Soil Characterization via Methods of Functional Data Analysis

Andrew John Gilliland

Submitted in fulfilment of the requirements for the Degree of Master of Science in Statistics

School of Mathematics and Statistics

College of Science and Engineering

University of Glasgow

June 2017

Abstract

Soil is a fundamental natural resource which is relied upon globally for its vital ecological and economic functions. It is important for many reasons including the production of food, support of wildlife and in supporting the mitigation of global warming. With an increasing world population, tremendous pressure is placed on the worlds natural resources. In order to keep up with the agricultural needs of a growing global population, soil management and monitoring practices need to be put in place. However, the standard procedures for monitoring soil quality are prohibitively expensive and slow, with an additional hazard to the environment through use of harmful chemicals. Thus, there has been a widespread interest into the use of diffuse reflectance spectroscopy for the prediction of physical and chemical properties in the soil. This method of recording soil data is cost-effective, rapid, requires minimal sample preparation and does not involve the use of hazardous chemicals. Currently, multivariate analyses such as partial least squares regression are routinely used to predict a wide range of soil properties from spectral data obtained from a mid- and near-infrared diffuse reflectance spectroscopy of soil samples. Whilst this method has been shown to successfully predict a multitude of soil quantities, methods of functional data analysis provide an alternate way of studying continuous data, recognising that it is sometimes more natural, and often fruitful, to view a collection of data points as observed realisations of random functions. In this thesis, the main focus is to compare standard multivariate techniques of analysing soil spectra to methods of functional data analysis. Chapter 1 provides an introduction to the importance of soil monitoring, mid-infrared spectroscopy, a description of the data and the objectives of the thesis. Following this, Chapter 2 demonstrates the performances of principal component analysis, linear discriminant

analysis and support vector machines in investigating the variability of the soil spectra across the mid-infrared range. These multivariate methods are assessed on their ability to distinguish differences between groups of spectra based on various grouping variables. In Chapter 3, functional data analysis is introduced and methods of functional principal component analysis and functional hypothesis testing are implemented. Functional principal component analysis is applied to identify regions of the spectra which contain the principal modes of variation which could be pertinent to explaining differences between samples of different land-uses or sampling sites. Functional hypothesis tests are used to directly test for differences between groups of spectra and pointwise permutation F-tests are used to locate regions of the spectra where these group differences are prominent. Chapter 4 introduces functional linear regression as an alternative to the industry standard of partial least squares regression for relating the spectra to the physical wet chemistry properties of the soil. In this chapter, it is of interest to identify physical soil properties which can be successfully predicted by functional and partial least squares regression; and what the achievable performances of these predictions are. Comparisons between the two approaches are made and the advantages of each approach are considered. Finally, Chapter 5 provides a summary of the work presented and discusses the limitations and remaining challenges for the use of functional data analysis for the characterization of soil.

Acknowledgements

Firstly, I would like to thank my supervisors Prof. Marian Scott and Dr. Claire Miller for their invaluable guidance and support throughout this year. I am extremely grateful for their patience and encouragement, without which this thesis would not have been possible. I would also like to thank CSIRO for providing the data and helpful advice, in particular Peter Toscas and Lynne MacDonald. In addition, I greatly acknowledge the funding from the School of Mathematics and Statistics which allowed me to undertake this thoroughly enjoyable year of work. I thank my fellow postgraduates of Room 420, Cillian and Andy, for sharing their ideas, offering suggestions and providing fine company on many a *Fika* this past year. A final word of thanks to my parents, they have always been my number one fans and I thank them for their limitless support and encouragement.

Declaration

This thesis has been composed by myself and it has not been submitted in any previous application for a degree. The work reported within was executed by myself, unless otherwise stated.

Signed:

Date:

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Background

Soil is a fundamental natural resource which is relied upon globally for its vital ecological and economic functions. Soil is important for the production of food, support of wildlife and the regulation of water movement in the landscape (Stenberg et al., 2010). It provides the foundations of our buildings and a basis for plant life. Furthermore, there is scope for the development of soil management practices which can support carbon sequestration to reduce atmospheric carbon dioxide in the mitigation of global warming (Bricklemyer et al., 2005; Mooney et al., 2004; Stenberg et al., 2010).

Soil health and soil quality are terms that are used interchangeably to describe soils that are not only fertile but also possess adequate physical and biological properties to sustain productivity, maintain environmental quality and promote plant and animal health (Doran and Parkin, 1994). However, soil quality is often considered a complex characteristic and thus there are multiple definitions of what soil quality encompasses. For instance, Larson and Pierce (1991) define soil quality as the capacity of a soil to function whereas (Anderson and Gregorich, 1984) define soil quality as the soil's ability to sustain, accept, store and recycle nutrients, water and energy. However, regardless of how soil quality is defined, regular soil quality monitoring is imperative.

Over the years, scientists have developed and applied various methods for the characterization of soil constituents. For example, traditional methods for quantifying Total Carbon include methods of chromate oxidation (Walkley and Black, 1934) and combustion (Allison et al., 1965). However, these methods are expensive and slow (McDowell et al., 2012). Furthermore, recent estimates suggest that more than 200,000 soil samples are analysed commercially in Australia each year and thus there is much need for the development of more time- and cost-effective alternatives to conventional lab-based analysis (Viscarra-Rossel et al., 2006). The majority of these conventional, laboratory-based analyses involve soil pH, salinity, extractable phosphorus and nitrate nitrogen, exchangeable cations, organic carbon and extractable trace elements (Merry and Janik, 2001). With an increasing demand for good quality and inexpensive data, the method of diffuse reflectance mid-infrared (MIR) spectroscopy has been presented in the literature as a promising new technology for the characterization of soil.

1.2 Mid-Infrared Soil Spectroscopy

The use of spectroscopy for soil analyses is simple, fast, cost-effective, and non-destructive. There is no need for hazardous chemicals, and it requires minimal sample preparation (Batten, 1998; Viscarra-Rossel et al., 2006). Furthermore, the spectrum obtained from one sample contains comprehensive information on a wide range of soil properties and can be used to predict these simultaneously (Cozzolino and Moron, 2006; Nocita et al., 2015). This means that the method is less expensive than conventional laboratory analyses with lower per analysis costs and quick turnaround times (Merry and Janik, 2001; Viscarra-Rossel et al., 2006). This is especially true when it is necessary to analyse a large number of samples. Furthermore, the rapid development of portable spectrometers will permit the method to be implemented in-situ.

The mid-infrared (MIR) range spans the electromagnetic spectrum between 4000 and 400cm^{-1} in wavenumber or 2.5 to 50μm in wavelength. In the context of spectroscopy, wavenumbers are most commonly reported and thus from here on only wavenumbers will be considered. However, the two measures can be related as follows:

$$\tilde{v} = \frac{1}{\lambda}$$

where $\tilde{v}$ represents the wavenumber as the number of wavelengths per unit distance, and λ is the wavelength.

MIR spectroscopy obtains spectra from the mid-infrared range where direct information about the elements of a sample is provided (Etzion et al., 2004). In the generation of a MIR soil spectrum, radiation of the relevant frequencies (4000-400cm^{-1}) is directed at a soil sample. The spectral signature obtained depends on the constituents present in the soil which influence the frequencies at which light is absorbed. The directed radiation causes individual molecular bonds of the soil chemistry to vibrate, either by bending or stretching, and they absorb light to various degrees. The frequencies at which light is absorbed are reported as % reflectance (R). However, this is commonly transformed to a value of apparent absorbance: $A = \log(1/R)$ (Stenberg et al., 2010). The objective of mid-infrared spectroscopy is to associate specific absorbances with functional groups common in the soil mineral and organic matter. However, with soil having a complex nature due to its composition, the combined contributions from various soil components can result in very complicated spectral signatures which can be difficult to interpret. The spectral signatures from soil samples are largely nonspecific due to the overlapping absorption regions of different soil constituents. Additionally, some soil constituents such as quartz can cause a scattering effect which further complicates the interpretation of spectra (Stenberg et al., 2010).

1.3 Spectroscopic Calibrations

Due to the lack of specificity and the presence of highly correlated neighbouring wavenumbers, multivariate calibration techniques are commonly used to correlate the spectra with various physical and chemical soil properties of interest (Martens and Naes, 1990; Wetterlind et al., 2013). Once a calibration model is established it can be used for the prediction of soil quantities not used in the original calibration of the model (Cobo et al., 2010).

First introduced for spectral data by Haaland and Thomas (1988), partial least squares regression (PLSR) is the most widely used calibration method in soil science. It is based on the assumption of a multilinear relationship between predictor variables (e.g. the absorbance peaks in the spectra) and the response variable of interest (e.g. Total Carbon in soil) (Niazi et al., 2015). The PLS regression analysis reduces the number of predictors by constructing linear combinations (components) of the original predictor variables. Combined with spectroscopic techniques, PLSR has been frequently reported for the successful prediction of soil pH, Soil Organic Matter (SOM), Total Organic Carbon (TOC), Phosphorus (P), Potassium (K), Iron (Fe), Calcium (Ca), Sodium (Na), Magnesium (Mg), Total Carbon, Total Nitrogen, texture as well as biological properties (Bellino et al., 2016; Feyziyev et al., 2016; Stenberg et al., 2010; Zornoza et al., 2008; Viscarra-Rossel et al., 2006; Yang et al., 2012; Heinze et al., 2013; Conforti et al., 2015). Forrester et al. (2015) successfully predicted phosphorus via the phosphorus buffer index (PBI) through the implementation of PLSR to mid-infrared spectra of 601 Australian agricultural soils. Reeves et al. (2001) also showed that results using partial least-squares regression gave accurate calibrations for the determination of a number of compositional parameters of soil including Total Carbon, Total Nitrogen, pH, and different measures of biological activity. PLSR has also been applied to the near-infrared range of spectra and this method has been shown to have a good capacity to predict soil organic matter, even for samples of different types (Fidencio et al., 2002; Nocita et al., 2013).

Mid-infrared spectroscopy has been introduced as a cheaper alternative to the prohibitively expensive laboratory based techniques for soil analysis. Partial least squares regression has previously been successful in the prediction of many soil properties from mid-infrared soil spectra. However, despite the success of partial least squares regression and the popularity within the soil science community, there is an argument that methods of functional data analysis are more suited to this problem. Advances in technology enables large volumes of data to be recorded and often these observations are recorded over high frequencies of time, space and various other continua. This allows the idea that these collections of data points are observed realisations of underlying smooth functions. Philosophically, functional data analysis presents a more natural and more advantageous method of studying continuous data. The main focus of this work is to compare these approaches and assess the information which is provided by an FDA approach.

1.4 What can be measured?

A comprehensive knowledge of soil optical characterisation is essential to finely interpret the output from calibration techniques such as PLSR. In this work, the guide provided by Soriano-Disla et al. (2014) which outlines approximate wavenumbers of spectral absorptions in the MIR region for some of the major soil components will be used. These are summarised in Table 1.1.

Major Soil Component	Wavenumber (cm^{-1})
Quartz (sand)	1100-1000
Clay Minerals:	3690-3620 (Kaolinite)
	3620-3630 (Smectite)
	3400-3300 (Illite)
Carbonates	1430 and 2520
Iron Oxides	600-700
Iron oxyhydroxides	3100, 900, 800
Organic Matter:	2930-2850 (Alkyl ($-CH_2$))
	1670 and 1530 (Protein Amide ($OC - NH$))
	1720 (Carboxylic Acid ($COOH$))
	1630 (Water associated)
	1600 and 1400 (Carboxylate anion ($-COO^-$)
	1600-1570 (Aromatic Groups)

Table 1.1: Approximate wavenumbers of spectral absorptions in the MIR region for major soil components as identified by Soriano-Disla et al. (2014)

This table is only used as a general guide for the interpretation of output from the statistical analyses of this thesis. The major soil components identified in the mid-infrared range comprise of quartz, clay minerals (Kaolinite, Smectite/Illite), carbonates, iron oxides and soil organic matter. Following discussions with soil scientists at CSIRO it has also been learned that at either end of the MIR spectral range, mineral signals may be found at approximately 3800-3500cm^{-1} and 1300-400cm^{-1}. Nitrates can also be found to be absorbed at approximately 1370cm^{-1} and lignin at 900-860cm^{-1}. Given the complexity of soils and the non-specificity of absorption regions this information should not be treated as exhaustive. The MIR spectra are difficult to finely characterize with many constituents having frequencies which overlap with each other. For example, it has been found that Fe oxide minerals have frequencies that overlap with other soil mineral peaks near 600-700cm^{-1}, making them difficult to identify. The COO- anion and water frequencies have also been found to overlap with aromatic groups near 1600-1570cm^{-1} (Soriano-Disla et al., 2014). Additionally, the scattering effect caused by minerals such as quartz can also complicate interpretation.

1.5 Description of the Data

Geographically, this study focuses on sampled data from three paired farmland sites in the Bookham area within the region of New South Wales, Australia (Figure 1.1). Used extensively for sheep grazing, the sample sites were on farms that had been established for more than forty years and were located within 15km of one another: Bogo (34.813746°S, 148.704558°E), Glenrock (34.858413°S, 148.56724°E) and Talmo (34.936976°S, 148.625293°E).

