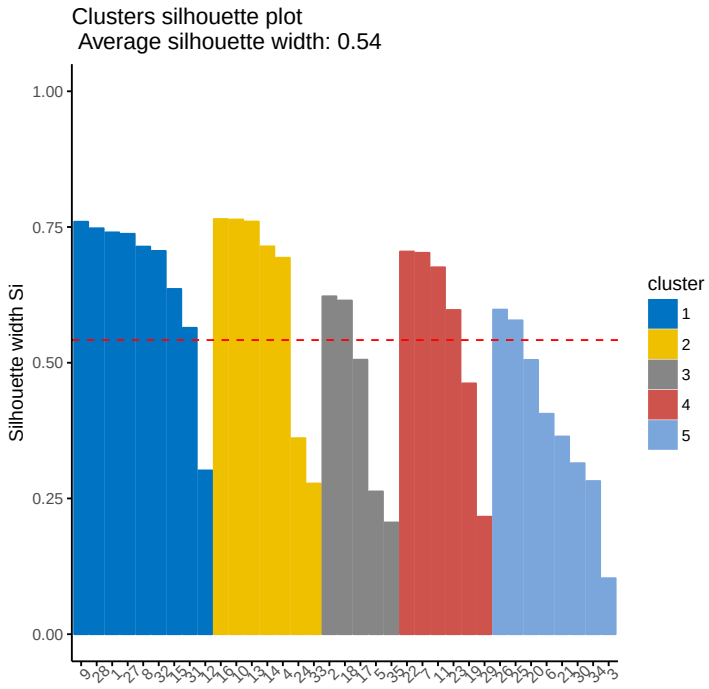


Figure 10.9: Silhouette plot for the 5-cluster partition derived by the k-means clustering method for the European data. The average silhouette widths for the five clusters are 0.66, 0.62, 0.44, 0.56 and 0.39 respectively, and the average silhouette width for the entire data set is 0.54.

to the green cluster and in k-means they merge with the Bulgaria samples 2, 17 and 18. HCA is able to discriminate the samples from Bulgaria in one group. There is a spread of samples from the UK. Also, there is an overlap between samples 13 and 14 from Lithuania and samples 4, 10 and 16 from the UK in both methods.

In general, k-means clustering is faster to run than that of HCA, and requires fewer choices to be made, so is simpler to use.

Self-organising map (SOM) analysis represents another unsupervised multivariate technique suitable for the cluster metabolomics data. SOM will now be used and the results compared with those from HCA and k-means. In this case, N_U is approximately 30 ($5\sqrt{35}$) < 35 (number of samples) (see Section 9.1.5 in chapter 9); therefore, the size of the map is taken as 30 to ensure a better response from the map quality criteria, whereas the ratio of the two largest eigenvalues of the covariance matrix is approximately $\frac{3}{1}$, thus the grid dimension can be 6×5, to approximate the map size as closely as possible without violating the ratio rule.

A two-dimensional projection of the European data superimposed with the clustering solution derived by the 5-cluster SOM partition can be seen in Figure 10.10 (iii). Comparing the clusters in terms of the first two principal component scores (Figure 10.10) according to the results from HCA, k-means and the partition derived by the selected SOM clustering model, it can be seen that the SOM clustering model discriminates the samples much like k-means clustering, and the clusters of samples are the same as for the k-means method. However, SOM is a less easily understood method for general use and so again the k-means method has an advantage.

10.3 Conclusions

This chapter has applied a range of linear and non-linear pattern recognition techniques attempting to identify any natural clustering in the European honey bee propolis data

set and compare between the results of three clustering approaches. More specifically, initially, a linear dimension-reduction technique, i.e. PCA, was applied to the Europe data set, to reduce the dimensionality of the input space of the data to two or three dimensions, making the pattern recognition procedure easier by visualising the data in a lower dimensional representation. Results indicated that two PCs (or dimensions) are sufficient to describe most of the total variation of the Europe data. In general, PCA was useful in obtaining a good picture of the general structure of the Europe data set. The first PCs explain 82.23% of the variation in European data. The samples are broadly spread in terms of PC1 and PC2 because the samples have different composition, so the bees may have used different plant sources (Alotaibi et al., 2019). PC1 discriminates the samples from Bulgaria (2, 17 and 18) and samples 13 and 14 from Lithuania, which are at opposite ends of the PC1 axis.

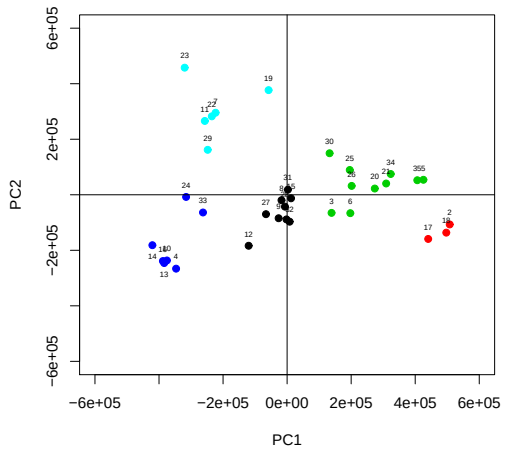
The next step involved application of a range of unsupervised clustering techniques to the Europe data set, to classify if possible, the samples to groups in terms of their location, and particularly exploring the possibility of finding any clustering of the samples. These clustering techniques were hierarchical agglomerative algorithms, optimal partitioning using k-means clustering, and self-organising maps (Chapter 7, 8 and 9).
For hierarchical clustering (HCA), the overall best fitting result found was the cluster partition derived by the Euclidean - Average method for the European data. The Silhouette coefficient was 0.54, 0.48 and 0.55 for methods Euclidean- Average, Canberra- Average and Manhattan- Average respectively. The best method was Euclidean - Average and the cluster sizes were 9, 5, 8, 7 and 6. The best defined cluster for the European data is the red group in Figure 10.10(i). This plot represents regions which are geographically far apart.

The k-means algorithm described in chapter 8 was also used on the Europe data set. As for the HCA results, a 5-cluster partition was the best partition for k-means. The

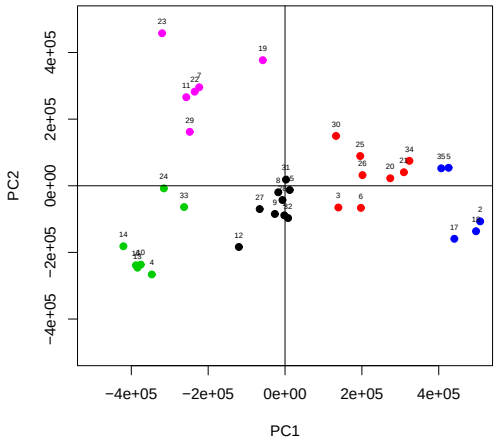
cluster sizes found were 9, 7, 5, 6 and 8. From the HCA and k-means results, HCA and k-means give a slight difference where UK samples 5 and 35 in HCA belong to the green cluster and in k-means they merge with the Bulgaria samples 2, 17 and 18. HCA is able to discriminate the samples from Bulgaria in one group, separate from the rest. There is a spread of samples from the UK in both clustering methods. Also, there is an overlap between samples 13 and 14 from Lithuania and samples 4, 10 and 16 from the UK in both clustering methods.

Finally, a category of clustering algorithms which has not been used widely in metabolomics is that of the competitive learning algorithms. The classic self-organising maps (SOM) algorithm was used here. One map size was chosen for the data set for comparison and analysis purposes. The map size was 6×5 for the Europe data set. The sizes of the 5-clusters for the Europe map were 9, 5, 8, 7 and 6 for clusters 1-5, respectively.

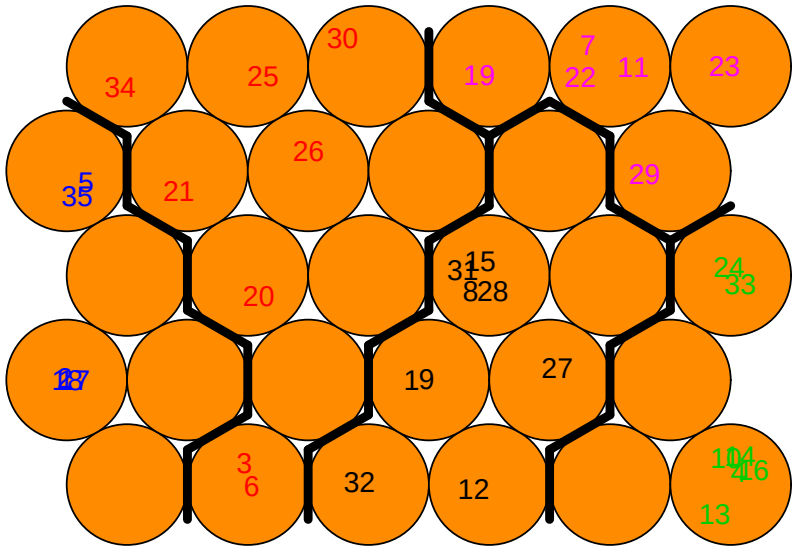
Comparing the three clustering approaches, the k-means and SOM methods gave the same results for the Europe data set with 5 clusters, while k-means is a simpler method to understand. The k-means method is simpler to implement than HCA and gave a similar but not identical solution to HCA. HCA did separate the Bulgaria samples than the rest.