At each farm, one pasture plot was selected that lay adjacent to an area of native woodland, which was not grazed or actively managed apart from fence maintenance to keep livestock out. With the sites being within 15km of each other, the local climactic conditions are quite similar. The approximate

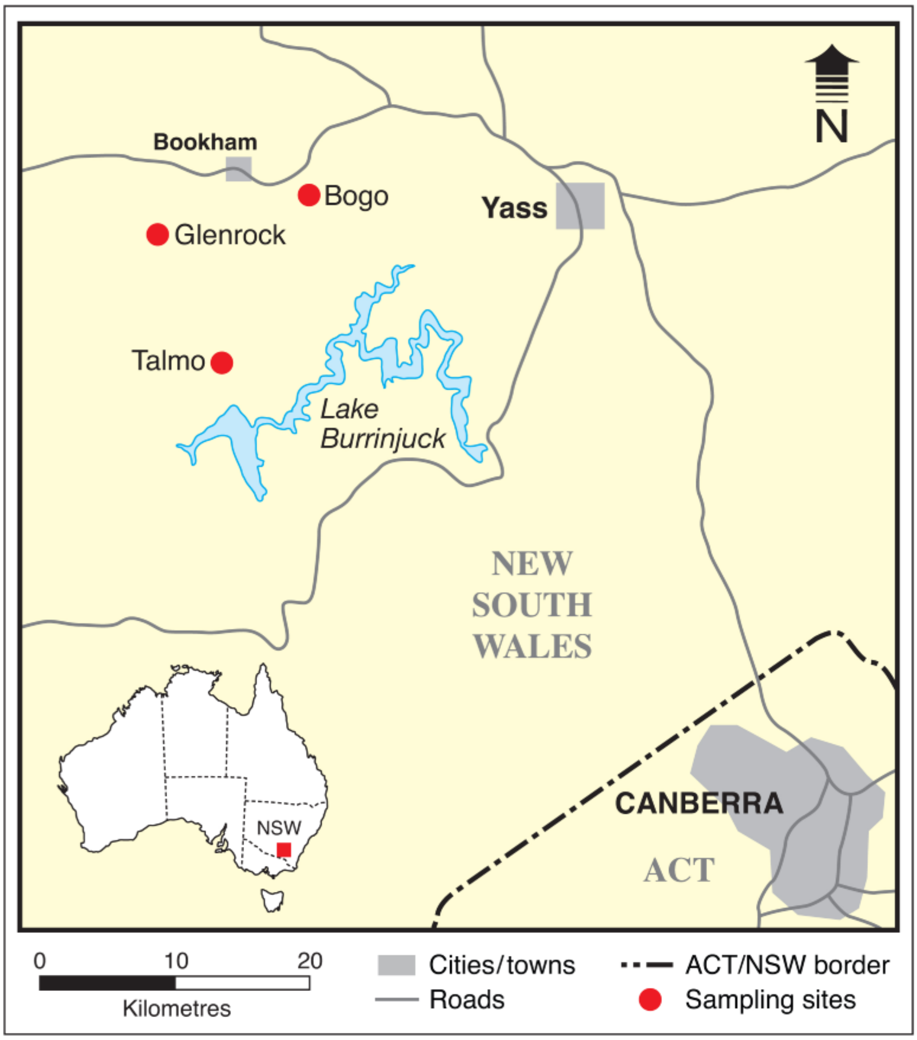


Figure 1.1: Geographic map of the sampling region (de Menezes et al. (2015))

distances between the woodland and pasture plots at each site were 160m (Talmo), 200m (Glenrock) and 440m (Bogo). Adjacent sites were chosen for analysis to minimize the differences in parent soil type between land uses.

A total of 240 soil samples were collected for analysis, and a third of these samples ($n = 60$) were obtained from each farmland. Within each farmland, the sixty samples were obtained equally between their woodland and pasture sampling plots. These data were supplied by CSIRO, and a more comprehensive description of the study sites and soil sampling procedures may be found in De Menezes et al. (2015). In short, at each woodland or pasture site, a 100 × 100m sampling plot was measured out with 25m intervals and three subplots also marked out at 12.5m intervals. Two soil cores of approximately 10cm depth and 2cm diameter were taken at the intersections of the grids and composited. The soil cores were kept cold (4°C) until they were processed and the samples were later sieved, homogenized and separated into aliquots. These aliquots were then prepared differently depending on the aspect of wet chemistry being investigated.

The MIR soil reflectance spectra obtained from the samples had a short range with a mean coarse resolution of about 4cm^{-1} between each wavenumber. This allowed for 921 diffuse reflectance measurements per spectra to be defined in the MIR range from approximately 4000-450cm^{-1} . In the presentation of results, these diffuse reflectance measurements may also be referred to in this thesis and in the wider literature as spectral bands, bandwidths or as the wavenumber. An example of the type of data analysed in this thesis is presented in Figure 1.2 where the soil spectra obtained from the soil samples of the Talmo farmland can be observed. In the laboratory, methods of wet chemistry were also carried out on the same soil samples and the data recorded was paired with the soil spectra data. The soil properties investigated are summarised in Table 1.2.

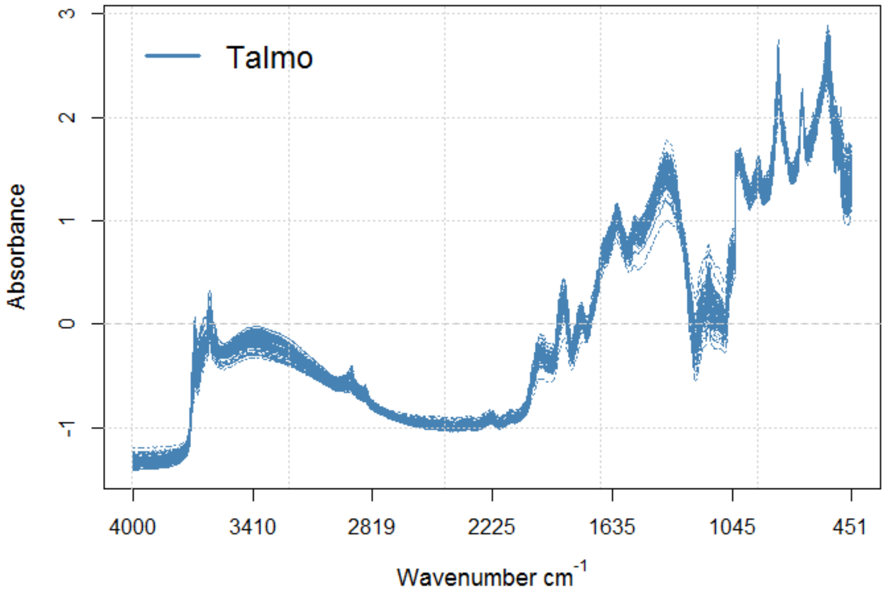


Figure 1.2: Exploratory plot of the spectral signatures resulting from an MIR diffuse reflectance spectrscopy of the soil samples obtained from the Talmo woodland

Variable No.	Wet Chemistry Variables	Units
1	pH	pH
2	Moisture	%
3	Carbon (C)	mg/g
4	Nitrogen (N)	mg/g
5	Total Dissolved Nitrogen (TDN)	mg/g
6	Dissolved Organic Nitrogen (DON)	mg/g
7	Amino-N	mg/g
8	NH4-N	mg/g
9	NO3.N	mg/g
10	Biomass N	mg/g
11	Total Dissolved Carbon (TDC)	mg/g
12	Microbial Carbon	mg/g
13	Inorganic Phosphorus	mg/g
14	Organic Phosphorus	mg/g

Table 1.2: Soil Wet Chemistry Variables investigated via standard laboratory procedures

1.6 Aims and Objectives

Research has previously been published in evaluating the potential of mid-infrared (MIR) spectroscopy for the determination of physical soil quantities using the data described in the previous sections. However, this research has predominantly investigated multivariate methods and in particular, partial least squares regression as a means of calibrating the spectra with specific soil physical and chemical attributes. Partial least squares regression has been shown to facilitate mid-infrared spectroscopy as an effective method of soil characterization which is far cheaper than the laboratory based alternatives. In this study, the general overarching goal is to introduce functional data analysis (FDA), presenting it as an alternative approach to the classical multivariate statistical methods currently employed for the analysis of soils. Functional data analysis is investigated as it appears to be a theoretically more sound approach to the MIR spectroscopy problem with the current multivariate approaches violating assumptions about the data. Furthermore, with growth in the field functional data analysis has the potential to unlock more information beneath the data and has additional exploratory benefits.

Using the data which has been provided by CSIRO; soil spectra and wet chemistry data will be considered for the investigation of the following general objectives:

1. Using FDA, explore the variability across the soil spectra and between different groups of spectra

2. Investigate functional principal components analysis as a means of identifying the areas of highest variability across the spectra

3. Investigate functional regression as an alternative to PLSR for the prediction of soil properties

In Chapter 2, only the soil spectra data are considered and standard statistical methods of analysis are used to investigate the variability across the mid-infrared range. Multivariate methods of principal components analysis, linear discriminant analysis and support vector machines are assessed on their ability to distinguish differences between groups of spectra based on land-use, sampling site location and an interaction between these grouping variables.

In Chapter 3, functional data theory is introduced and methods of functional data analysis are applied to the soil spectra data in the same vein as Chapter 2. After a functional exploration of the data, a functional principal components analysis is carried out and contrasted with the multivariate equivalent. Additionally, functional hypothesis testing is used to test for differences between groups of spectra.

Chapter 4 introduces the various forms of functional regression to relate the soil spectra with the physical wet chemistry data. Functional regression models are formed to predict the MIR spectra from wet chemistry variables, and the more interesting scalar-on-function regression is pursued for the individual predictions of specific soil constituents. In this chapter, it is of interest to find which physical soil properties can be predicted by a functional regression approach and what the achievable performance of these predictions are. Additionally, comparisons with a partial least squares approach are made and the advantages of each approach are considered.

Finally, Chapter 5 will conclude and discuss the findings of the study, present the advantages and disadvantages of a functional data approach, and suggest the steps which could be taken to overcome the remaining challenges of functional data analysis.

Chapter 2

Exploring the Spectra via Multivariate Methods

In this chapter, standard approaches such as Principal Component Analysis, Linear Discriminant Analysis and Support Vector Machines are used to explore the soil spectra. Predominantly, the focus is on identifying regions of the spectra which may be responsible for differences between samples from different groups (i.e. site and land-use combinations). These regions of difference may then be related to particular soil constituents which are absorbed in these spectral ranges.

Firstly, some necessary theoretical foundations of multivariate analyses are introduced. It is crucial for these concepts to be familiarised since they provide the basis for functional equivalents introduced in Chapter 3. Although the statistical theory is widely available in the literature, it has been adapted and presented in such a way to suit this thesis.

2.1 Multivariate Theory

2.1.1 Principal Component Analysis

First introduced by Pearson (1901), Principal Components Analysis (PCA) is now an established multivariate statistical method which is well described in the literature. It is commonly used as a dimension reduction technique and a way of summarising the principal modes of variation within the data. Whilst reducing the complexity of a dataset, the goal is to minimize information loss by finding the best low-dimensional representation of the variation. This is done by transforming the original correlated variables into a smaller number of uncorrelated variables called principal components (PCs) (Jolliffe, 2002). These principal components are linear combinations of the original variables given in decreasing order of importance, where the importance is determined by the percentage of variation of the original data that is explained by each component. Thus, the first principal component (PC1) explains the most of the total variability. In the case that the original variables are highly correlated, then the first few PCs will account for most of the variation and the remaining PCs can be discarded with minimal information loss. It is then hoped that the first few components, containing most of the information, will be meaningful.

Principal Components

The first principal component, z_1, is simply a linear combination of the original variables $x_1, x_2,\ldots, x_p$, and can be defined as follows:

$$z_1 = \alpha_1^T\mathbf{x} = \alpha_{11}\mathrm{x}_1 + \alpha_{12}\mathrm{x}_2 + \ldots + \alpha_{1\mathrm{p}}\mathrm{x}_\mathrm{p} = \sum_{\mathrm{j}=1}^{\mathrm{p}} \alpha_{1\mathrm{j}}\mathrm{x}_\mathrm{j} \tag{2.1}$$

The coefficients in Equation 2.1, $\alpha_{11}, \alpha_{12}, \ldots, \alpha_{1p}$, are known as the loadings. These indicate the relative importance of the variable in the component. These loadings, also referred to as weights, are mathematically determined to maximise the variation of the original data in $\mathbf{x}$, subject to the normalization constraint

$$\alpha_{11}^2 + \alpha_{12}^2 + \ldots + \alpha_{1p}^2 = 1. \tag{2.2}$$

Next, the second principal component, $z_2 = \alpha_2^T \mathbf{x}$, is determined and has the maximum variance that is uncorrelated with z_1. This process is continued to identify the rest of the principal components, all subject to the same normalization constraint (Equation 2.2). The number of principal components formulated is equal to the number of original variables, p . In highly correlated datasets, most of the variation will be accounted for in just a few principal components. However, there may be valuable information contained within trailing PCs and thus caution must be taken when choosing the number of PCs to extract.

Eigenvectors, Eigenvalues and Eigen-analysis

For an $n \times n$ matrix, $\mathbf{C}$, and nonzero vector $\mathbf{q}$, the values of λ satisfying the equation,

$$C\mathbf{q} = \lambda \mathbf{q} \tag{2.3}$$

are defined as the eigenvalues of $\boldsymbol{C}$. The vectors $\mathbf{q}$ satisfying Equation 2.3 are the corresponding eigenvectors. At an absolute maximum, the number of non-zero eigenvalues of $\boldsymbol{C}$ is equal to the total number of linearly independent columns of $\boldsymbol{C}$ - the definition of a full rank matrix. This can be rewritten as,

$$(\mathbf{C} - I_n)\mathbf{q} = 0$$

Provided $\boldsymbol{C}$, and $(\mathbf{C} - I_n)$, are $n \times n$ square and full rank matrices the eigenvalues of $\boldsymbol{C}$ can be found by solving,

$$\det(\mathbf{C} - I_m) = \mathit{0}$$

where $\det(\boldsymbol{S})$ denotes the determinant of a square matrix, $\boldsymbol{S}$.
With n linearly independent eigenvectors $[q_1, \ldots, q_n]$, $\boldsymbol{C}$ can be expressed as the product of the matrices,

$$\mathbf{C} = \mathbf{Q}\boldsymbol{\Lambda}\mathbf{Q}^{\mathbf{-1}} \tag{2.4}$$

where $\boldsymbol{\Lambda}$ is a diagonal matrix with the eigenvalues of $\boldsymbol{C}$ decreasing along the main diagonal $(\lambda_1, \ldots, \lambda_n)$. The matrix $\boldsymbol{Q}$ is the matrix of eigenvectors, $\mathbf{Q} = [q_1, \ldots, q_n]$, with the i^{th} eigenvector corresponding to the i^{th} largest eigenvalue. Equation 2.4 above is known as the eigen-decomposition of $\boldsymbol{C}$ and is used to perform PCA. The eigen-decomposition can be carried out in a variety of methods including singular value decomposition, the Jacobi method, QR or Choleskey decomposition.

The eigenvalues represent how much of the variation in the data is described by the corresponding principal component, making it possible to rank the principal components in order of importance. The associated eigenvectors provide the coefficients (weights) referred to as the loadings.

Visualisations in PCA

In this thesis, scores and loadings plots are produced in order to determine which variables appear to explain the differences between grouping variables. Scores plots project an observation onto the first two or three components, and can be used to look for class separations in the data. If a particular principal component is found to consistently exhibit class separations in its score plots with other principal components then its loadings plot should be examined. The loadings plot is a simple visualisation method which shows the weights of the original variables in each principal component. The original variables associated with loadings of largest magnitudes are identified as having a strong influence on their corresponding principal component.

Selection of Principal Components

A key step in PCA is the determination of the reduced number of dimensions which adequately describe the variation in the data. This is most easily determined by consulting the scree plot as developed by Cattell (1966). Figure 2.1 gives an example of a screeplot where the eigenvalues are plotted against the number of principal components. Cattell suggests identifying the point where the smooth decrease in eigenvalues levels off to the right as the appropriate number of principal components. All components prior to this point should be retained, with the variance represented in the tail of the curve assumed to represent only the random variability in the data, i.e. the noise.

Proposed by Kaiser (1960), the Kaiser criterion is another widely used selection criteria stating that principal components should only be retained with eigenvalues greater than one. This is essentially saying that only com-

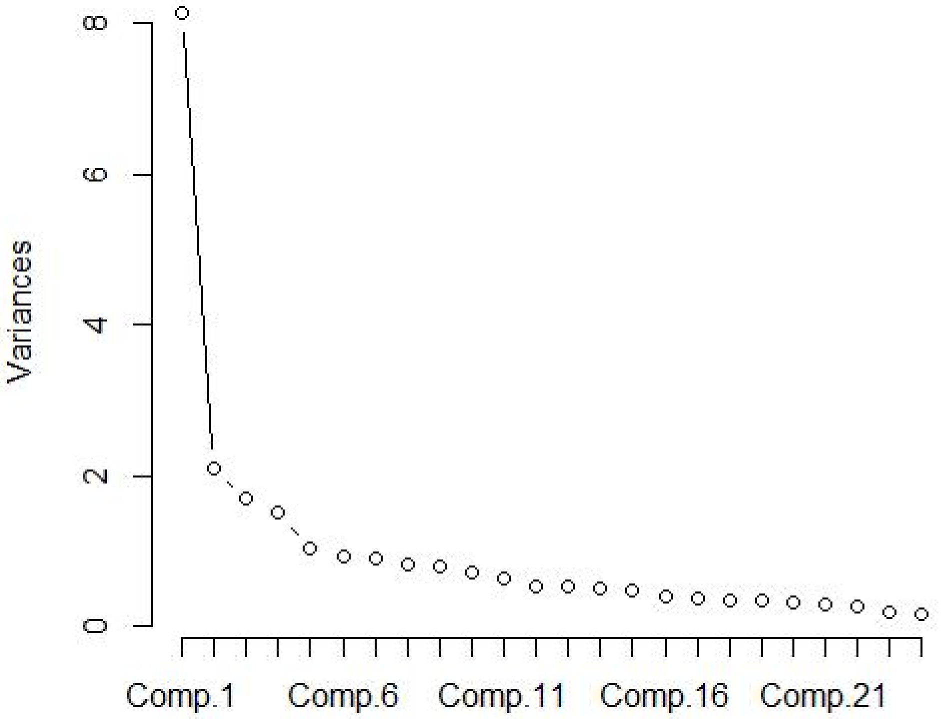


Figure 2.1: An example of a scree plot as developed by Cattell (1966). The number of principal components should be selected at the point where the decrease in eigenvalues levels off.

.

ponents which extract at least as much variance as the equivalent of one original variable should be extracted. Another popular proposal is to consider the proportion of variance explained. It is standard to decide *a priori* that a certain amount of variance is to be explained (usually 80-90%). Only the leading components which contribute to this arbitrary threshold are kept.

These methods work when there are relatively few clearly defined principal components. However, in practice, the context of a problem should also be considered before deciding on the optimal number of components to retain. For example, it may be of interest to retain additional principal components further than the recommended 90% variability threshold. These additional PCs could contain information crucial to explaining the subtleties between classes of data that are not given by larger preceding PCs. This is common in high dimensional data such as hyperspectral data. Thus, in order to minimize the risk of information loss in the forthcoming analyses it was decided that principal components should be retained by the following criteria:

1. Leading PCs were retained accounting for a cumulative 90% of the total variance explained.

2. Thereafter, trailing PCs were retained providing they satisfied Kaisers criterion and explained at least 0.01% of the variance.

Rotation Methods

The interpretability of principal components can sometimes be problematic. Rotation is a method by which the interpretation of components can be made easier. It is a procedure usually carried out after the extraction of principal components to maximize high correlations between components and variables, and minimize low correlations (Tabachnick and Fidell., 2007). In this way, the solution is made more interpretable without changing its mathematical properties. Thurstone (1947) and Cattell (1978) recognised the procedure for its ability to achieve simple structure. Essentially, simple structure is achieved when each variable has a substantial loading on one and only one factor with the rest of the loadings being zero or close to zero. The initial loadings matrix is transformed in order to achieve simple structure, and the adopted transformations are orthogonal or oblique rotations which gives rise to the methods' name.

Thurstone (1947) suggests a matrix of loadings is simple if it satisfies the following five criteria:

1. Each row contains at least one zero;
2. For each column, there are at least as many zeros as there are columns;
3. For any pair of factors, there are some variables with zero loadings on one factor and large loadings on the other factor;
4. For any pair of factors, there is a sizeable proportion of zero loadings;
5. For any pair of factors, there are only a small number of large loadings.

There are many methods which could be used to rotate the initial factor loadings, both oblique and orthogonal. However, Gorsuch (1983) suggests that if simple structure is made clear, any one of the more popular procedures can be expected to lead to the same interpretations. Developed by Kaiser (1958), Varimax is unquestionably the most popular rotation method. Varimax aims to maximise the sum of variances of squared loadings in the columns of the factor matrix producing loadings in each column which are either very high or near zero.

2.1.2 Linear Discriminant Analysis

Basic LDA

In the classification of data into a known number of groups, one possible way is to fit a linear model to determine a decision boundary which would discriminate a set of K independent classes. Originally developed by Fisher (1936), Linear Discriminant Analysis (LDA) is a well-established statistical technique used for the classification of multivariate data. It is a method of finding linear combinations of variables which maximize the separation between two or more classes. It is similar to principal components analysis in that it is a data driven technique for dimensionality reduction. However, unlike PCA, linear discriminant analysis is a supervised learning method in which the class labels are known *a priori* in an attempt to preserve as much of the class discriminatory information as possible. In calculating the linear discriminants, important variables are identified and the functions can be used to allow new observations to be classified.

The linear combinations of the original variables which achieve the best possible separation between classes are known as the discriminant functions. These discriminant functions are found such that the ratio of between-class variance to within-class variance is maximized to allow for adequate class separability. Additionally, the linear combinations are sorted by the relative importance they have in distinguishing between groups. Geometrically, LDA has the goal of finding the axis which provides the maximum separation between the distributions of the discriminant scores of different groups as illustrated in Figure 2.2. In this two-class problem, a great deal of overlapping in the classes can be observed. However, in the direction of the first linear discriminant (LD1) the classes can be seen to be well separated. Since this is a two-class problem, no other linear discriminant function can be obtained to achieve a better discrimination.

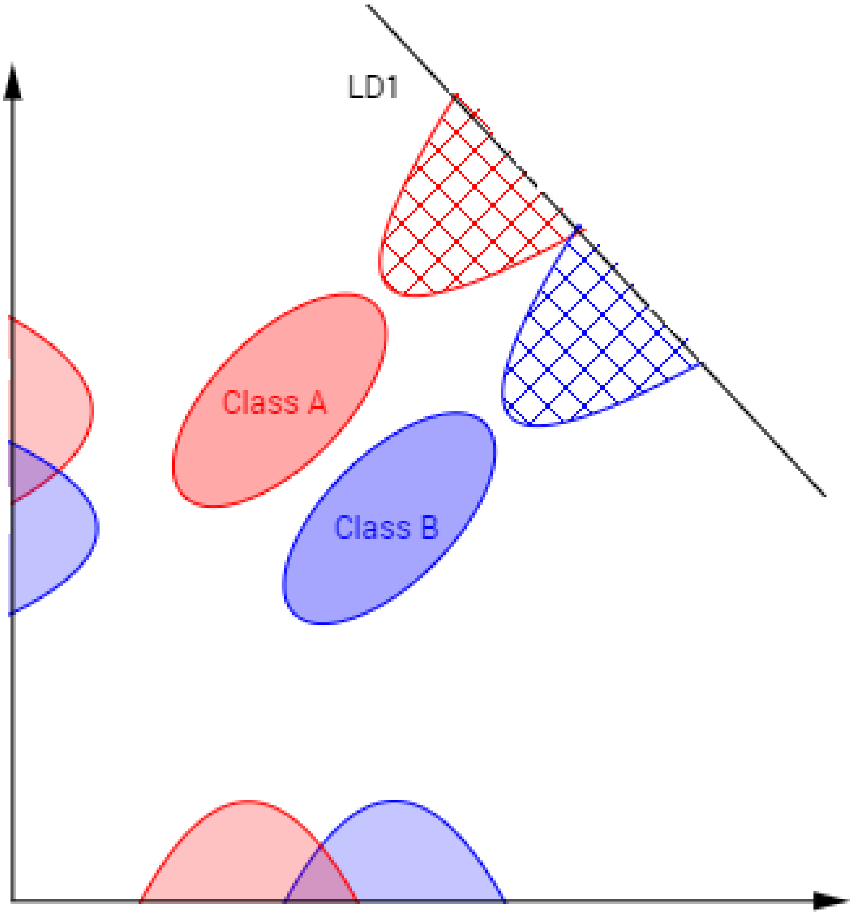


Figure 2.2: Geometric representation of linear discriminant analysis. In the direction of the first linear discriminant the classes become linearly separable.

The application of LDA has some important assumptions regarding the groups including multivariate normality and the sharing of a common covariance matrix σ (Balakrishnama, 1998). Suppose there are K different groups, each obeying these assumptions with mean vectors $\mu_k (k = 1, \ldots, K)$. The idea of LDA is to classify observations x_i to the group k, which minimize the within-group variance, i.e.,

$$k = \text{argmin}_k (x_i - \mu_k)^T \sigma^{-1} (x_i - \mu_k)$$

Under multivariate normal assumptions, this is equivalent to finding the group that maximizes the likelihood of the observation. In this problem, N objects are observed and for each object the values and class membership of variables $X_1, \ldots, X_p$ are known. Providing the assumptions of multivariate normality and homogeneity of the covariance matrices hold, the linear discriminant function provides an optimal classification rule to minimise the probability of misclassification. The original linear discriminant was described for a two-class problem, and later generalised by Rao (1948) for multi-class linear discriminant analysis.

Stepwise Linear Discriminant Analysis

With high dimensional data, linear discriminant analysis may involve constructing linear combinations of many predictor variables. Thus, a more practical method of performing LDA could be to select significant variables via a stepwise procedure. In stepwise discriminant analysis, the model of discrimination or decision rule is built step by step. At each of these steps, variables are evaluated on their contribution to the discrimination between classes. The variable identified as having the greatest contribution is selected for inclusion before repeating the process for the remaining variables. This particular form is known as forward stepwise selection. In a backward stepwise selection the process works similarly but removes the variables with the least contribution to discrimination one by one. In both cases the process identifies the most influential variables for discrimination between classes.

2.1.3 Support Vector Machines

Introduced by Cortes and Vapnik (1995), support vector machines are a relatively new supervised learning technique for solving difficult classification problems. The technique is used across a wide range of application domains and is well known for its strong theoretical foundations, generalization performance and ability to handle high dimensional data (Batuwita and Palade, 2013; Devos et al., 2009). Despite its advanced underlying theory, SVMs are mostly treated as a black box technique and thus this section only presents an overview of support vector learning including the popular extension to non-linear problems. A more in depth description of the mathematical theory can be found in Vapnik's *The Nature of Statistical Learning Theory* and in section 4.5 *Separating Hyperplanes* of the book: *Elements of Statistical Learning* by Hastie et al. (2009).

The Linearly Separable Case

In the binary classification problem, the objective of an SVM classifier is to derive a function, $f : \mathbb{R}^N \mapsto \{\pm\}$, that describes the decision boundary or hyperplane separating the classes by maximising a margin between them (Figure 2.3). A two-class problem is linearly separable if the hyperplane can be positioned such that all data of one class fall on one side and all data of another class fall on the opposite side. There can be many linearly separating hyperplanes as illustrated in Figure 2.3 (right); however the goal is to find the optimal hyperplane that maximizes the margin where the margin is the distance between the nearest datum and the hyperplane.

The Soft Margin Approach

For linearly separable data, the SVM produces an optimal hyperplane with the largest possible margin with the decision boundary separating the two classes without error. This is referred to as a hard margin SVM. However in practice, data points are not frequently linearly separable and even in the lin-

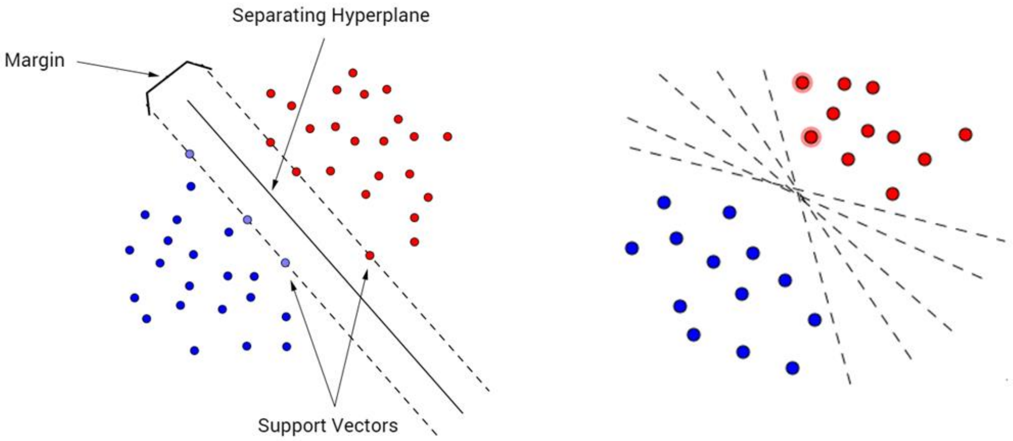


Figure 2.3: The margin and separating hyperplanes

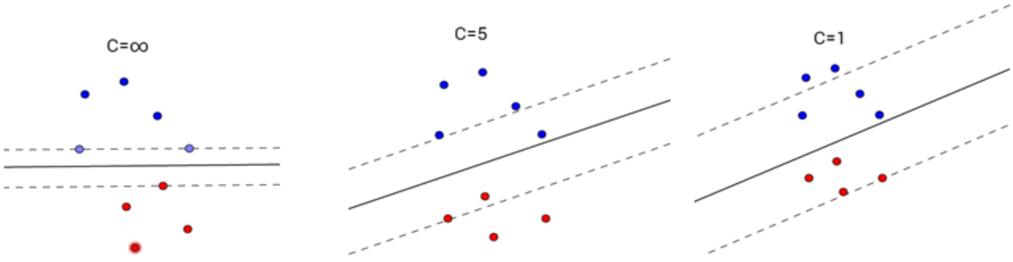


Figure 2.4: The effect of varying C on the hyperplanes obtained

early separable case a greater margin with better generalization performance can be achieved by allowing the classifier to misclassify points. Proposed by Cortes and Vapnik (1995), this type of SVM is referred to as the soft margin SVM.