(i) PCA with HCA clustering



(ii) PCA with k-means



(iii) SOM

Figure 10.10: Results of SOM and HCA (using the Euclidean-Average method) and k-means clustering on the European data; the colour coding shows the groups of samples.

Chapter 11

Conclusions and Further work

In this chapter the approaches used are summarised, as are the results, advantages and disadvantages of the methods, and further work is suggested.

11.1 Approaches

This thesis studies a wide range of unsupervised multivariate statistical procedures, which are reviewed. The intention was to study the application of these to metabolomics data from honey bee propolis samples. A review and explanation of MS (an essential analytical technique in metabolomics) is presented. A detailed description is also given of the available pre-processing and pre-treatment methods for metabolomics data. The analysis involved a novel assessment of pre-treatment methods and of statistical clustering methodologies and an evaluation of their appropriateness for examining the MS metabolomics data from propolis. For MS metabolomics data, the methods mostly utilised include PCA, HCA and k-means clustering. Novel methods rarely implemented in analysis of metabolomics data include MDS and SOM, both of which were also examined here. These methods were used on highly multivariate metabolomics data to systematically compare

how useful the methods are for such data and to identify what each method could offer in this context.

11.2 Results

The data used for the analyses was generated by MS and before the application of any statistical technique, the data was pre-processed (by Dr.Watson with his team). In this thesis the data sets were pre-treated in ways such as column-scaling, as well as element transformations of the data matrix. Various scaling and transformation methods were compared on the data, with the conclusion that column-scaling by mean-centring and Pareto scaling of the variables was best. The variables are considerably more correlated, and consequently more suitable for PCA, in the mean-centred and Pareto-scaled data than when using other column-scaling methods.

Two unsupervised exploratory data methods, namely principal components analysis (PCA) and multi-dimensional scaling (MDS), were used initially (the techniques and their results are described in detail in Chapters 5 and 6 respectively), to project the original input space for the data to 2D or 3D output space to facilitate any pattern identification in the data. PCA is restricted to Euclidean space, but on the other hand it allows the investigation of any relationship between variables and samples. MDS can be used with any dissimilarity (or similarity) measure, but it is very difficult to extract any information about variables from the results of the MDS analysis. In general, PCA has been proved to be a useful method for visualisation or dimensionality reduction of the data, with the ability to extract any relationship between samples and variables.

The second data visualisation and data reduction techniques used was the classical MDS method, and the derived 2-dimensional configuration was used as input to Sammon's non-linear mapping (NLM) method. An important advantage of these methods is that the required between-samples distances can be calculated using any dissimilarity (or similar-

ity) measure. Comparing the MDS configurations derived by various distance metrics, it was shown that classical MDS was not capable of giving more information than PCA about the clustering behaviour of the samples. Nevertheless, two MDS configurations, based on the the Euclidean and the Manhattan distances for data sets I, II, III and IV, and on the Euclidean and the maximum distances for the Libya data were retained, with both having better results in locating the cluster patterns for the samples compared to any other distance matrix that had been previously examined and implemented for future analysis as an input to NLM. It was found that the results derived were very consistent with those of the PCA; however, on the whole MDS does not improve any findings of PCA or even add information. From a general point of view, the results of the Euclidean model showed some resemblance to the second-best metric. On data sets I, II, III, IV and Libya, results of the classical MDS were slightly different from NLM. The two mentioned models were capable of successfully reproducing some of the PCA findings, in terms of clustering the data; however, they did not provide further information on the clustering of the samples compared to the two models of NLM.

The data exploration part of the thesis gave good indications concerning the dimensionality reduction and visualisation of the samples. The next step was to apply the data clustering methods. Four different categories of unsupervised classification methods were assessed. More specifically, in the relevant literature for metabolomics data, hierarchical clustering algorithms, optimal partitioning methods and competitive learning algorithms are generally the most popular and suitable unsupervised classification techniques, therefore they were chosen for the clustering analysis of the propolis data.

Hierarchical clustering algorithms such as agglomerative nesting algorithms are important in the area of metabolomics and are commonly used. These methods involve more than one step to establish the clustering patterns of the data, and each sample is assigned to one and only cluster. The data is clustered in the form of a dendrogram, showing the relationships between the samples. The procedure initially assigns one sample per cluster

and ends when all samples are contained in a single group. Four distance measures were used to calculate the distances between the samples, and five agglomerative nesting algorithms were used, so that 20 HCA methods were constructed and their results compared. Among these clustering methods, the best approach was found to be the 2-cluster partition derived by the Euclidean-Average combination for data set I. The 4-cluster solution derived by the Euclidean-Average method for data set II, the 4-cluster partition obtained by the Canberra-Average method for data set III, the 4-cluster separation derived by the Manhattan-Average method for data set IV, and the 3-cluster partition acquired by the Canberra-Average method for the Libya data set, were found to be best. The best linkage method with all data sets was Average linkage. These methods provided the best overall fit to the data sets.

The k-means hard clustering algorithm was used in the analysis of the propolis data sets, also a common clustering approach, although not widely used until now in metabolomics applications. Because of the nature of optimisation, it usually produces tighter clusters than HCA. The results of the analyses showed that the 2, 4, 4, 4 and 3 cluster partitions were the best partitions derived by k-means clustering of data sets I, II, III, IV and Libya respectively, and these were found to be the same partitions as the best HCA clustering partitions, although there were some differences in the discriminating of samples and the silhouette width values of the two methods. Therefore, HCA and k-means have different clustering ability for samples for data sets III and IV.

Competitive learning algorithms form a different category of clustering methods than those previously mentioned. In this case, for each object presented to the algorithm, all the predefined representatives in a set compete with each other, and the winner is the closer representative, using some distance measure, to the object. Consequently, the winner representative is being updated to be closer to the target, with the procedure continuing for all objects until no updates can occur in any representatives. The self-organising maps (SOM) technique, which was used to analyse the propolis data sets, is

one such algorithm. In this case, the representatives are called codebook vectors.

The methods (HCA, k-means and SOM) were able to discriminate the samples according to the replicate analysis where a set of three samples come from the same hive (or colony) for data sets I, II and IV (excluding data set III). Also, SOM discriminates the samples depending on their location, for data set IV. In some cases k-means gave the same results as SOM, for data sets III and IV.

11.3 Advantages and Disadvantages of the methods

11.3.1 Advantages

- Principal components analysis investigates the relationship between variables and samples of data.
- An essential advantage of MDS is that the sample distance required can be calculated by using any similarity or dissimilarity measures.
- An essential advantage of HCA algorithms is that the derived clusters are not restricted to a spherical shape; therefore, depending on the data in question, they might be more useful and flexible than other clustering methods. For example, a single linkage produces non-compact elongated clusters, whereas Ward's method produces compact spherical clusters.
- We do not need to specify in advance the number of clusters required for HCA algorithms.
- The dendrogram produced from HCA can be very useful in understanding the data.
- K-means is computationally faster than HCA when the number of variables is considerable, as in the case of metabolomics data.

- K-means usually produces tighter clusters than HCA, due to the nature of the optimisation, which may be considered to be an advantage.
- An advantage of SOM is that the data can be expressed in a map-like visualization form; however, such a map needs to contain several units exceeding two nodes to accurately describe the data.

11.3.2 Disadvantages

- Multi-dimensional scaling is implemented with similarity or dissimilarity measures; however, it becomes difficult to extract information about the variables from the MDS results already analysed.
- The time complexity for HCA clustering can result in very long computation times, in comparison with efficient algorithms, such as k-means.
- If we have a large data set, HCA can become difficult to interpret in terms of determining a suitable number of clusters from the dendrogram.
- HCA requires more choices than k-means (assuming that the Euclidean distance is used in k-means).
- K-means does require the number of clusters to be specified before starting.
- The 1D (proven) topological ordering property of SOM does not extend to 2D.
- SOM clustering is less transparent than k-means or HCA.

11.4 Further work

There is a lot of possible future work which could be done. The following are some suggestions:

- Look at propolis samples from more geographical areas and compare results with those from this study.
- Look at other types of metabolomics data from other fields such as food (Cevallos et al., 2009).
- Consider Nuclear Magnetic Resonance Spectroscopy (NMR Spectroscopy) data as well as MS data (Chao et al., 2010; Kealey and Haines, 2002). There are various advantages to using NMR data, such as:
    1. It is a non-destructive technique.
    2. After an NMR analysis, the samples can be reused for other analyses.
    3. In "omics" studies involving complex bio-mixtures, measurements can often be made with minimal sample preparation.
    4. NMR can provide detailed information on molecular structure for pure compounds and complex mixtures.
    5. It can provide information on absolute or relative concentrations.
    6. It can be conducted in vivo on whole live organisms, which is useful when metabolic profiling for studies of diseases is required.
    7. It is particularly useful for distinguishing isomers, for obtaining molecular information and for studies of molecular dynamics and compartmentation.
- Further research study of supervised methods could also be done such as using PLS (Partial Least Squares) (Salerno Jr et al., 2017).
- Finally, extending the work done in Chapter 6, other clustering methods could be applied such as fuzzy clustering (Theodoridis and Koutroumbas, 2003).

Appendix

Aberdeenshire			
Variables	PC1	Variables	PC2
28	3.023264e-01	8	2.937449e-01
710	-1.830806e-01	28	-2.842272e-01
177	1.745528e-01	124	2.009967e-01
597	-1.717367e-01	72	1.859650e-01
109	1.648305e-01	104	-1.723078e-01
659	-1.592129e-01	155	1.520515e-01
190	1.558905e-01	295	1.490662e-01
147	1.449312e-01	182	1.472805e-01
730	-1.416703e-01	128	1.357622e-01
262	1.348034e-01	183	1.348234e-01
138	1.345622e-01	586	1.342230e-01
30	1.303075e-01	290	1.175583e-01
140	1.267925e-01	206	1.165936e-01
703	-1.250714e-01	458	1.165124e-01
131	-1.247750e-01	420	1.125211e-01
104	-1.240556e-01	591	1.109521e-01
180	1.236175e-01	226	1.047067e-01
127	1.138927e-01	131	-1.044153e-01
106	1.124810e-01	597	-9.836205e-02
267	1.124453e-01	174	9.138940e-02

Table B: Loadings of top 20 variables, contribution for the first 2 PCs of the Aberdeenshire data set.

Fort William			
Variables	PC1	Variables	PC2
1	2.640640e-01	435	-4.000679e-01
10	2.508412e-01	91	1.991250e-01
147	2.130767e-01	79	1.845410e-01
436	2.121357e-01	261	-1.738810e-01
188	1.929692e-01	413	1.640545e-01
142	-1.864050e-01	265	-1.635933e-01
48	1.678272e-01	318	-1.621488e-01
150	1.604268e-01	354	-1.566707e-01
38	1.489143e-01	17	1.557720e-01
226	1.437981e-01	142	-1.519627e-01
16	1.416007e-01	488	-1.448385e-01
152	1.352637e-01	15	1.300509e-01
41	1.349252e-01	25	1.270345e-01
318	-1.298792e-01	478	-1.256072e-01
488	-1.270778e-01	404	-1.227070e-01
46	1.249067e-01	467	1.213367e-01
67	1.191299e-01	491	-1.192057e-01
405	-1.183279e-01	465	1.148107e-01
437	1.146619e-01	319	-1.137072e-01
192	1.145292e-01	111	1.082623e-01

Table C: Loadings of top 20 variables, contribution for the first 2 PCs of the Fort William data set.

Dunblane			
Variables	PC1	Variables	PC2
1	-3.176318e-01	5	3.184075e-01
2	2.418677e-01	14	1.701281e-01
4	-2.204017e-01	18	1.646092e-01
3	-2.165158e-01	407	-1.540400e-01
290	-2.081985e-01	13	-1.498369e-01
291	-2.081985e-01	2	-1.481781e-01
6	-1.988893e-01	22	1.447486e-01
19	-1.773254e-01	41	-1.446827e-01
105	-1.773254e-01	43	1.384270e-01
15	-1.690997e-01	21	-1.347753e-01
10	-1.484089e-01	6	-1.341368e-01
12	-1.421798e-01	39	1.327072e-01
41	1.417223e-01	9	-1.321058e-01
33	1.412573e-01	35	1.273445e-01
16	-1.371802e-01	62	1.171151e-01
7	-1.358884e-01	32	1.164324e-01
25	1.322734e-01	33	-1.136759e-01
493	1.317611e-01	366	-1.133593e-01
11	-1.279336e-01	12	-1.123830e-01
50	1.257055e-01	50	-1.118863e-01

Table D: Loadings of top 20 variables, contribution for the first 2 PCs of the Dunblane data set.

Data set IV			
Variables	PC1	Variables	PC2
28	-0.34197683	710	-0.21617334
104	-0.20154529	597	-0.20205004
597	-0.14933172	28	0.19463689
710	-0.13673366	659	-0.18736138
586	-0.12922662	104	-0.16948687
659	-0.12631046	177	0.16402134
109	-0.12205404	730	-0.16312456
1854	0.11899703	131	-0.14710513
190	-0.10955477	703	-0.14580545
177	-0.10785221	109	0.14404208
1420	0.10613728	190	0.13911488
703	-0.10458065	147	0.13120155
730	-0.10332095	262	0.12517052
140	-0.10127861	465	-0.12375504
147	-0.09791590	180	0.12016597
290	-0.09639725	30	0.11866233
138	-0.09635535	138	0.11825102
127	-0.09374130	378	-0.11245931
465	-0.09331245	267	0.10932637
479	-0.09215044	106	0.10836259

Table E: Loadings of top 20 variables, contribution for the first 2 PCs of data set IV.

Libya Data set			
Variables	PC1	Variables	PC2
2	0.23442901	2	-0.1892930
3	0.21158097	3	-0.1661720
4	0.19912671	4	-0.1640764
7	0.18380293	7	-0.1511923
11	0.17644545	1	-0.1500795
13	0.17467132	5	0.14580270
19	0.15292024	6	0.14515552
22	0.14864540	11	-0.1446442
1	-0.14126536	13	-0.1440468
29	0.13328688	8	-0.1363142
8	-0.13253284	16	0.13387707
12	-0.12767630	12	-0.1301353
9	-0.11710957	19	-0.1236243
51	0.10911097	17	0.12319929
38	0.10883625	22	-0.12186181
49	0.10695338	9	-0.11732736
10	-0.10677448	18	-0.11187630
18	-0.10649155	10	-0.10742843
63	0.10124756	26	-0.10464477
26	-0.09953632	30	0.10434028

Table F: Loadings of top 20 variables, contribution for the first 2 PCs of the Libya data set.

Libya Data set			
Variables	PC1	Variables	PC2
6	0.31186751	15	-0.65874132
27	0.28011028	29	-0.30636089
10	0.27679186	28	-0.21290451
25	0.26403997	63	-0.21106910
7	0.25407788	35	-0.15477690
51	0.24931476	20	-0.14052841
2	0.21039137	109	-0.13563744
11	0.18225787	13	-0.13343524
3	0.17730861	23	-0.11988440
38	0.15204046	123	-0.11892614
8	0.14904500	44	-0.11856306
52	0.14788887	226	-0.10779035
57	0.14466075	53	-0.10042005
34	0.12894093	55	-0.09762297
4	0.11882182	52	-0.08855174
23	0.10848704	51	-0.08660413
40	0.10679517	8	-0.08521822
15	-0.10644823	84	-0.08020147
21	0.10295879	70	-0.07982559
12	0.10024946	18	-0.07864044

Table G: Loadings of top 20 variables, contribution for the first 2 PCs of the European data set.

Bibliography

Adams, M. J. (2007). *Chemometrics in Analytical Spectroscopy*. Royal Society of Chemistry, Cambridge, UK.

Ahmadi-Nedushan, B., St-Hilaire, A., Bérubé, M., Robichaud, É., Thiémonge, N., and Bobée, B. (2006). A review of statistical methods for the evaluation of aquatic habitat suitability for instream flow assessment. *River Research and Applications*, 22(5):503–523.

Alonso, A., Marsal, S., and Julià, A. (2015). Analytical methods in untargeted metabolomics: state of the art in 2015. *Frontiers in Bioengineering and Biotechnology*, 3:23.

Alotaibi, A., Ebiloma, G. U., Williams, R., Alenezi, S., Donachie, A.-M., Guillaume, S., Igoli, J. O., Fearnley, J., De Koning, H. P., and Watson, D. G. (2019). European propolis is highly active against trypanosomatids including crithidia fasciculata. *Scientific Reports*, 9(1):1–10.

Amanor, K. S. and Pabi, O. (2007). Space, time, rhetoric and agricultural change in the transition zone of Ghana. *Human Ecology*, 35(1):51–67.