With an infinite number of separating hyperplanes available there is a risk of selecting a classifier based on a hyperplane which separates the training data perfectly but will perform poorly on unseen data. Within the SVM framework, this is referred to as overfitting (Cortes and Vapnik, 1995). By artificially separating the data through projecting the training data to higher dimensional feature spaces, support vector machines are at risk of finding these trivial solutions that overfit the data. Misclassifying some training points allows a separating hyperplane to be obtained with an overall better position. The misclassification rate is regulated by a penalty weight C, known as the cost function. Figure 2.4 demonstrates the effect of varying the value of C on the choice of hyperplane.

A large C makes the cost of misclassification high causing the SVM to behave as a hard margin-SVM. The data are explained very strictly and the decision boundary results in a perfect classification rate. This risks overfitting and the small margin limits the generalization of the SVM. Reducing the penalty C increases the margin and some points can now fall within it. When the penalty value is reduced even further, an even larger margin is obtained. It is intuitively clear that by lowering the cost values, the underlying distribution of the data is captured much better and greater generalization behaviour can be achieved.

Extension to Non-Linear Class Boundaries

One of the reasons support vector machines have risen in popularity is the use of the so-called *kernel trick* (Kivinen et al., 2004). It is a convenient method to solve the more realistic non-linear classification problems in arbitrarily high dimensional feature spaces (Shivaswamy, 2007). Figure 2.5 illustrates how kernel functions map data into a higher dimensional feature space in which these non-linear problems become linearly separable.

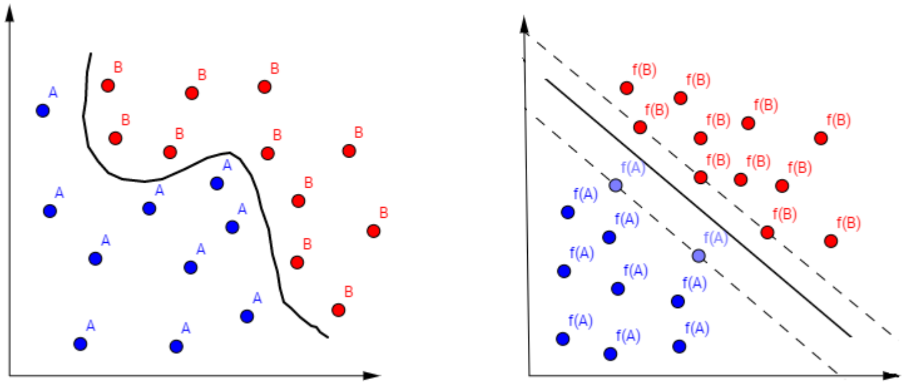


Figure 2.5: The mapping of data into higher dimensional feature spaces via kernel functions to linearly separate data.

Though new kernels are being proposed all the time, three of the most commonly implemented functions found in the literature are:

- Linear
- Polynomial
- Gaussian Radial Basis Function (RBF)

Depending on the type of kernel, specific parameters must be set. With fewer numerical difficulties, Hsu and Lin (2010) recommend using the RBF kernel, and an advantage of this kernel is that there are only two hyper-parameters to consider, the cost function C and gamma (γ)- where γ defines how far the influence of a single training point reaches.

2.2 Multivariate Applications

Now that some multivariate theory has been introduced, the focus switches to applying these multivariate methods to the soil spectra data. As a starting point some exploratory analyses are investigated to gain a general impression of the data structure. The raw spectra are presented in Figure 2.6, and the general spectral signature is shown to be quite variable throughout the entire mid-infrared range exhibiting many peaks and troughs in the data. However, the variability in the curves is different depending on the MIR region observed. For example, in the 3500-2000cm^{-1} range the variability is quite low in comparison with the variability observed at approximately 1000-450cm^{-1} where there is a high degree of variability and a large number of local features.

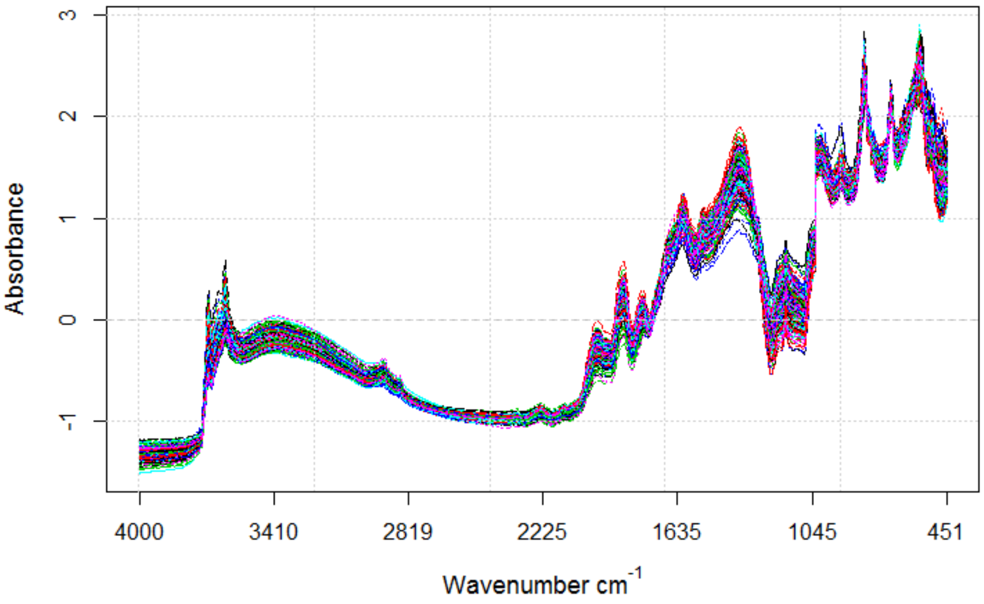


Figure 2.6: Exploratory plot of the raw spectral data investigating the variability in the curves

Furthermore, in examining the shape of the spectral signatures, there is no regular periodic nature to the data. However, the curves of each replicate tend to follow each other very tightly. Boxplots of the data were also produced and Figure 2.7 (top) gives the absorption ranges across the MIR range for each subgroup of the data (i.e. site and land-use combinations). Descriptive statistics associated with each of these boxplots are also presented in Table 2.1.

	All Data	Pasture	Woodland	Bogo	Talmo	Glenrock
Minimum	-1.51	-1.43	-1.51	-1.47	-1.41	-1.51
Median	-0.26	-0.26	-0.27	-0.30	-0.25	-0.24
Maximum	2.90	2.90	2.81	2.85	2.90	2.85

Table 2.1: Descriptive Statistics for the Absorption Ranges of the Spectra for each class

The descriptive statistics and boxplots of each subgroup are very similar demonstrating that if differences between each class exist then they are likely to be very subtle. Boxplots corresponding to each individual wavenumber are also presented in Figure 2.7 (bottom) across the approximate wavenumber range 3770-3650cm^{-1}. The 3770-3650cm^{-1} range has been presented since the dataset is too large to present all boxplots for the full mid-infrared range. Furthermore, plotting over this wavenumber range is useful for demonstrating that there is existing correlation between neighbouring wavenumbers with boxplots in close proximity which exhibit very similar ranges and shapes. This plot is also useful for highlighting that the range of absorption values is different depending on the wavenumber range considered, which is an aspect overlooked in the original boxplots of all data. Furthermore, the variability exhibited in the boxplots is different across the spectral range. For example, boxplots in the 3770-3720cm^{-1} range exhibit very tight ranges in comparison

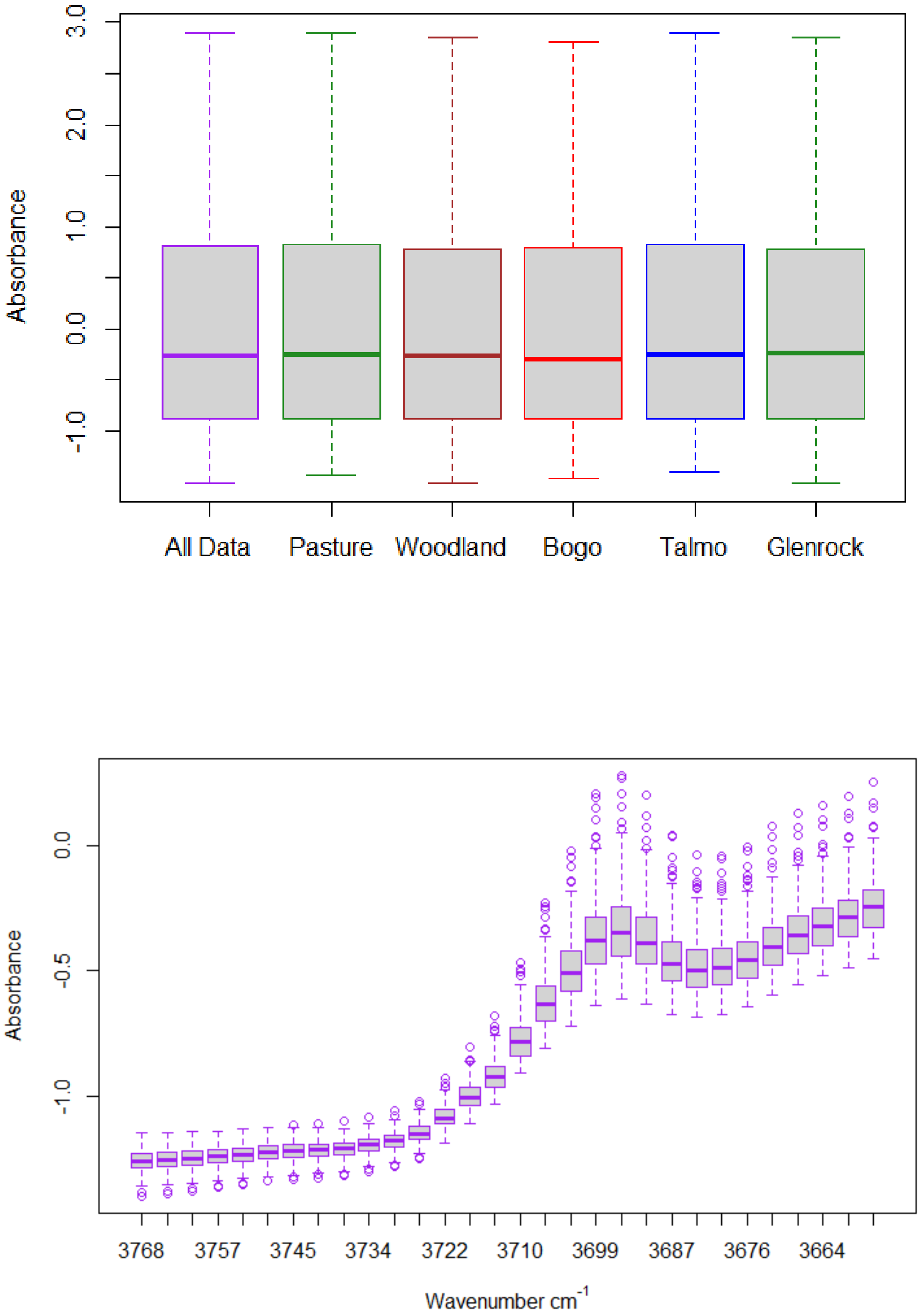


Figure 2.7: Boxplots of the absorption ranges exhibited by the spectra. The top plot gives the overall range of absorptions investigated by subgroup; the lower plot investigates the range of absorptions observed per wavenumber over the approximate 3770-3650cm^{-1} range.

with those at approximately 3700-3650cm^{-1}. Note, these differences are exhibited across the entire mid-infrared region and not just within the spectral range presented.

Given the high dimensionality of the spectral data and the close proximity between each diffuse reflectance measurement (wavenumber measurement), the correlation exhibited in Figure 2.7 (bottom) was not unexpected. This correlation can be likened to temporal correlation found within time series data. Since it is not likely that each of these wavenumbers contains completely independent information it is necessary to apply statistical techniques to both reduce the dimensionality of the dataset and identify the influential information within. In the next sections methods of Principal Components Analysis, Linear Discriminant Analysis and Support Vector Machines are applied to address this issue and also to investigate class discrimination.

2.2.1 Principal Components Analyses

Principal component analysis (PCA) is used to transform the MIR data set into linear combinations of the original spectra variables and attempt to reduce the dimensionality. In the process, combinations responsible for most of the variation in the data are identified. Basic PC analyses are implemented on the full spectra as well as on subsets of the data split by land-use and sampling site. The objective is to examine the data for possible class separations and locate regions of the spectra responsible for any observed differences. It is hoped that the identified MIR band regions can point towards specific aspects of the soil mineralogy which may be responsible for differences between land-uses, sites and the interaction between.

Suitability of Principal Components Analysis

With such high dimensional data, graphical display of data is not a luxury and gaining a general impression of the relationships between variables is not easily achieved. Since the MIR dataset is too vast to fully explore, Figure 2.8 gives pairwise scatterplots for the first nine of the 921 wavenumber MIR measurements. The spectral bands appear almost perfectly positively correlated reporting Pearson correlation coefficients very close to one. Additionally, the distributions of the diffuse reflectance measurements appear approximately normal confirming the Gaussian assumptions made in PCA.