Apostolescu, N. and Baran, D. (2016). Sammon mapping for preliminary analysis in hyperspectral imagery. *INCAS Bulletin*, 8(1):13.

Arumugam, M., Raes, J., Pelletier, E., Le Paslier, D., Yamada, T., Mende, D. R., Fer-

nandes, G. R., Tap, J., Bruls, T., Batto, J.-M., et al. (2011). Enterotypes of the human gut microbiome. *Nature*, 473(7346):174.

Bankova, V. (2005). Recent trends and important developments in propolis research. *Evidence based Complementary and Alternative Medicine*, 2(1):29–32.

Bankova, V., Bertelli, D., Borba, R., Conti, B. J., da Silva Cunha, I. B., Danert, C., Eberlin, M. N., I Falcão, S., Isla, M. I., Moreno, M. I. N., et al. (2016). Standard methods for Apis mellifera propolis research. *Journal of Apicultural Research*, 58(2):1–49.

Banskota, A. H., Tezuka, Y., and Kadota, S. (2001). Recent progress in pharmacological research of propolis. *Phytotherapy Research*, 15(7):561–571.

Barwick, V., Langley, J., Mallet, T., Stein, B., and Webb, K. (2006). Best Practice Guide for Generating Mass Spectra. LGC, Teddington, UK.

Bauer, H.-U., Herrmann, M., and Villmann, T. (1999). Neural maps and topographic vector quantization. *Neural Networks*, 12(4-5):659–676.

Beckonert, O., Bollard, M. E., Ebbels, T. M., Keun, H. C., Antti, H., Holmes, E., Lindon, J. C., and Nicholson, J. K. (2003). NMR-based metabonomic toxicity classification: hierarchical cluster analysis and k-nearest-neighbour approaches. *Analytica Chimica Acta*, 490(1-2):3–15.

Beebe, K., Pell, R., and Seasholtz, M. (1998). *Chemometrics: A Practical Guide*. Wiley-Interscience Series on Laboratory Automation. John Wiley and Sons, New York, USA.

Bertelli, D., Papotti, G., Bortolotti, L., Marcazzan, G. L., and Plessi, M. (2012). 1H-NMR simultaneous identification of health-relevant compounds in propolis extracts. *Phytochemical Analysis*, 23(3):260–266.

Besse, P. (1992). PCA stability and choice of dimensionality. *Statistics & Probability Letters*, 13(5):405–410.

Besse, P. and De Falguerolles, A. (1993). Application of resampling methods to the choice of dimension in principal component analysis. In Hardle, W. and Simar, L., editors, *Computer Intensive Methods in Statistics,* Statistics and Computers, pages 167–176. Physica-Verlag, Heidelberg, Germany.

Bino, R. J., Hall, R. D., Fiehn, O., Kopka, J., Saito, K., Draper, J., Nikolau, B. J., Mendes, P., Roessner-Tunali, U., Beale, M. H., et al. (2004). Potential of metabolomics as a functional genomics tool. *Trends in Plant Science*, 9(9):418–425.

Blanck, A., Tedesco, P. A., and Lamouroux, N. (2007). Relationships between life-history strategies of European freshwater fish species and their habitat preferences. *Freshwater Biology*, 52(5):843–859.

Boccard, J., Veuthey, J., and Rudaz, S. (2010). Knowledge discovery in metabolomics: An overview of MS data handling. *Journal of Separation Science*, 33(3):290–304.

Borcard, D., Gillet, F., and Legendre, P. (2011). *Numerical Ecology with R.* Use R! Springer, New York, USA.

Borg, I., Groenen, P. J., and Mair, P. (2017). Applied Multidimensional Scaling and Unfolding. Springer, New York.

Borman, S., Russell, H., and Siuzdak, G. (2003). A mass spec timeline. *Today's Chemist at Work*, September 2003:47–49, USA.

Bouchereau, A., Guenot, P., and Larher, F. (2000). Analysis of amines in plant materials. *Journal of Chromatography B: Biomedical Sciences and Applications*, 747(1):49–67.

Brereton, R. G. (2009). *Chemometrics for Pattern Recognition.* John Wiley and Sons, West Sussex, UK.

Burdock, G. (1998). Review of the biological properties and toxicity of bee propolis (propolis). *Food and Chemical Toxicology*, 36(4):347–363.

Campos, P. M., Praca, F. S. G., and Bentley, M. V. L. B. (2016). Quantification of lipoic acid from skin samples by HPLC using ultraviolet, electrochemical and evaporative light scattering detectors. *Journal of Chromatography B*, 1019:66–71.

Cangelosi, R. and Goriely, A. (2007). Component retention in principal component analysis with application to cDNA microarray data. *Biology Direct*, 2(1):2.

Castillo, S., Gopalacharyulu, P., Yetukuri, L., and Oresic, M. (2011). Algorithms and tools for the preprocessing of LC-MS metabolomics data. *Chemometrics and Intelligent Laboratory Systems*, 108(1):23–32.

Catalan, J., Barbieri, M. G., Bartumeus, F., Bitusik, P., Botev, I., Brancelj, A., Cogalniceanu, D., Manca, M., Marchetto, A., Ognjanova, N., et al. (2009). Ecological thresholds in European alpine lakes. *Freshwater Biology*, 54(12):2494–2517.

Cattell, R. B. (1966). The scree test for the number of factors. *Multivariate Behavioural Research*, 1(2):245–276.

Cebolla, V. L., Membrado, L., Vela, J., and Ferrando, A. C. (1997). Evaporative light-scattering detection in the quantitative analysis of semivolatile polycyclic aromatic compounds by high-performance liquid chromatography. *Journal of Chromatographic Science*, 35(4):141–150.

Cevallos, J. M., Reyes-De-Corcuera, J. I., Etxeberria, E., Danyluk, M. D., and Rodrick, G. E. (2009). Metabolomic analysis in food science: a review. *Trends in Food Science and Technology*, 20(11-12):557–566.

Chao, Z., Liang, Q., Yi-Ming, W., and Guo-An, L. (2010). Integrated development of metabonomics and its new progress. *Chinese Journal of Analytical Chemistry*, 38(7):1060–1068.

Charrad, M., Ghazzali, N., Boiteau, V., Niknafs, A., and Charrad, M. (2014). Package NbClust. *Journal of Statistical Software*, 61:1–36.

Chatfield, C. (2018). Introduction to Multivariate Analysis. Routledge, UK.

Chen, C., Gonzalez, F., and Idle, J. (2007). LC-MS-based metabolomics in drug metabolism. *Drug Metabolism Reviews*, 39(2-3):581–597.

Cheng, Y., Liang, Q., Hu, P., Wang, Y., Jun, F., and Luo, G. (2010). Combination of normal-phase medium-pressure liquid chromatography and high-performance counter-current chromatography for preparation of ginsenoside from panax ginseng with high recovery and efficiency. *Separation and Purification Technology*, 73(3):397–402.

Claeson, P., Tuchinda, F., and Reutrakul, V. (1993). Some empirical aspects on the practical use of flash chromatography and medium pressure liquid chromatography for the isolation of biologically active compounds from plants. *Journal of the ScientiTc Society of Thailand*, 19:73–86.

Coombes, K., Tsavachidis, S., Morris, J., Baggerly, K., Hung, M., and Kuerer, H. (2005). Improved peak detection and quantification of mass spectrometry data acquired from surface-enhanced laser desorption and ionization by denoising spectra with the undecimated discrete wavelet transform. *Proteomics*, 5(16):4107–4117.

Cox, T. and Cox, M. (2001). *Multidimensional Scaling*, volume 88 of Monographs on Statistics and Applied Probability. Chapman and Hall/CRC, Florida, USA.

Crane, E. et al. (1990). *Bees and Beekeeping: Science, Practice and World Resources.* Heinemann Newnes, USA.

Cremer, S., Armitage, S., and Schmid-Hempel, P. (2007). Social immunity. *Current Biology*, 17(16):693–702.

Cuesta, O., Piccinelli, A., and Rastrelli, L. (2012). Tropical propolis: recent advances in chemical components and botanical origin. *Medicinal Plants: Biodiversity and Drugs. 1st ed. New York: CRC Press, Taylor and Francis Group*, pages 209–240.

da Cunha, M., Franchin, M., Galvao, L., de Ruiz, A., de Carvalho, J., Ikegaki, M.,

de Alencar, S., Koo, H., and Rosalen, P. (2013). Antimicrobial and antiproliferative activities of stingless bee melipona scutellaris geopropolis. *BMC Complementary and Alternative Medicine*, 13(1):23.

Daniel, L. (1992). Bootstrap methods in the principal components case. In Proc. Annual Meeting of the American Educational Research Association, San Francisco, USA. AERA.