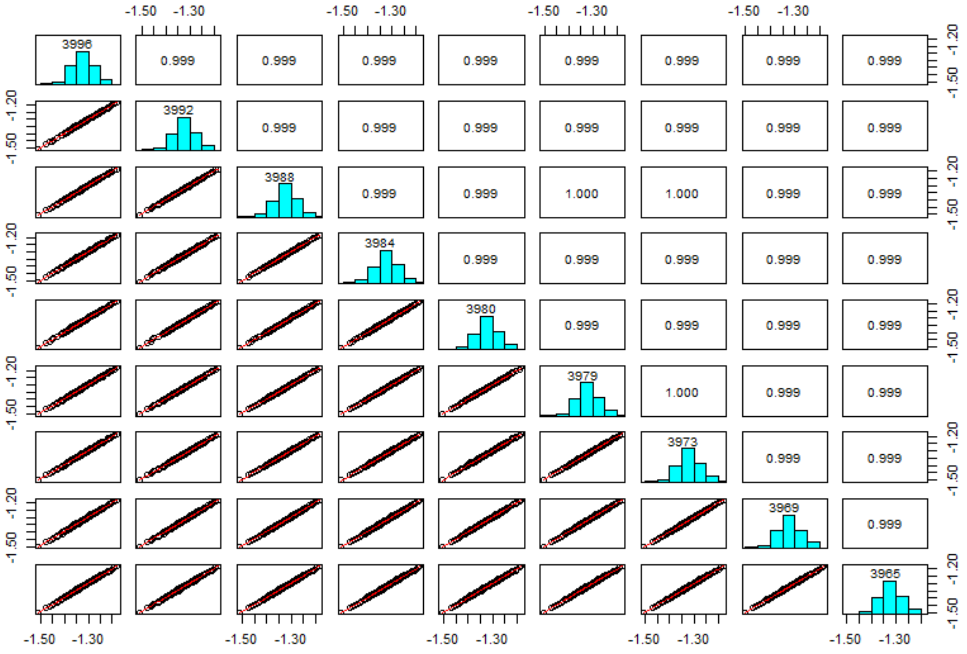


Figure 2.8: Scatterplot matrix of the first nine of 921 MIR spectral bands with associated Pearson Correlation Coefficients displayed in the upper panel

PCA of the MIR data

A standard principal components analysis was carried out for determining the appropriate dimension reduction. Table 2.2 lists the percentage of cumulative variability for the first 12 principal components defined. In all analyses, a singular value decomposition (SVD) algorithm was implemented and following the selection of principal components, a varimax rotation was applied to the resulting loadings matrices in order to achieve simple structure. All analyses were also investigated using the statistics of both the standardized correlation and covariance matrices. However, yielding very similar results, only the use of the covariance matrix is reported.

PC	Standard Deviation	Proportion of Variance	Cumulative Proportion
PC1	19.2	0.409	0.409
PC2	14.2	0.225	0.634
PC3	11.9	0.158	0.792
PC4	7.92	0.0699	0.862
PC5	6.84	0.0521	0.914
PC6	5.02	0.0281	0.945
PC7	4.52	0.0228	0.964
PC8	3.08	0.0106	0.975
PC9	2.88	0.00923	0.984
PC10	2.23	0.00555	0.99
PC11	2.18	0.00528	0.995
PC12	2.10	0.00492	1

Table 2.2: General summary of the PCA on the complete MIR spectra dataset

Based on the scree plot in Figure 2.9, one might choose to retain five principal components since the curve appears to flatten after this point. Opting to retain these five components would explain 91.36% of the total variability, thus achieving the cumulative variance criterion requiring at least 90% of the total variability to be explained. The Kaiser criterion is also satisfied with all components achieving eigenvalues greater than one. However, in order to minimize the risk of information loss it was decided *a priori* to retain subsequent trailing PCs contributing to at least 0.01% of the total variability. By this additional criterion, eight PCs explaining 97.50% of the total variability (Table 2.2) were extracted for further analysis. Each of these eight components define a linear combination of the 921 spectral bands. The weight (loading) of each spectral band is the related number in the eigenvectors. Next, scores plots are visually inspected for potential class separations and for the identification of principal components that contain useful information.

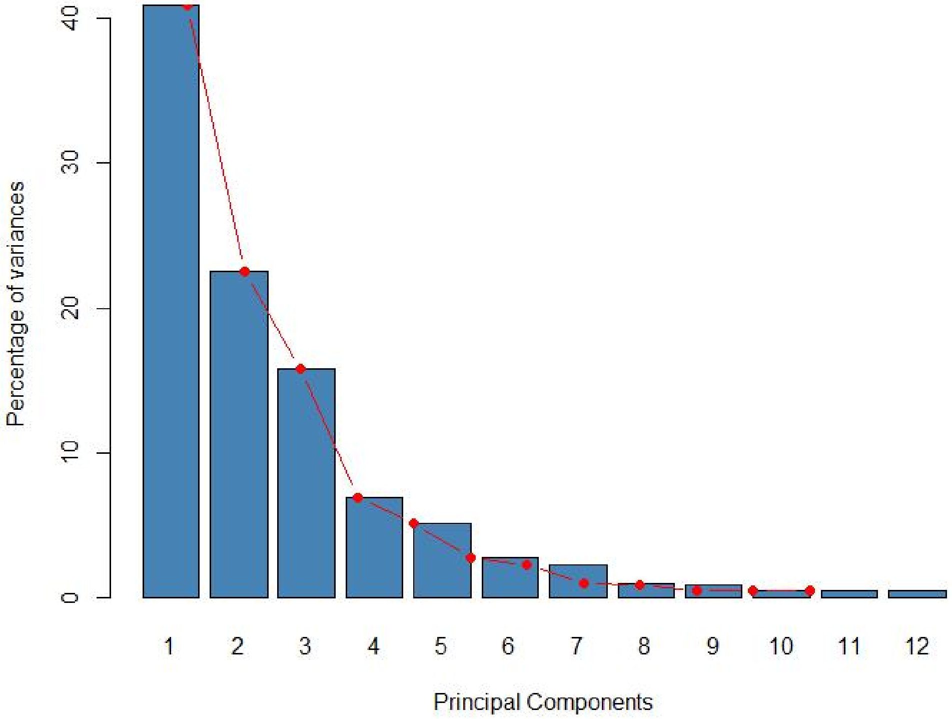


Figure 2.9: Scree plot of the first 12 eigenvalues of the covariance matrix for the complete MIR spectra

Figure 2.10 displays the scores plots in the space of the first eight principal components. The scores are coloured by the site that each soil sample originated from. In inspection of these scores plots there does not appear to be any definite separation between the sites. On the whole, there is no apparently clear group separation in any of the pairwise principal component scores plots displayed. Similarly, scores plots were produced identifying classes of land-use and the site*land-use interaction. However, no definite class separations were apparent in either visualisation. Although it has not been possible to view any class separation in the principal component scores, it is still possible to find the areas of the spectra most responsible for the sources of variation across the spectra. Figure 2.11 gives the loadings plots of all eight principal components retained following the PCA. The higher the loading of a particular spectral variable onto a PC, the more it contributes to that PC.

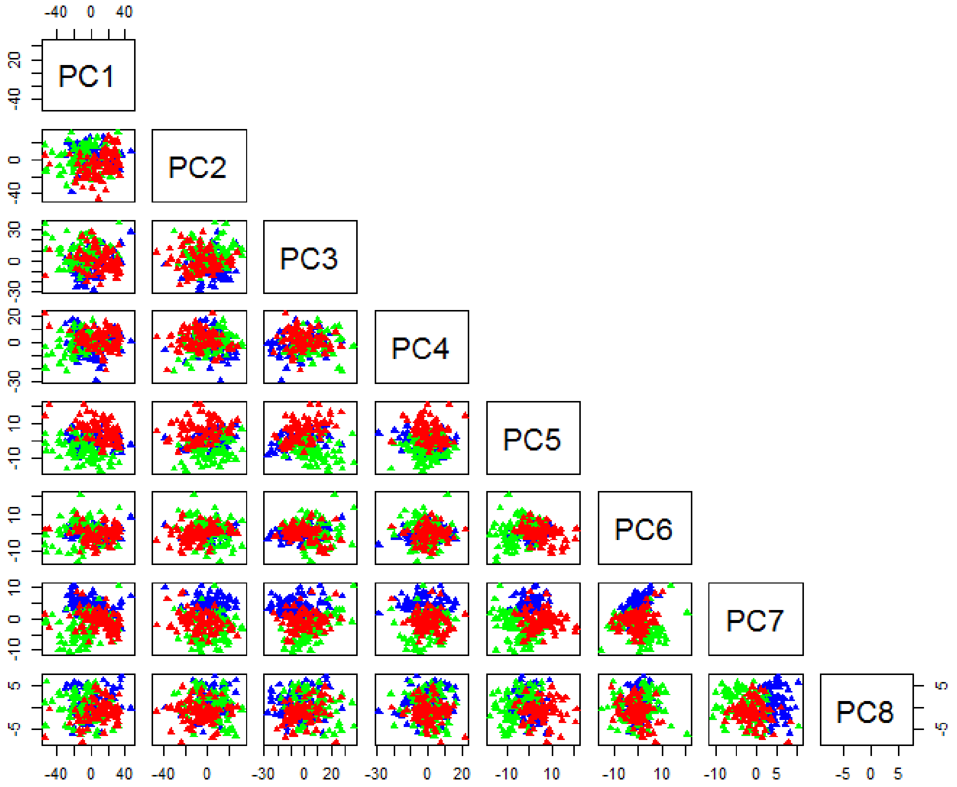


Figure 2.10: Scores plots of PCs 1-8 from the PCA on the complete MIR spectra dataset. Coloured by site. (Bogo: Red, Talmo: Blue, Glenrock: Green)

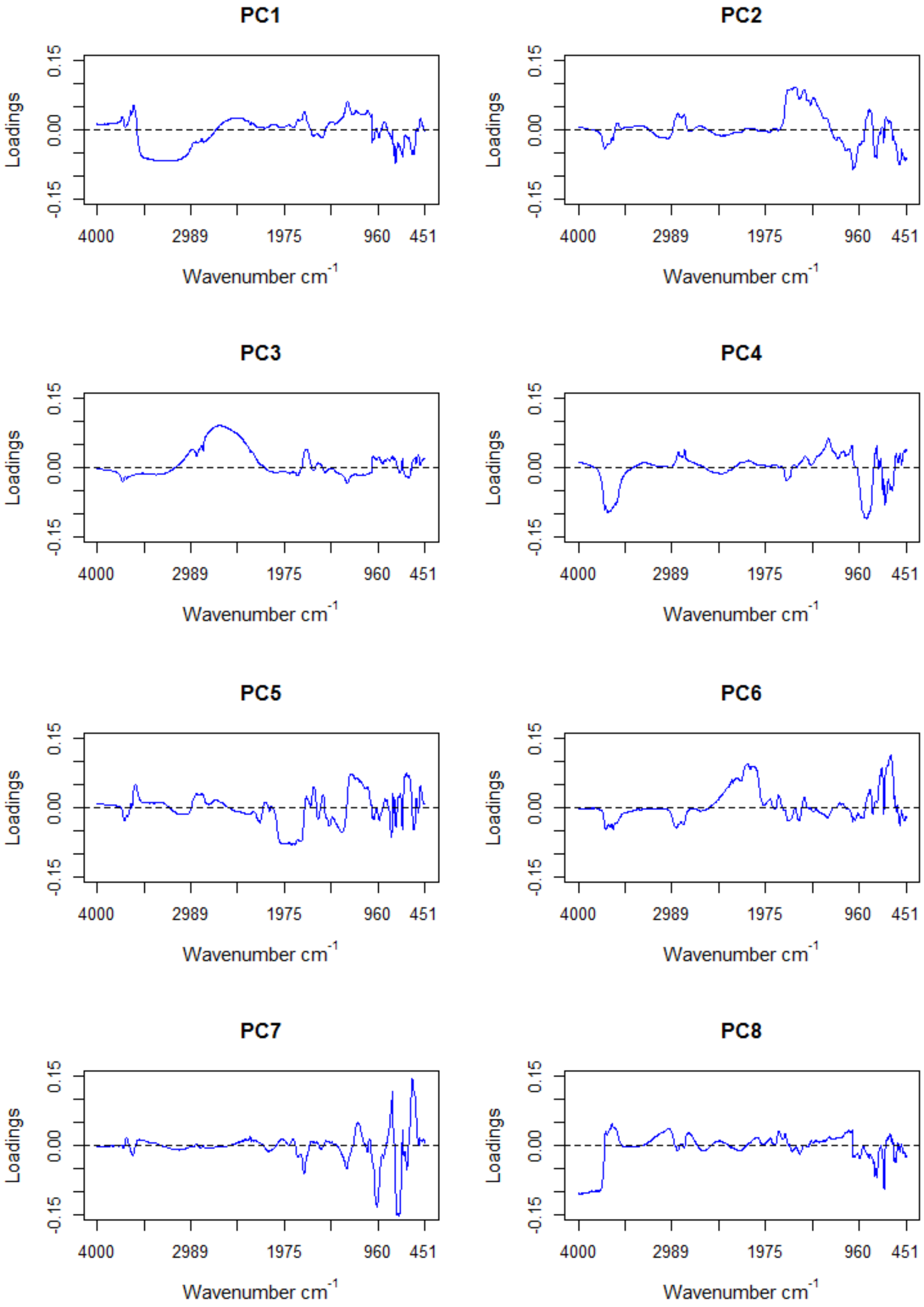


Figure 2.11: Loadings plots for PCs 1-8 of the original PCA on the complete MIR spectra dataset

Since the first principal component is the direction along which there is greatest variation, the loadings plot for PC1 indicates the spectral variables most responsible for the variation across the entire spectra. The first principal component captured 40.9% of the variation and the spectral variables associated with the majority of this variation have been found to lie in the 3500-3000cm^{-1} range. Consulting Table 1.1 of Chapter 1, it appears that these wavenumbers could relate to the spectral absorptions of clay minerals, and in particular Illite would fall within the 3400-3300cm^{-1} range. The second principal component gives the orthogonal direction to PC1 along which the majority of the remaining variation in the spectra is captured. The loadings plot for PC2 indicates that the 22.5% of variation it has captured originates from the 1800-1300cm^{-1} range, and this region could similarly be related to organic matter such as Protein Amide (1670 & 1530cm^{-1}), Carboxylic Acid and aromatic groups (1600-1570cm^{-1}). Similar interpretations can be made for the remainder of the loadings plots. However, since no clean clusterings were observable in the scores plots, these sources of variation cannot be attributed to a certain difference in land-use, site or site*land-use interaction.

PCA of the MIR data: Sampling Site Subsets

The principal component analysis applied to the entire MIR spectra revealed no concretely clear group separation either by land use, sampling site or the interaction between. In this section, the original MIR data are divided into subsets investigating principal component analyses for each farmland site independently. Using the same selection criteria used previously, the results of the PC analyses on the Bogo, Talmo and Glenrock subsets are summarised in Table 2.3.

Site	No. PCs retained	Cumulative Proportion of Variance Explained
Bogo	6	98.35%
Talmo	7	98.29%
Glenrock	6	97.97%

Table 2.3: Number of PCs retained and Cumulative Proportion of Variance Explained for each Site-Specific PCA

The principal component component analyses for Bogo, Talmo and Glenrock reduced their spectral variable data to just 6, 7 and 6 principal components respectively. The original PCA managed to capture 97.5% of the variation in eight components, but the site-specific PC analyses retained less principal components whilst capturing marginally more of the total variability.