De Carvalho, F., Lechevallier, Y., and De Melo, F. (2012). Partitioning hard clustering algorithms based on multiple dissimilarity matrices. *Pattern Recognition*, 45(1):447–464.

De Vos, R., Moco, S., Lommen, A., Keurentjes, J., Bino, R. J., and Hall, R. (2007). Untargeted large-scale plant metabolomics using liquid chromatography coupled to mass spectrometry. *Nature Protocols*, 2(4):778–791.

Dettmer, K., Aronov, P., and Hammock, B. (2007). Mass spectrometry-based metabolomics. *Mass Spectrometry Reviews*, 26(1):51–78.

Diamantaras, K. and Kung, S. (1996). *Principal Component Neural Networks: Theory and Applications*. Adaptive and Learning Systems for Signal Processing, Communications and Control. John Wiley and Sons, New York, USA.

Ding, C. and He, X. (2004). K-means clustering via principal component analysis. In Proceedings of the twenty-first International Conference on Machine Learning, page 29. ACM.

Dittenbach, M., Rauber, A., and Merkl, D. (2002). Uncovering hierarchical structure in data using the growing hierarchical self-organizing map. *Neurocomputing*, 48(1-4):199–216.

Duran, A., Yang, J., Wang, L., and Sumner, L. W. (2003). Metabolomics spectral formatting, alignment and conversion tools (MSFACTs). *Bioinformatics*, 19(17):2283–2293.

Dvovravckova, E., Snoblova, M., and Hrdlivcka, P. (2014). Carbohydrate analysis: From sample preparation to HPLC on different stationary phases coupled with evaporative light-scattering detection. *Journal of Separation Science*, 37(4):323–337.

El-Aneed, A., Cohen, A., and Banoub, J. (2009). Mass spectrometry, review of the basics: electrospray, maldi, and commonly used mass analyzers. *Applied Spectroscopy Reviews*, 44(3):210–230.

El-Soud, N. H. A. (2012). Honey between traditional uses and recent medicine. *Macedonian Journal of Medical Sciences*, 5(2):205–214.

Ellis, D., Dunn, W., Griffin, J., Allwood, W., and Goodacre, R. (2007). Metabolic fingerprinting as a diagnostic tool. *Pharmacogenomics*, 8(9):1243–1266.

Everitt, B. (1993). *Cluster Analysis*. Arnold, London, UK.

Everitt, B. and Rabe-Hesketh, S. (1997). *The Analysis of Proximity Data*. Series: Kendall's Library of Statistics 4. Arnold, London, UK.

Ferre, L. (1995). Selection of components in principal component analysis: A comparison of methods. *Computational Statistics and Data Analysis*, 19(6):669–682.

Fiehn, O. (2001). Combining genomics, metabolome analysis, and biochemical modelling to understand metabolic networks. *Comparative and Functional Genomics*, 2(3):155–168.

Fiehn, O. (2002). Metabolomics–the link between genotypes and phenotypes. *Plant Molecular Biology*, 48(1-2):155–171.

Franklin, S., Gibson, D., Robertson, P., Pohlmann, J., and Fralish, J. (1995). Parallel analysis: a method for determining significant principal components. *Journal of Vegetation Science*, 6(1):99–106.

Gardana, C., Scaglianti, M., Pietta, P., and Simonetti, P. (2007). Analysis of the polyphe-

nolic fraction of propolis from different sources by liquid chromatography–tandem mass spectrometry. *Journal of Pharmaceutical and Biomedical Analysis*, 45(3):390–399.

Gemperline, P. (2006). Principal component analysis. *In Practical Guide to Chemometrics*, chapter 4, pages 69-104. Taylor and Francis, New York, Second edition.

Go, E., Uritboonthai, W., Apon, J., Trauger, S., Nordstrom, A., O'Maille, G., Brittain, S., Peters, E., and Siuzdak, G. (2007). Selective metabolite and peptide capture/mass detection using fluorous affinity tags. *Journal of Proteome Research*, 6(4):1492–1499.

Gomez-Casati, D. F., Zanor, M. I., and Busi, M. V. (2013). Metabolomics in plants and humans: applications in the prevention and diagnosis of diseases. *BioMed Research International*, 2013:792527.

Goodacre, R., Broadhurst, D., Smilde, A., Kristal, B., Baker, J., Beger, R., Bessant, C., Connor, S., Capuani, G., Craig, A., Ebbels, T., Kell, D., Manetti, C., Newton, J., Paternostro, G., Somorjai, R., Sjostrom, M., Trygg, J., and Wulfert, F. (2007). Proposed minimum reporting standards for data analysis in metabolomics. *Metabolomics*, 3(3):231–241.

Goodacre, R., Vaidyanathan, S., Dunn, W., Harrigan, G., and Kell, D. (2004). Metabolomics by numbers: acquiring and understanding global metabolite data. *Trends in Biotechnology*, 22(5):245–252.

Gordon, A. (1981). *Classification*. Monographs on Applied Probability and Statistics. Chapman and Hall, London, UK.

Gordon, A. (1996). Hierarchical classification. In Arabie, P., Hubert, L., and De Soete, G., editors, *Clustering and Classification*, pages 65–121. World Scientific Publishing, River Edge, New Jersey.

Green, P. E. (2014). Mathematical Tools for Applied Multivariate Analysis. Academic Press, New Jersey.

Gregoris, E., Fabris, S., Bertelle, M., Grassato, L., and Stevanato, R. (2011). Propolis as potential cosmeceutical sunscreen agent for its combined photoprotective and antioxidant properties. *International Journal of Pharmaceutics*, 405(1-2):97–101.

Griffin, J., Walker, L., Garrod, S., Holmes, E., Shore, R., and Nicholson, J. (2000). NMR spectroscopy based metabonomic studies on the comparative biochemistry of the kidney and urine of the bank vole (Clethrionomys glareolus), wood mouse (Apodemus sylvaticus), white toothed shrew (Crocidura suaveolens) and the laboratory rat. *Comparative Biochemistry and Physiology Part B: Biochemistry and Molecular Biology*, 127(3):357–367.

Griffin, J. L. (2004). The potential of metabonomics in drug safety and toxicology. *Drug Discovery Today: Technologies*, 1(3):285–293.

Groenen, P. and Velden, M. (2004). Multidimensional scaling. Econometric Institute Report EI 2004-155, Econometric Institute, Erasmus University Rotterdam, Rotterdam, The Netherlands.

Gulcin, I., Bursal, E., Sehitoglu, M., Bilsel, M., and Goren, A. (2010). Polyphenol contents and antioxidant activity of lyophilized aqueous extract of propolis from Erzurum, Turkey. *Food and Chemical Toxicology*, 48(8-9):2227–2238.

Han, S. and Park, H. (1995). A study on the preservation of meat products by natural propolis: effect of EEP on protein change of meat products. *Korean Journal of Animal Science*, Korea Republic, 37:551-557.

Harrigan, G., LaPlante, R., Cosma, G., Cockerell, G., Goodacre, R., Maddox, J., Luyendyk, J., Ganey, P., and Roth, R. (2004). Application of high-throughput Fourier-transform infrared spectroscopy in toxicology studies: contribution to a study on the development of an animal model for idiosyncratic toxicity. *Toxicology Letters*, 146(3):197–205.

Hartigan, J. (1975). Clustering Algorithms. John Wiley and Sons, New York, USA.

Hartigan, J. and Wong, M. (1979). Algorithm AS 136: A k-means clustering algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 28(1):100–108.

Hausen, B., Wollenweber, E., Senff, H., and Post, B. (1987). Propolis allergy. *Contact Dermatitis*, 17(3):163–170.

Hebert, P., Masson, M., and Denoeux, T. (2006). Fuzzy multidimensional scaling. *Computational Statistics and Data Analysis*, 51(1):335–359.

Hilario, M., Kalousis, A., Pellegrini, C., and Mueller, M. (2006). Processing and classification of protein mass spectra. *Mass Spectrometry Reviews*, 25(3):409–449.

Hipple, J., Sommer, H., and Thomas, H. A. (1949). A precise method of determining the faraday by magnetic resonance. *Physical Review*, 76(12):1877.

Hong, J., Yang, L., Zhang, D., and Shi, J. (2016). Plant metabolomics: an indispensable system biology tool for plant science. *International Journal of Molecular Sciences*, 17(6):767.

Honkela, T., Pulkki, V., and Kohonen, T. (1995). Contextual relations of words in Grimm tales analyzed by self-organizing map. *Proceedings of ICANN-95, International Conference on Artificial Neural Networks*, 2:3–7.

Horai, H., Arita, M., Kanaya, S., Nihei, Y., Ikeda, T., Suwa, K., Ojima, Y., Tanaka, K., Tanaka, S., Aoshima, K., et al. (2010). Massbank: a public repository for sharing mass spectral data for life sciences. *Journal of Mass Spectrometry*, 45(7):703–714.