Following the same protocol, scores plots were visually inspected for potential class separations. This time there were arguably slight clusterings by land use revealed. With respect to the Bogo PCA, there was notable land-use class separation along the sixth principal component. And there was greatest separation noted in the scores plot between PC1 and PC6. The Talmo and Glenrock PC analyses also revealed some land-use class separation but not as much as for the scores plots of Bogo. The strongest spectral distinction observed in the PCA of the Talmo data was found along the scores plots for its second principal component. Lastly, the PCA of the Glenrock subset revealed land-use class separation along its fourth principal component. Figure 2.12 displays the scores plots of the land-use class separations mentioned for each independent PCA by sampling site.

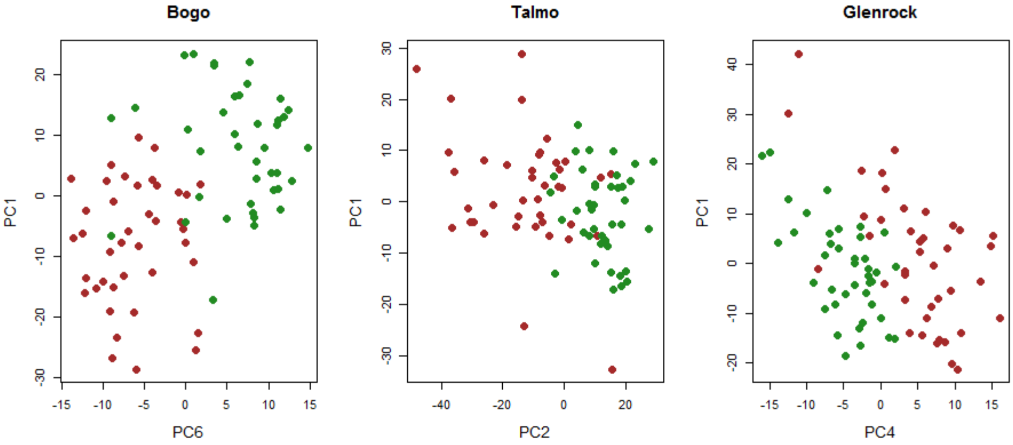


Figure 2.12: Scores plots investigating differences by land-use within each site. Woodland sites are identified in brown and Pasture sites in green.

Figure 2.12 illustrates that the strongest spectral distinction with land-use is evident at Bogo, with lesser land-use differences in the Talmo and Glenrock scores plots. With Bogo, the separation on PC6 is noted with samples from woodland sites (brown) having more positive scores on PC6 compared with pasture sites (green). Talmo displays smaller land-use differences but class separation can be observed along the second principal component with pasture sites giving more positive scores on PC2. Finally, the PCA of the Glenrock MIR data also appears to show some class separation with samples from pastures giving on the whole more negative scores on PC4.

There is potentially useful information contained within the identified principal components and thus there should be an examination of their loadings plots. These loadings plots may reveal the sources of variation across the spectra for driving land-use differences within each farmland site. Figure 2.13 (top) gives the loadings plot for PC6 of the Bogo solution. The plot indicates that the majority of spectral differences originated from regions dominated by mineral signals (3700-3590cm^{-1}). This is an indication that these minerals could be having an influence on the differences observed by land-use in Bogo. Similarly, loadings plots for the second principal component of the Talmo PCA and the fourth principal of the Glenrock PCA have been reported in Figure 2.13. The differences in land-use plots of Talmo appear to be driven by spectral absorptions in the 3500-2900cm^{-1} range, and this also loosely corresponds to mineral signals. The land-use differences of Glenrock are harder to be related to soil components since the loadings plot is not dominated largely by a single span of wavenumbers. The most influential spectral absorptions appear to be scattered in the 900-450cm^{-1} range. However, with only weak class separations observed in the scores plots for Glenrock, these interpretations are very subjective.

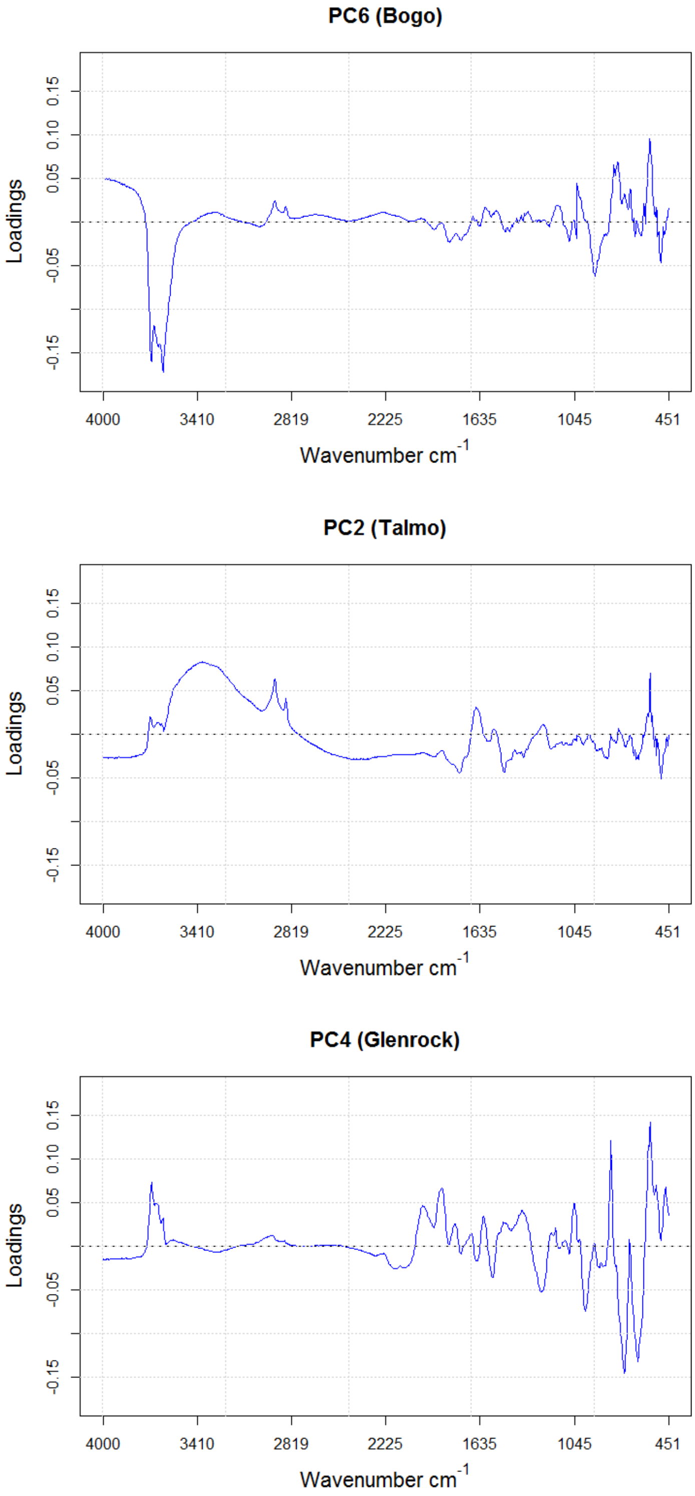


Figure 2.13: Loadings plots of PC6 of the Bogo solution (top), Talmo solution (centre) and Glenrock (bottom). These loadings are used to identify regions of the spectra which dominate PCs. These regions may be related to soil constituents with known absorption regions.

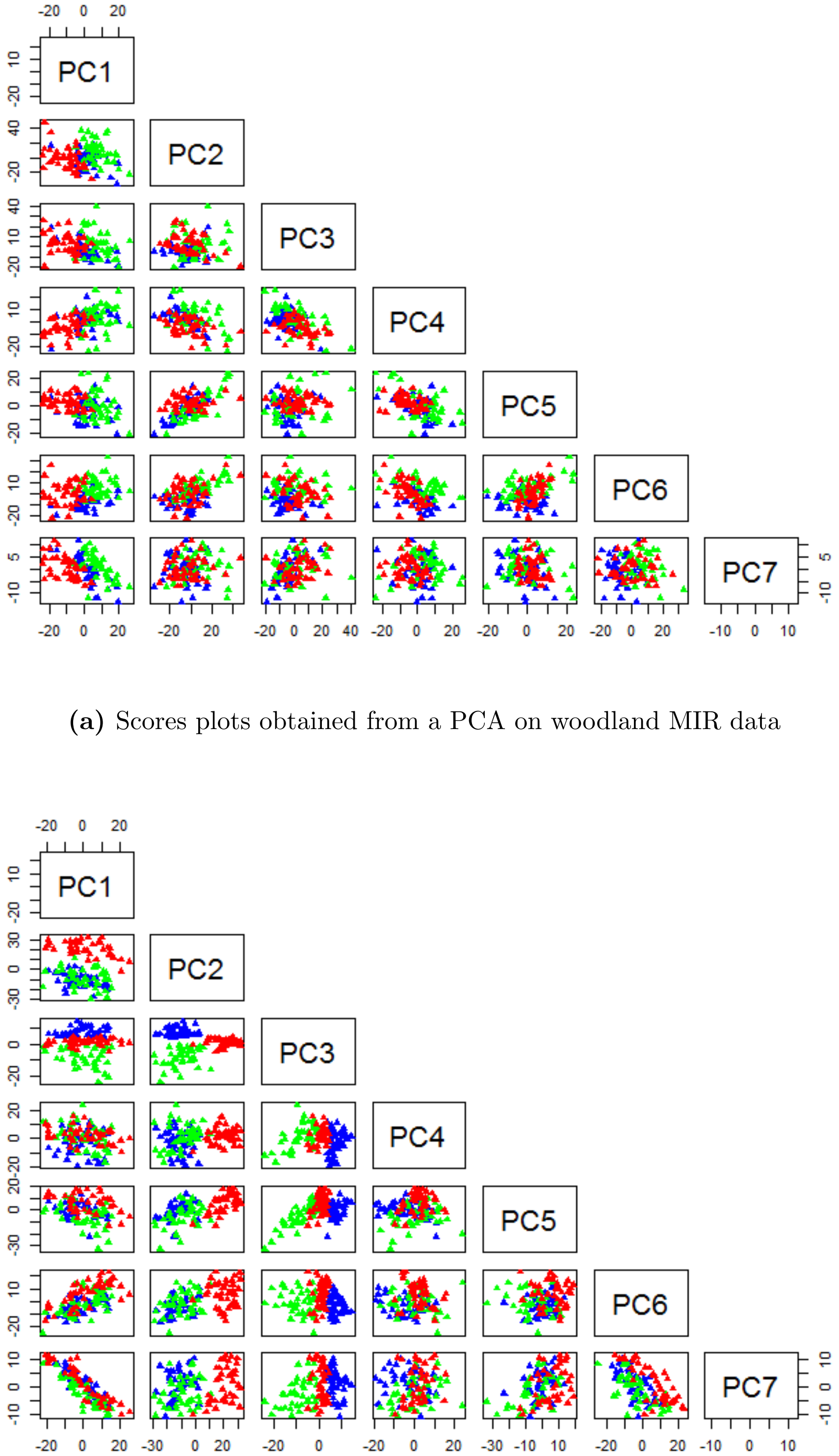
PCA of the MIR data: Land-Use Subsets

Independent principal component analyses were also applied to subsets of the MIR data based on land-use. In both the woodland and pasture cases, PCA successfully reduced the dimensionality of the datasets to just seven principal components and coincidentally they both explained 98.5% of the total variability in the original data (Table 2.4).

Site	No. PCs retained	Proportion of Variance Explained
Woodland	7	98.5%
Pasture	7	98.5%

Table 2.4: Number of PCs retained and Cumulative Proportion of Variance Explained for each Land use-Specific PCA

This time the resulting scores plots were inspected for class differences between the sample sites; Bogo, Talmo and Glenrock. These are presented in Figure 2.14a and Figure 2.14b. In Figure 2.14a, the woodland scores plots do not indicate any clear separation between the scores of different sampling sites for all extracted PCs. This does not mean that there are no differences in the woodland plots of different sampling sites as the differences could just be very subtle. To the eye there are no obvious clusters with a great deal of mixing exhibited. However, on closer inspection there does appear to be some vague separation between the sampling sites along the scores associated with PC1. In green, the scores associated with the Glenrock soil samples tend to be more positive along PC1 whilst scores belonging to Bogo (in red) tend to be more negative along PC1. Scores associated with Talmo (blue) appear to closely lie either side of the zero line but appear to cluster somewhere in between the clouds of Glenrock and Bogo scores.



(a) Scores plots obtained from a PCA on woodland MIR data

(b) Scores plots obtained from a PCA on pasture MIR data

Figure 2.14: Scores plots obtained from a PCA on woodland and pasture MIR data separately. Coloured by soil sample site (Bogo: Red, Talmo: Blue, Glenrock: Green)

Figure 2.15 (top) gives the loadings associated with PC1 of the woodland PCA. It is thought that PC1 may represent the variability in the data that describes the differences between the woodland plots of Bogo, Talmo and Glenrock. The loadings plot indicates that the most important variables for explaining the variability captured by PC1 are the wavenumber variables between 1300-1000cm^{-1}. With the scores plots of PC1 in Figure 2.14a highlighting differences by site, it is possible that soil constituents related to these wavenumbers could be responsible. Of the major soil components listed in Table 1.1 of Chapter 1, quartz is the only soil component which falls within this range. With attention turned to the scores plots corresponding to the PCA of the pasture subset (Figure 2.14b) there also appears to be some distinctive separations observed between scores of different sample sites. The scores plot with perhaps the most obvious display of score clustering can be found between PC2 and PC3. In this particular score plot, Glenrock pasture sites (in green) tend to have more positive scores on PC3 and more negative scores on PC2. Talmo pasture sites represented in blue are more widely spread but on the whole have more negative scores on PC2 and PC3, whilst pasture sites of Bogo (red) show the opposite- clustering tighter with more positive scores on PC2 and PC3.

In the remaining scores plots, there was no detection of class distinction with plots presenting highly mixed scores between different sample sites. It appears that the best class separation is between PC2 and PC3, with discernible distinctions along the scores plots with at least one of these principal components. Thus, it is reasonable to suggest that the loadings plots of these two principal components could indicate the individual wavenumber variables or wavenumber regions responsible for the differences between the woodland plots of Bogo, Talmo and Glenrock. Figure 2.15 also gives the loadings plots for PC2 (centre) and PC3 (bottom) of the pasture PCA. The highest loadings are observed in the 1000-500cm^{-1} range of PC3 and these can be attributed to organic signals. With respect PC2, the influential loadings are quite spread but the highest is observed around 3400cm^{-1} which can be related to clay mineral signals.