Horgan, G. W. (2000). Principal component analysis of random particles. *Journal of Mathematical Imaging and Vision*, 12(2):169–175.

Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika*, 30(2):179–185.

Hothorn, T. and Everitt, B. (2014). A Handbook of Statistical Analyses using R. Chapman and Hall/CRC.

Hu, Q., Noll, R., Li, H., Makarov, A., Hardman, M., and Graham, R. (2005). The orbitrap: a new mass spectrometer. *Journal of Mass Spectrometry*, 40(4):430–443.

Hubert, M., Rousseeuw, P., and Branden, K. (2005). ROBPCA: A new approach to robust principal component analysis. *Technometrics*, 47(1):64–79.

Inui, S., Hosoya, T., Shimamura, Y., Masuda, S., Ogawa, T., Kobayashi, H., Shirafuji, K., Moli, R., Kozone, I., Shin-ya, K., et al. (2012). Solophenols B–D and solomonin: New prenylated polyphenols isolated from propolis collected from the Solomon Islands and their antibacterial activity. *Journal of Agricultural and Food Chemistry*, 60(47):11765–11770.

Ivanauskas, L., Jakstas, V., Radusiene, J., Lukosius, A., and Baranauskas, A. (2008). Evaluation of phenolic acids and phenylpropanoids in the crude drugs. *Medicina*, 44(1):48–55.

Izenman, A. (2008). Modern Multivariate Statistical Techniques. *Regression, Classification and Manifold Learning*. Springer Texts in Statistics. Springer, New York, USA.

Jackson, J. (2003). *A User's Guide to Principal Components*. Wiley Series in Probability and Statistics. Wiley-Interscience, Hoboken, New Jersey.

Jambu, M. (1978). *Classification Automatique Pour L'analyse des Donnees (Tome 1)*. Dunod, Paris, France.

Janzekovic, F. and Novak, T. (2012). PCA–a powerful method for analyze ecological niches. In Principal Component Analysis - Multidisciplinary Applications. IntechOpen, London.

Jin, H., Shum, W., Leung, K., and Wong, M. (2004). Expanding self-organizing map for data visualization and cluster analysis. *Information Sciences*, 163(1-3):157–173.

Johnson, K., Wright, B., Jarman, K., and Synovec, R. (2003). High-speed peak matching algorithm for retention time alignment of gas chromatographic data for chemometric analysis. *Journal of Chromatography A*, 996(1):141–155.

Jolliffe, I. (2011). *Principal Component Analysis*. Springer, Berlin, Heidelberg, pages 1094–1096.

Kalteh, A., Hjorth, P., and Berndtsson, R. (2008). Review of the self-organizing map (SOM) approach in water resources: Analysis, modelling and application. *Environmental Modelling and Software*, 23(7):835–845.

Kang, J. (2012). Principles and applications of LC-MS/MS for the quantitative bioanalysis of analytes in various biological samples. *In Tandem Mass Spectrometry-Applications and Principles*, IntechOpen, London.

Kasiotis, K., Anastasiadou, P., Papadopoulos, A., and Machera, K. (2017). Revisiting Greek propolis: Chromatographic analysis and antioxidant activity study. *PloS One*, 12(1): e0170077.

Kaski, S. (1997). Data exploration using self-organizing maps. *Acta Polytechnica Scandinavica: Mathematics, Computing and Management in Engineering Series no. 82.* The Finnish Academy of Technology.

Kassambara, A. (2017). *Practical Guide to Cluster Analysis in R: Unsupervised Machine Learning*, volume 1. STHDA, UK.

Katajamaa, M., Miettinen, J., and Oresic, M. (2006). MZmine: toolbox for processing and visualization of mass spectrometry based molecular profile data. *Bioinformatics*, 22(5):634–636.

Katajamaa, M. and Oresic, M. (2007). Data processing for mass spectrometry-based metabolomics. *Journal of Chromatography A*, 1158(1):318–328.

Kaufman, L. and Rousseeuw, P. J. (2009). *Finding Groups in Data: An Introduction to Cluster Analysis*, volume 344. John Wiley and Sons, New York, USA.

Kealey, D. and Haines, P. (2002). *Analytical Chemistry.* Instant notes. BIOS Scientific Publishers, Oxford, UK.

Keun, H. (2006). Metabonomic modeling of drug toxicity. *Pharmacology and Therapeutics*, 109(1-2):92–106.

Kissinger, P. (2002). Analytical chemistry. *Clinical Chemistry*, 48(12):2303–2303.

Kiviluoto, K. (1996). Topology preservation in self-organizing maps. In *Proceedings of the IEEE International Conference on Neural Networks*, volume 1, pages 294–299.

Kohonen, T. (1990). The self-organizing map. *Proceedings of the IEEE*, 78(9):1464–1480.

Korfmacher, W. (2005). Foundation review: Principles and applications of LC-MS in new drug discovery. *Drug Discovery Today*, 10(20):1357–1367.

Krauss, M., Singer, H., and Hollender, J. (2010). LC–high resolution MS in environmental analysis: from target screening to the identification of unknowns. *Analytical and Bioanalytical Chemistry*, 397(3):943–951.

Krzanowski, W. and Marriott, F. (1995). *Multivariate Analysis.* Kendall's Library of Statistics, Volume 2. Arnold, London, UK.

Kuropatnicki, A., Szliszka, E., and Krol, W. (2013). Historical aspects of propolis research in modern times. *Evidence-Based Complementary and Alternative Medicine*, 2013(1):964149.

Kvalheim, O., Brakstad, F., and Liang, Y. (1994). Preprocessing of analytical profiles in the presence of homoscedastic or heteroscedastic noise. *Analytical Chemistry*, 66(1):43–51.

Legendre, L. and Legendre, P. (1998). *Numerical Ecology, volume 20 of Developments in*

Environmental Modelling, second English edition. Elsevier Science B.V., Amsterdam, The Netherlands.

Legendre, P. and Legendre, L. (2012). *Numerical Ecology*, volume 24. Elsevier, The Netherlands.

Li, Y. and Pan, F. (2013). Application of improved SOM neural network in manufacturing process quality control. *Proceedings of the 2nd International Conference on Computer Science and Electronics Engineering. Paris: Atlantis Press*, pages 1154–1157.

Lindon, J., Holmes, E., and Nicholson, J. (2001). Pattern recognition methods and applications in biomedical magnetic resonance. *Progress in Nuclear Magnetic Resonance Spectroscopy*, 39:1–40.

Lindon, J., Holmes, E., and Nicholson, J. (2006). Metabonomics techniques and applications to pharmaceutical research and development. *Pharmaceutical Research*, 23(6):1075–1088.

Lindon, J., Nicholson, J., and Holmes, E. (2011). Metabolic profiling: Applications in plant science. *The Handbook of Metabonomics and Metabolomics*, pages 443–487, Elsevier, Amsterdam.

Lithio, A. and Maitra, R. (2018). An efficient k-means-type algorithm for clustering datasets with incomplete records. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 11(6):296–311.

Lloyd, S. (1982). Least squares quantization in PCM. *IEEE Transactions on Information Theory*, 28(2):129–137.

Looney, P. (2012). Quantification of Characterization of Biologically Active Components of Actaea Racemosa L.(black Cohosh) for Identifying Desirable Plants for Cultivation. PhD Thesis. Western Carolina University.

Lotfy, M. (2006). Biological activity of bee propolis in health and disease. *Asian Pacific Journal of Cancer Prevention*, 7(1):22–31.

Lu, L.-C., Chen, Y.-W., and Chou, C.-C. (2005). Antibacterial activity of propolis against staphylococcus aureus. *International Journal of Food Microbiology*, 102(2):213–220.

Lukasova, A. (1979). Hierarchical agglomerative clustering procedure. *Pattern Recognition*, 11(5-6):365–381.

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. In Le Cam, L. and Neyman, J., editors, *Mathematical Statistics and Probability, Proceedings of the fifth Berkeley symposium, June 21 - July 18, 1965*, volume I: Theory of Statistics, pages 281–297. Statistical Laboratory, University of California, USA. University of California Press, USA.

Makinen, V., Soininen, P., Forsblom, C., Parkkonen, M., Ingman, P., Kaski, K., Groop, P., and Korpela, M. (2008). 1H NMR metabonomics approach to the disease continuum of diabetic complications and premature death. *Molecular Systems Biology,* 4 (Article number 167). Available at: http://onlinelibrary.wiley.com/doi/10.1038/msb4100205/pdf. Last visited on 2019-03-18.

Marcucci, M. (1995). Propolis: chemical composition, biological properties and therapeutic activity. *Apidologie*, 26(2):83–99.