Summary of the Principal Component Analyses

The use of principal components analysis has shown that it is possible to significantly reduce the dimensionality of the MIR spectra to a much smaller set of variables. A principal components analysis on the complete set of soil samples did not reveal convincing class separation in the data. However, on applying PCA to subsets based on a sampling site and land-use basis it was found that class separations became apparent. The clearest differences were found not based on geographic differences but by land-use differences within particular sites. However, these differences were small and the PCA projections yielded were not linearly separable. Despite no concretely clear exhibitions of class differences observed in the scores plots this does not mean clean differences do not exist. Differences may be very subtle and whilst principal component analysis extracts the most descriptive information this does not necessarily mean that the variation captured will be responsible for differences between individual spectra.

Overall, principal component analysis has performed successfully as a method of dimension reduction. However, its ability to identify areas of the spectra responsible for driving differences in the classes has been limited. In the next section Linear Discriminant Analysis (LDA), as a supervised learning

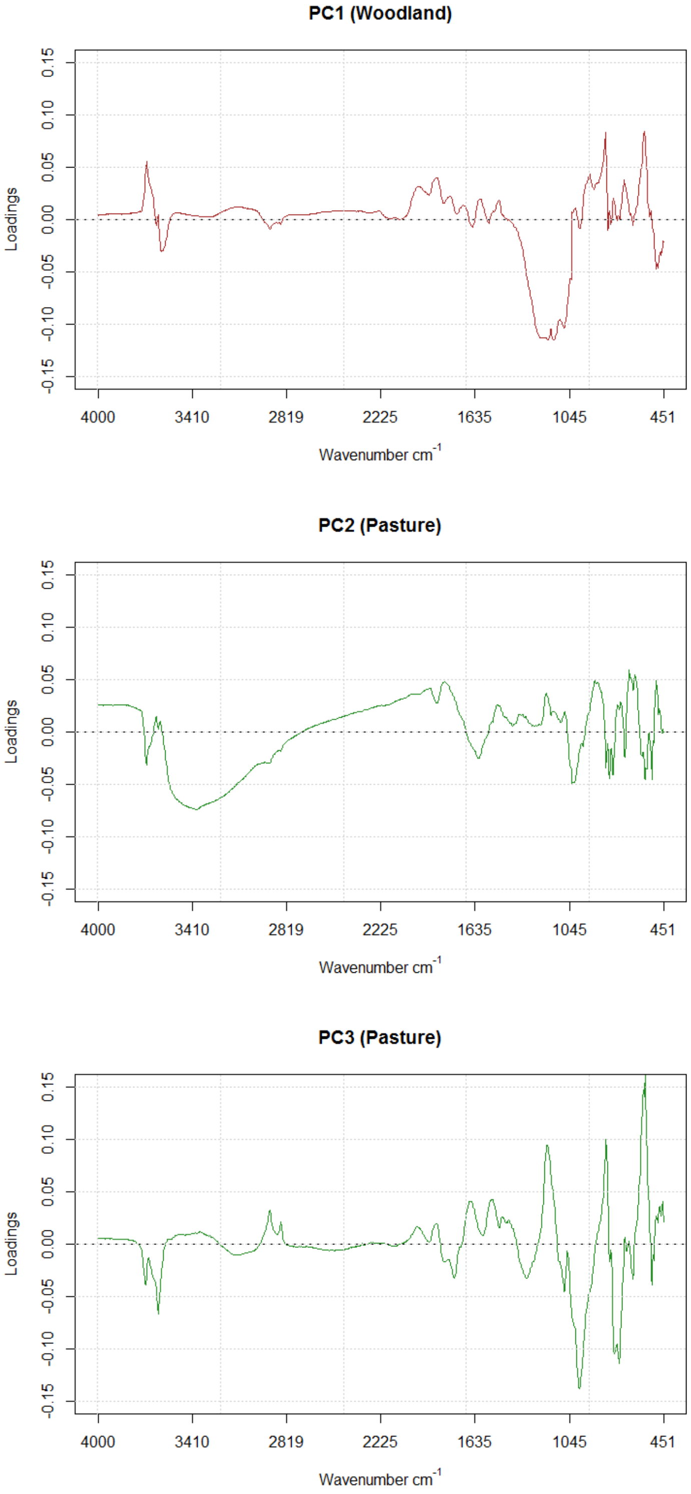


Figure 2.15: Loadings plots of PC1 of the Woodland solution (top), PC2 of the Pasture solution (centre) and PC3 of the Pasture solution (bottom). These loadings are used to identify regions of the spectra which dominate PCs. These regions may be related to soil constituents with known absorption regions.

technique, makes use of class labels known *a priori* and it is expected that LDA should perform well in classifying the soil spectra.

2.2.2 Linear Discriminant Analyses

Linear discriminant analysis searches for linear combinations of the original variables that best discriminate among classes rather than those that best describe the data. More formally, given a number of independent features relative to which the data are described, LDA creates a linear combination of these which yields the largest mean differences between the desired classes (Martínez and Kak, 2001). If effective classifiers can be reliably achieved, then regions of the spectra which drive class separability may be identified.

LDA vs PCA

In contrast to PCA, LDA explicitly attempts to model the differences between classes whereas PCA does not take into account any differences in class. Thus, PCA is often described as an unsupervised algorithm since class labels are ignored and the only goal is to find the directions (principal components) that maximize the variance in a dataset.

Linear discriminant analysis is routinely performed on a set of training data to establish a classification rule before assessing the performance of this classifier on a previously unseen set of test data. This approach will be taken in the next section but first a comparison will be made between the original PCA of the previous section and a linear discriminant analysis performed on the entire MIR soil spectra dataset. Figure 2.16 gives the scores plot between the first and second principal components together with a plot of the first two linear discriminants of an LDA.

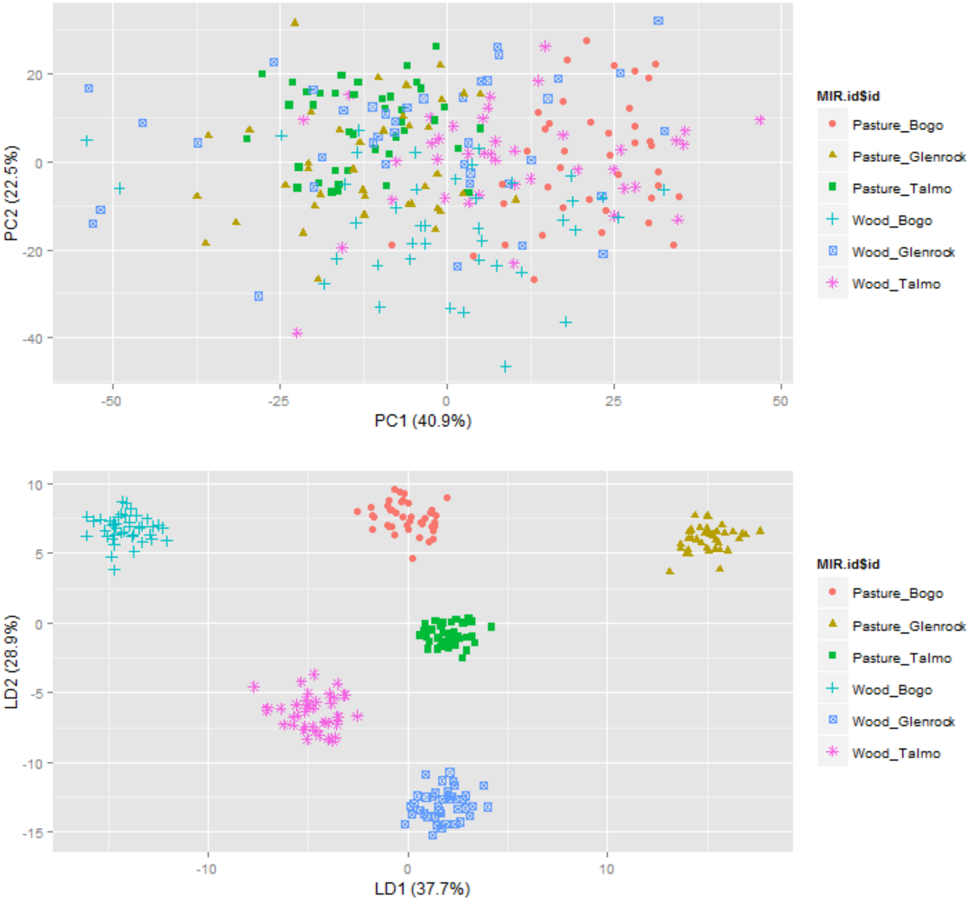


Figure 2.16: Comparison of LDA and PCA for identifying groups of different classes in the data.

The PCA ignored class labels to find the directions that maximise the variance in the data set, and the first two principal components accounted for 40.9% and 22.5% of the total variance respectively. The scores plot revealed no patterns or clustering in any of the classes that can be observed. The classes in this scores plot are identified as the interaction between land-use and sample site. As reported in the previous sections, the results were very similar when considering either land-use or sampling site classes independently. Since LDA takes into account class labels *a priori* the superior class separation observed in the lower panel is not unexpected. Here, the first two linear discriminants cumulatively explain more than 66% of the between-group variance in the data whilst the first two PCs explain 63.4% of the total variability in the data.

LDA: Class Separation problems

The performance of LDA is evaluated for the three different class separation problems using different proportional splits of training and test data. The separation problem between sampling sites involves three different sites, so the number of groups (G) is 3, and the number of variables is equal to the number of uniquely defined spectral bands (921). With three classes, the maximum number of useful discriminant functions that can separate the data is two ($G - 1 = 2$). Similarly, the land-use separation problem gives one discriminant function and the interaction problem gives five.

It is typical for classification problems to make use of training and test data. The training set is a set of data used to build a prediction model which can be employed to predict the outcome or class membership of future unseen objects. A test set is a set of data that is used to evaluate the prediction performance of the classification rule created. By varying the percentages of data allocated to the training and test data subsets then the performance of LDA in distinguishing classes can be evaluated. The performance of the linear discriminant analyses were assessed for 100 random splits into: 50% labelled, 50% unlabelled data; 25% labelled, 75% unlabelled and 10% labelled, 90% unlabelled data. It was chosen to obtain 100 random splits so that average misclassification rates could be calculated and these are reported in Table 2.5. The classification performance of LDA is impressive in all class separation problems giving very low misclassification rates. Unsurprisingly, classification accuracy decreased when a classification rule was based on a smaller

LDA	Training/Test Split	Average Misclassification Rate
	50/50	0.009
Site	25/75	0.034
	10/90	0.127
	50/50	0.0104
Land-use	25/75	0.052
	10/90	0.138
	50/50	0.003
Site*Land-use	25/75	0.028
	10/90	0.200

Table 2.5: Performance of Linear Discriminant Analyses on classifying Site, Land-use and Site*Land-use test data.

set of training data. With less points to train a model on, it is intuitive that a models predictive power would be lower. As expected, all three LDA analyses give much greater performance after training on 50% and 25% of the data compared with just 10%. The classification performance of LDA is greatest separating the data into classes based on the interaction between site and land-use, and poorest separating the land-use classes independently. However, only 1% of the land-use test data are missclassified when using a 50/50 split. The interaction class is separated better than the other class variables in the 50/50 and 25/75 splits, but reducing to a 10% training set and the LDA on the Site*Land-use interaction actually performs the poorest.

Figure 2.17 gives a scores plot between the first two linear discriminant functions of a linear discriminant analysis on the interaction class performed on the full data set (no training and test data split). The LDA has achieved excellent class separation in the space of the first two linear discriminant functions explaining 37.7% and 28.9% of the total variation respectively. It is clear to see that there are definite differences between the spectra belonging to different farmland plots. Figure 2.17 only presents the class separation of the interaction classes. However, it should be noted that LDAs performed to discriminate by land-use and sampling site independently yielded results just as emphatic.

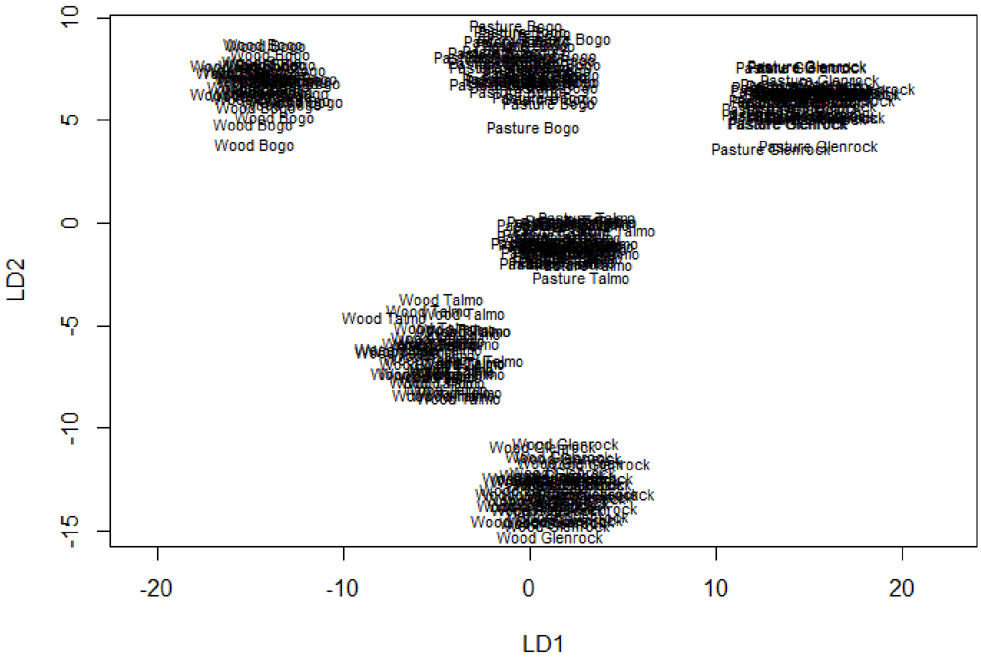


Figure 2.17: Scores plot between the first two linear discriminant functions of an LDA on the Site*Land use class for the full dataset.

The first linear discriminant function explains the majority of the total variance. To find the most influencing wavenumbers for class discrimination, the coefficients can be examined and interpretated similarly to that of PCA loadings. Since each linear discriminant function is a linear combination of 921 wavenumber variables, displaying the full formulations is not appropriate. However, the top five wavenumbers with the highest influence for each class separation problem have been identified graphically in Figure 2.18 and also summarised in order of importance in Table 2.6.