Mariey, L., Signolle, J., Amiel, C., and Travert, J. (2001). Discrimination, classification, identification of microorganisms using FTIR spectroscopy and chemometrics. *Vibrational Spectroscopy*, 26(2):151–159.

Midorikawa, K., Banskota, A., Tezuka, Y., Nagaoka, T., Matsushige, K., Message, D., Huertas, A., and Kadota, S. (2001). Liquid chromatography–mass spectrometry analysis of propolis. *Phytochemical Analysis*, 12(6):366–373.

Miljkovic, D. (2017). *Brief Review of Self-Organizing Maps.* In 2017 40th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), pages 1061–1066.

Miyagi, A., Takahashi, H., Takahara, K., Hirabayashi, T., Nishimura, Y., Tezuka, T., Kawai-Yamada, M., and Uchimiya, H. (2010). Principal component and hierarchical clustering analysis of metabolites in destructive weeds; polygonaceous plants. Metabolomics, 6(1):146–155.

Mizuno, M. (1989). Food packaging materials containing propolis as a preservative. *Japanese Patent No. JP Ol*, 243(974):89.

Mullner, D. (2011). Modern hierarchical, agglomerative clustering algorithms. *arXiv preprint arXiv:1109.2378.*

Myatt, G. (2007). *Making Sense of Data: A Practical Guide to Exploratory Data Analysis and Data Mining.* John Wiley and Sons, New Jersey, USA.

Neme, A. and Miramontes, P. (2005). Statistical properties of lattices affect topographic error in self-organizing maps. In *Artificial Neural Networks: Biological Inspirations - ICANN 2005.* Lecture Notes in Computer Science, pages 427–432. Springer, Heidelberg, Berlin.

Nielsen, J. and Jewett, M. C. (2007). *Metabolomics: A Powerful Tool in Systems Biology*, volume 18. Springer Science & Business Media, New York.

Nielsen, J. and Oliver, S. (2005). The next wave in metabolome analysis. *Trends in Biotechnology*, 23(11):544–546.

Oldiges, M., Noack, S., and Paczia, N. (2013). Metabolomics in biotechnology (microbial metabolomics). *Metabolomics in Practice: Successful Strategies to Generate and Analyze Metabolic Data*, pages 379–391, Wiley Online Library.

Olive, D. J. (2017). Principal Component Analysis. Springer, USA.

Oliveira, A., Franca, H., Kuster, R., Teixeira, L., and Rocha, L. (2010). Chemical composition and antibacterial activity of Brazilian propolis essential oil. *Journal of Venomous Animals and Toxins including Tropical Diseases*, 16(1):121–130.

Park, Y., Tison, J., Lek, S., Giraudel, J., Coste, M., and Delmas, F. (2006). Application of a self-organizing map to select representative species in multivariate analysis: A case study determining diatom distribution patterns across France. *Ecological Informatics*, 1(3):247–257.

Peres, P., Jackson, D., and Somers, K. (2005). How many principal components? Stopping rules for determining the number of non-trivial axes revisited. *Computational Statistics and Data Analysis*, 49(4):974–997.

Petrova, A., Popova, M., Kuzmanova, C., Tsvetkova, I., Naydenski, H., Muli, E., and Bankova, V. (2010). New biologically active compounds from Kenyan propolis. *Fitoterapia*, 81(6):509–514.

Polani, D. (1999). On the optimization of self-organizing maps by genetic algorithms. In Oja, E. and Kaski, S., editors, *Kohonen Maps*, pages 157–169. Elsevier, Amsterdam, The Netherlands.

Pölzlbauer, G. (2004). Survey and comparison of quality measures for self-organizing maps. In WDA 2004, pages 67-82. Elfa Academic Press, Kosice. Vortrag: Workshop on Data Analysis, Vysoke Tatry, Slovakia, 2004-06-24 - 2004-06-27.

Prelorendjos, A. (2014). Multivariate Analysis of Metabonomic Data. PhD Thesis, University of Strathclyde.

Qureshi, K., Khan, R., and Osman, G. (2014). Chemoprevention of oral cancer by Saudi Arabian propolis: An overview. *Asian Journal of Pharmaceutical Technology and Innovation*, 2347:8810.

R Core Team (2013). R: A Language and Environment for Statistical Computing. R

Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/.

Raghukumar, R., Vali, L., Watson, D., Fearnley, J., and Seidel, V. (2010). Antimethicillin-resistant staphylococcus aureus (MRSA) activity of 'Pacific propolis' and isolated prenylflavanones. *Phytotherapy Research*, 24(8):1181–1187.

Roessner, U. and Bowne, J. (2009). What is metabolomics all about? *Biotechniques*, 46(5):363.

Rousseeuw, P. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65.

Ryan, D. and Robards, K. (2006). Metabolomics: the greatest omics of them all? *Analytical Chemistry-Columbus*, 78(23):7954–7958.

Salas, R., Moreno, S., Allende, H., and Moraga, C. (2007). A robust and flexible model of hierarchical self-organizing maps for non-stationary environments. *Neurocomputing*, 70(16-18):2744–2757.

Salerno Jr, S., Mehrmohamadi, M., Liberti, M. V., Wan, M., Wells, M. T., Booth, J. G., and Locasale, J. W. (2017). Rrmix: A method for simultaneous batch effect correction and analysis of metabolomics data in the absence of internal standards. *PloS one*, 12(6):e0179530.

Sammon, J. (1969). A nonlinear mapping for data structure analysis. *IEEE Transactions on Computers*, 100(5):401–409.

Scalbert, A., Brennan, L., Fiehn, O., Hankemeier, T., Kristal, B., Ommen, B., Pujos, E., Verheij, E., Wishart, D., and Wopereis, S. (2009). Mass-spectrometry-based metabolomics: limitations and recommendations for future progress with particular focus on nutrition research. *Metabolomics*, 5(4):435.

Seidel, V., Peyfoon, E., Watson, D., and Fearnley, J. (2008). Comparative study of

the antibacterial activity of propolis from different geographical and climatic zones. *Phytotherapy Research*, 22(9):1256–1263.

Seltmann, G., Voigt, W., and Beer, W. (1994). Application of physico-chemical typing methods for the epidemiological analysis of salmonella enteritidis strains of phage type 25/17. *Epidemiology and Infection*, 113(3):411–424.

Sforcin, J., J, F., Lopes, C., Bankova, V., and Funari, S. (2000). Seasonal effect on Brazilian propolis antibacterial activity. *Journal of Ethnopharmacology*, 73(1-2):243–249.

Sforcin, J. M. and Bankova, V. (2011). Propolis: is there a potential for the development of new drugs? *Journal of Ethnopharmacology*, 133(2):253–260.

Sharaf, M., Illman, D., and Kowalski, B. (1986). Chemometrics, volume 82. John Wiley & Sons, New York, USA.

Shlens, J. (2003). A tutorial on principal component analysis: derivation, discussion and singular value decomposition. Available at: http://www.cs.princeton.edu/picasso/mats/pca-tutorial-intuition.pdf. Last visited on 2020-02-01.

Silva, B. and Marques, N. (2007). A hybrid parallel SOM algorithm for large maps in data-mining. Available at: http://ssdi.di.fct.unl.pt/ nmm/mypapers/sm2007.pdf. Last visited on 2018-09-24.

Smedsgaard, J. and Nielsen, J. (2005). Metabolite profiling of fungi and yeast: from phenotype to metabolome by ms and informatics. *Journal of Experimental Botany*, 56(410):273–286.

Smith, C., Want, E., Maille, G., Abagyan, R., and Siuzdak, G. (2006). XCMS: processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification. *Analytical Chemistry*, 78(3):779–787.

Smith, R., Ventura, D., and Prince, J. (2015). LC-MS alignment in theory and practice: a comprehensive algorithmic review. *Briefings in Bioinformatics*, 16(1):104–117.

Sokal, R. and Rohlf, F. J. (1995). Assumptions of analysis of variance. Sokal, R and Rohlf, F. J (editors). *Biometry: the Principles and Practice of Statistics in Biological Research*, pages 392–450, W. H. Freeman and Co., New York, USA.

Sturm, M., Bertsch, A., Gropl, C., Hildebrandt, A., Hussong, R., Lange, E., Pfeifer, N., Schulz, O., Zerck, A., and Reinert, K. (2008). Openms–an open-source software framework for mass spectrometry. *BMC Bioinformatics*, 9(1):163.

Suna, T., Salminen, A., Soininen, P., Laatikainen, R., Ingman, P., Mäkelä, S., Savolainen, M., Hannuksela, M., Jauhiainen, M., Taskinen, M., Kaski, K., and Ala-Korpela, M. (2007). 1H NMR metabonomics of plasma lipoprotein subclasses: elucidation of metabolic clustering by self-organising maps. *NMR in Biomedicine*, 20(7):658–672.

Tamilselvi, R., Sivasakthi, B., and Kavitha, R. (2015). A comparison of various clustering methods and algorithms in data mining. *International Journal of Multidisciplinary Research and Development*, 2(5):32–36.

Tan, H. and George, S. (2004). Investigating learning parameters in a standard 2-D SOM model to select good maps and avoid poor ones. In Webb, G. I. and Yu, X., editors, *AI 2004: Advances in Artificial Intelligence*. Lecture Notes in Computer Science, pages 425–437, Heidelberg, Berlin. Springer. 17th Australias Joint Conference on Artificial Intelligence, Cairns, Australia, December 4-6, 2004.

Taner, M. (1997). Kohonen's self organizing networks with 'conscience'. Technical report, Rock Solid Images.

Theodoridis, G., Gika, H., and Wilson, I. (2011). Mass spectrometry-based holistic analytical approaches for metabolite profiling in systems biology studies. *Mass Spectrometry Reviews*, 30(5):884–906.

Theodoridis, G., Gika, H., and Wilson, I. (2013). LC-MS-Based nontargeted metabolomics. *Metabolomics in Practice: Successful Strategies to Generate and Analyze Metabolic Data*, pages 3:93–115, Wiley Online Library.

Theodoridis, S. and Koutroumbas, K. (2003). *Pattern Recognition*, second edition. Academic Press, Elsevier, San Diego, USA.

Tinsley, H. and Tinsley, D. (1987). Uses of factor analysis in counseling psychology research. *Journal of Counseling Psychology*, 34(4):414.

Tomita, M. and Nishioka, T. (2006). *Metabolomics: The Trontier of Systems Biology*. Springer Science & Business Media, USA.

Toreti, V., Sato, H., Pastore, G., and Park, Y. (2013). Recent progress of propolis for its biological and chemical compositions and its botanical origin. *Evidence-Based Complementary and Alternative Medicine*, Hindawi Publishing Corporation.

Tsugawa, H., Cajka, T., Kind, T., Ma, Y., Higgins, B., Ikeda, K., Kanazawa, M., VanderGheynst, J., Fiehn, O., and Arita, M. (2015). MS-DIAL: data-independent MS/MS deconvolution for comprehensive metabolome analysis. *Nature Methods*, 12(6):523–526.

Vaher, M. and Koel, M. (2003). Separation of polyphenolic compounds extracted from plant matrices using capillary electrophoresis. *Journal of Chromatography A*, 990(1):225–230.

van den Berg, R., Hoefsloot, H., Westerhuis, J., Smilde, A., and van der Werf, M. (2006). Centering, scaling, and transformations: improving the biological information content of metabolomics data. *BMC Genomics*, 7(1):142.

Varmuza, K. and Filzmoser, P. (2016). *Introduction to Multivariate Statistical Analysis in Chemometrics*. CRC Press, Boca Raton, Florida, USA.

Velicer, W. (1976). Determining the number of components from the matrix of partial correlations. *Psychometrika*, 41(3):321–327.

Vesanto, J. (1999). SOM-based data visualization methods. *Intelligent Data Analysis*, 3(2):111–126.

Vesanto, J., Himberg, J., Alhoniemi, E., and Parhankangas, J. (2000). SOM toolbox for Matlab 5. Technical report A57, Helsinki University of Technology, Espoo, Finland.

Villmann, T., Der, R., Herrmann, M., and Martinetz, T. (1997). Topology preservation in self-organizing feature maps: Exact definition and measurement. *IEEE Transactions on Neural Networks*, 8(2):256–266.

Volpi, N. (2004). Separation of flavonoids and phenolic acids from propolis by capillary zone electrophoresis. *Electrophoresis*, 25(12):1872–1878.

Volpi, N. and Bergonzini, G. (2006). Analysis of flavonoids from propolis by on-line HPLC–electrospray mass spectrometry. *Journal of Pharmaceutical and Biomedical Analysis*, 42(3):354–361.

Wagh, V. (2013). Propolis: a wonder bees product and its pharmacological potentials. *Advances in Pharmacological Sciences*, 2013, Article ID 308249. Hindawi Publishing Corporation.

Wang, D., Ressom, H., Musavi, M., and Domnisoru, C. (2002). Double self-organizing maps to cluster gene expression data. In Verleysen, M., editor, *ESSAN 2002*. pages 45–50. 10th European Symposium on Artificial Neural Networks, Bruges, Belgium, 24-26 April 2002.

Wang, Y., Yang, C., Mathee, K., and Narasimhan, G. (2005). Clustering using adaptive self-organizing maps (ASOM) and applications. In Sunderam, V., van Albada, G. D., Sloot, P., and Dongarra, J., editors, *Computational Science ICCS 2005*. Lecture Notes in Computer Science, pages 944–951, Springer. Heidelberg, Berlin. 5th International Conference, Atlanta, May 22-25, 2005.

Want, E., O'Maille, G., Smith, C., Brandon, T., Uritboonthai, W., Qin, C., Trauger,

S., and Siuzdak, G. (2006). Solvent-dependent metabolite distribution, clustering, and protein extraction for serum profiling with mass spectrometry. *Analytical Chemistry*, 78(3):743–752.

Watson, D., Peyfoon, E., Zheng, L., Lu, D., Seidel, V., Johnston, B., Parkinson, J., and Fearnley, J. (2006). Application of principal components analysis to 1H-NMR data obtained from propolis samples of different geographical origin. *Phytochemical Analysis: An International Journal of Plant Chemical and Biochemical Techniques*, 17(5):323–331.

Webb, K., Bristow, T., Sargent, M., and Stein, B. (2004). Methodology for accurate mass measurement of small molecules. Best Practice Guide, LGC Limited, Teddington, UK.

Weber, P., Hamburger, M., Schafroth, N., and Potterat, O. (2011). Flash chromatography on cartridges for the separation of plant extracts: rules for the selection of chromatographic conditions and comparison with medium pressure liquid chromatography. *Fitoterapia*, 82(2):155–161.

Weckwerth, W. and Morgenthal, K. (2005). Metabolomics: from pattern recognition to biological interpretation. *Drug Discovery Today*, 10(22):1551–1558.

Wei, X., Sun, W., Shi, X., Koo, I., Wang, B., Zhang, J., Yin, X., Tang, Y., Bogdanov, B., Kim, S., et al. (2011). MetSign: a computational platform for high-resolution mass spectrometry-based metabolomics. *Analytical Chemistry*, 83(20):7668–7675.

Williams, C. (2002). On a connection between kernel PCA and metric multidimensional scaling. *Machine Learning*, 46(1-3):11–19.

Williams, D. and Fleming, I. (1995). *Spectroscopic Methods in Organic Chemistry*, fifth edition. McGraw-Hill, Berkshire, England.

Wilppu, E. (1997). Neural networks: an exploratory data analysis of logistics performance.

Technical report, Turku School of Economics and Business Administration, Turku, Finland.

Wishart, D. (2008). Applications of metabolomics in drug discovery and development. *Drugs in R & D*, 9(5):307–322.

Wold, S., Sjöström, M., and Eriksson, L. (2001). PLS-regression: a basic tool of chemometrics. *Chemometrics and Intelligent Laboratory Systems*, 58(2):109–130.

Wollenweber, E., Hausen, B., and Greenaway, W. (1990). Phenolic constituents and sensitizing properties of propolis, poplar balsam and balsam of Peru. *Bulletin de Liaison-Groupe Polyphenols*, 15:112–120.

Worley, B. and Powers, R. (2013). Multivariate analysis in metabolomics. *Current Metabolomics*, 1(1):92–107.

Xavier, N. and Rauter, A. (2008). Sugars containing α, β-unsaturated carbonyl systems: synthesis and their usefulness as scaffolds in carbohydrate chemistry. *Carbohydrate Research*, 343(10-11):1523–1539.

Xi, Y. and Rocke, D. (2008). Baseline correction for NMR spectroscopic metabolomics data analysis. *BMC Bioinformatics*, 9(1):324.

Xian, F., Hendrickson, C., and Marshall, A. (2012). High resolution mass spectrometry. *Analytical Chemistry*, 84(2):708–719.

Xu, R. and Wunsch, D. (2005). Survey of clustering algorithms. *IEEE Transactions on Neural Networks*, 16(3):645–678.

Young, C. and Dolan, J. (2003). Success with evaporative light-scattering detection. *LCGC North America*, 21(2):120–128.

Zhang, Z., Chen, S., and Liang, Y. (2010). Baseline correction using adaptive iteratively reweighted penalized least squares. *Analyst*, 135(5):1138–1146.

Zwick, W. and Velicer, W. (1986). Comparison of five rules for determining the number of components to retain. *Psychological Bulletin*, 99(3):432.