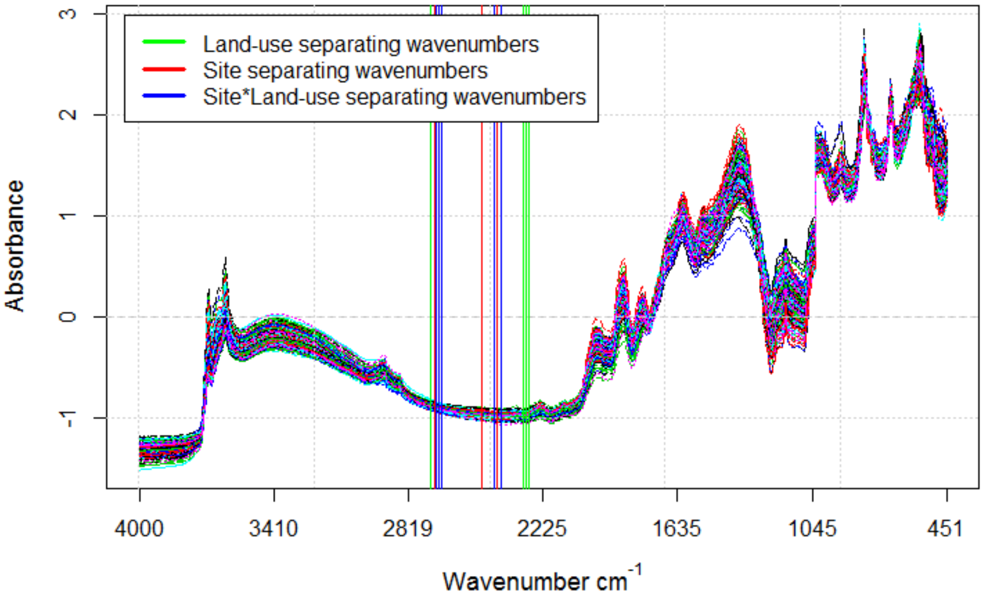


Figure 2.18: The most influencing regions of the spectra as identified by the top five contributing wavenumber variables to LD1 of each class separation problem

Figure 2.18 identifies the wavenumbers that are most influential to class distinction to be constrained within a central region of the MIR spectra ranging from 2720-2290cm^{-1}. This was not anticipated as it would be expected that there would be features across the entire spectral range that would be influencing differences between the classes. Furthermore, discussions with soil scientists at CSIRO prior to analysis informed that this region would not be flagged as it does not contain any elements which would be responsible for differences between groups. It was also suggested from a soil science stand point that the regions of greatest influence should originate from the upper and lower ends of the MIR spectral range. In comparison to other regions which exhibit a higher degree of variability and an abundance of local features, this central region of the spectra is relatively uneventful and uninteresting. There are no spikes over this central range which would indicate the presence of chemicals which would be responsible for differences between spectra of different land-use/site classification. However, these results and the identification of this region could be due to the very strong correlations between neighbouring wavenumbers.

Class Separated	Top 5 Wavenumbers for Class Distinction
Land-use	2680 (1) 2692 (2) 2430 (3) 2492 (4) 2700 (5)
Sampling Site	2290 (1) 2670 (2) 2719 (3) 2310 (4) 2302 (5)
Land-use*Sampling Site	2440 (1) 2680 (2) 2670 (3) 2407 (4) 2692 (5)

Table 2.6: Wavenumbers most responsible for class discrimination (site, land-use, land-use*site interaction) in linear discriminant analyses on the MIR dataset

Stepwise Linear Discriminant Analysis

With the high dimensional nature of the MIR dataset, a more practical method of performing LDA would be to select the most influencing wavenumber variables via a stepwise selection process. Forward stepwise linear discriminant analyses were performed to identify the wavenumbers that contribute most to the class separation of land-use, sample sites and in the interaction case.

In the forward selection process, variables are evaluated based on their contribution to the discrimination between classes with variables having the greatest contribution selected for inclusion first. The stepwise selection process used 10-fold cross-validation and only the most influencing wavenumbers were included if they improved the final correct classification rate by at least 2%. Table 2.7 summarises the wavenumber variables that were added iteratively to the discriminant functions for the three classification problems. The most influencing wavenumbers are listed in order of their selection based on the total amount of variation they explain. The cumulative correct classification rate is also reported for each additional variable retained.

Of the wavenumbers identified, none were common across each linear discriminant analysis. However, with such a high dimensional dataset it is not a surprise that there are no wavenumbers/spectral bandwidths with a common influencing impact on class separation. However, since the wavenumbers are form a continuous spectrum across the mid-infrared range it may be the case that a highly influencing variable lies within a region of high influence for class separation. Figure 2.19 shows the entire range of the mid-infrared spectra highlighting the variables identified from the stepwise linear discriminant analysis which are found to contribute significantly to the class separations between land-uses, sites and the interaction class.

Class Separated	Wavenumber Variables Retained	Classification Rate
Land-use	906.3793 (1)	0.771
	752.102 (2)	0.833
	2410.583 (3)	0.888
	740.5312 (4)	0.908
	3575.377 (5)	0.929
	2850.274 (6)	0.950
	2931.269 (7)	0.988
Sampling Site	3139.269 (1)	0.629
	983.518 (2)	0.775
	698.1049 (3)	0.842
	620.9662 (4)	0.908
	2919.698 (5)	0.950
	3282.25 (6)	0.971
Land-use*Sampling Site	2915.842 (1)	0.450
	2935.126 (2)	0.767
	3185.827 (3)	0.871
	3602.376 (4)	0.938
	3594.662 (5)	0.992

Table 2.7: Wavenumbers most responsible for class discrimination (site, land-use, land-use*site interaction) in linear discriminant analyses on the MIR dataset via a stepwise approach to LDA.

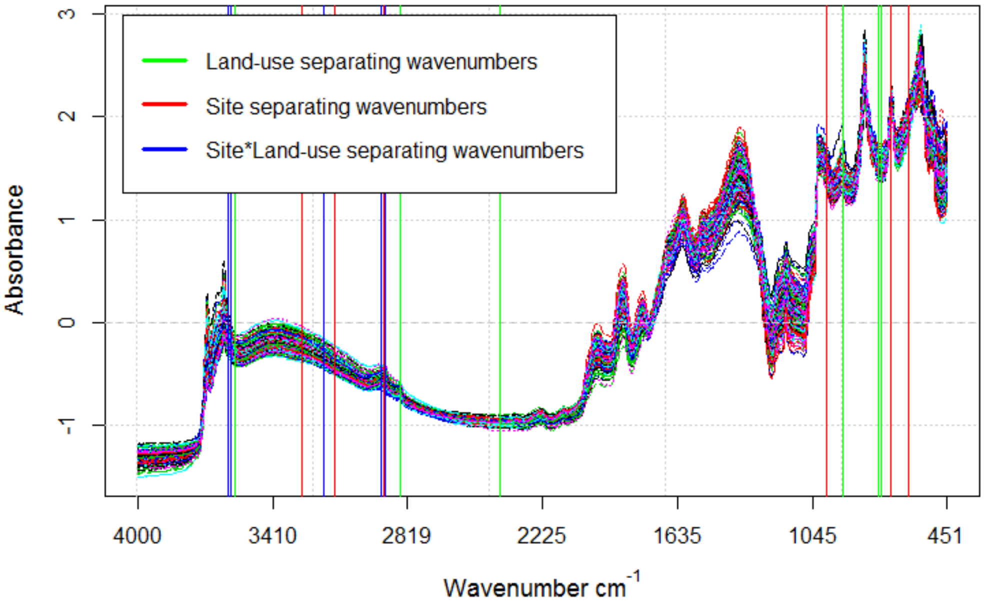


Figure 2.19: The most influencing wavenumbers to the class separation problems as identified by a Stepwise LDA approach

The results of the two approaches to the linear discriminant analyses are very different. The basic linear discriminant analyses indicate that the most important spectral variables for determining class membership of all classes (i.e. Site, Land-use and the interaction) are found in the central region of the mid-infrared range. With the stepwise linear discriminant analyses, the strongest influencing wavenumbers for determining class separations are found scattered across the entire mid-infrared range. The differences in these plots seem unusual as it is not expected that a basic linear discriminant analysis would pick out this central region which appears very uninformative. This central region has the smallest spectral variability in comparison with the rest of the mid-infrared range. There is also a lack of features in this portion of the MIR range and this is confirmed by a close-up in Figure 2.20.

The wavenumbers identified by the stepwise procedure in Figure 2.19 were more expected of this analysis identifying regions of higher variability and with areas with interesting features (peaks/troughs) that differ from their surrounding data. The wavenumbers which seemed responsible for differences between the classes based on sampling site were originally concentrated at wavenumbers of approximately 2300cm^{-1} and 2700cm^{-1}, indicated by the green bands in Figure 2.20.

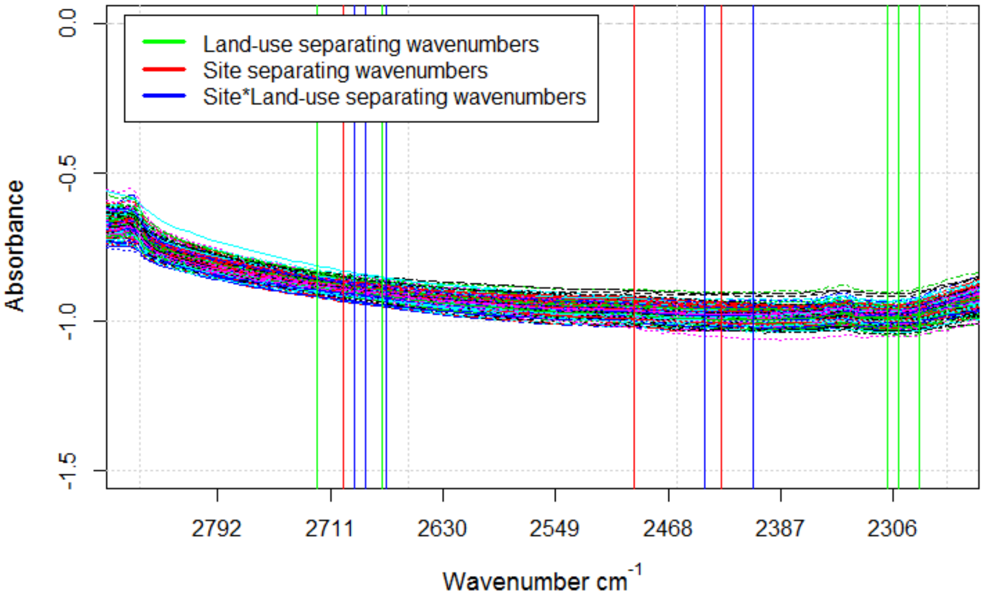


Figure 2.20: Close-up of the central region of the MIR spectra where the basic LDA identifies the most influencing wavenumbers for the class separation problems.

However, with the stepwise LDA it seems that the wavenumbers with the most influence on site differences are found at the following approximate wavenumbers: 3280, 3140, 2920, 985, 700 and 620 cm^{-1}. It seems more reasonable that these wavenumbers pick out areas of the spectra which are much more variable than the central region as identified by the straightforward LDA. However, it may be surprising that only one wavenumber now identifies with this central region via the stepwise approach. A wavenumber of 2410cm^{-1} is identified by the stepwise LDA as having a high influence on the discrimination between spectra of different sites, with additional approximate wavenumbers 3960, 2930, 2850, 905, 750 and 740cm^{-1} similarly identified. With respect to the wavenumbers identified as having influence on the discrimination between the six interaction groups of spectra, it is found that attention shifts to the 3600-2900 cm^{-1} range. The stepwise approach to LDA seems much more appropriate to this problem than the standard case. This is believed since it is much more intuitive that the spectra of different classes would differ across multiple different regions of the spectra rather than confined to a central region of discrimination.

Assumptions of LDA

Although Linear Discriminant Analysis via a stepwise selection procedure has performed well, there are assumptions of LDA that have to be relaxed. LDA requires an assumption that the classes report equal variance-covariance matrices to give accurate results. If these differ considerably then observations will tend to be assigned to the class where the variability is greater. Woodland plots could differ in terms of density of foliage and this would have an influence on the pulls of classification. However, there is literature to support that although relying on heavy assumptions which are not true in many applications, LDA has been proven to be effective (Lim et al., 2000), and this is mainly due to the fact that a simple linear model is more robust against noise, and most likely will not overfit (Gorecki and Luczak, 2013).

2.2.3 Support Vector Machines

The last classification technique explored in this chapter is the machine learning method of Support Vector Machines. Support Vector Machines (SVMs) are a fairly new method of classification for both linear and non-linear data. The popularity of SVMs rests firmly on their ability to non-linearly map original data points into higher dimensions. In the higher dimension, an optimal separating hyperplane can be found which defines decision boundaries separating data based on their class labels.

Hyper-parameter selection

The most criticial step for support vector machines is the tuning of the SVM hyperparameters and these are classically optimized using an exhaustive search algorithm or grid search (Devos et al., 2009). Since the MIR data are not linearly separable, the data were non-linearly mapped to a higher dimension by the RBF kernel function. With this kernel, there are two hyper-parameters which must be considered, the cost (C) and a gamma (γ) value. The goal is to select well-performing C and γ hyper-parameters such that the resulting classifier can accurately predict future unknown data. However, it may not be useful to achieve very high training accuracy as this could lead to overfitting and a classifier which is not generalizable.

In order to identify the best hyper-parameters a grid search on C and γ was performed using cross-validation. There are often high computational costs associated with exhaustive grid-searches. However, with only two hyper-parameters to locate the computation is still relatively quick and the grid-search is favoured over other approximate methods. The performance of various SVM classifiers are evaluated similarly to the linear discriminant analyses using different training and test data splits for the three class separation problems. Table 2.8 summarises the hyper-parameter values identified for each SVM by 10-fold cross validated grid searches.

Table 2.8: Values of the hyper-parameters as selected by an exhaustive grid search

Separating Classes	Cost	Gamma (γ)
Land-use	10	0.001
Site	10	0.0001
Land-use*Site	10	0.001

In the land-use classification problem, the best model in the parameter range is obtained using $C = 10$ and $\gamma = 0.001$. A graphical overview of these tuning results can be obtained by creating a contour plot of the error landscape, and this is presented in Figure 2.21. The areas of deepest blue show the areas where the optimal hyper parameters may be located and thus where the misclassification rates are lowest. The top and bottom plots show the same cost and gamma parameters for the land-use classification problem. However, the lower plot represents possible gamma hyper parameters through a log10 transform to improve the graphical presentation.

SVM Performance

Having identified the best hyper parameters, SVM classifiers were then trained on various splits and their predictive accuracy assessed on a test split of the data. The performance of the SVM classifiers were assessed in the same manner as with LDA. That is, assessment was carried out based on 100 random splits into: 50% labelled, 50% unlabelled data; 25% labelled, 75% unlabelled

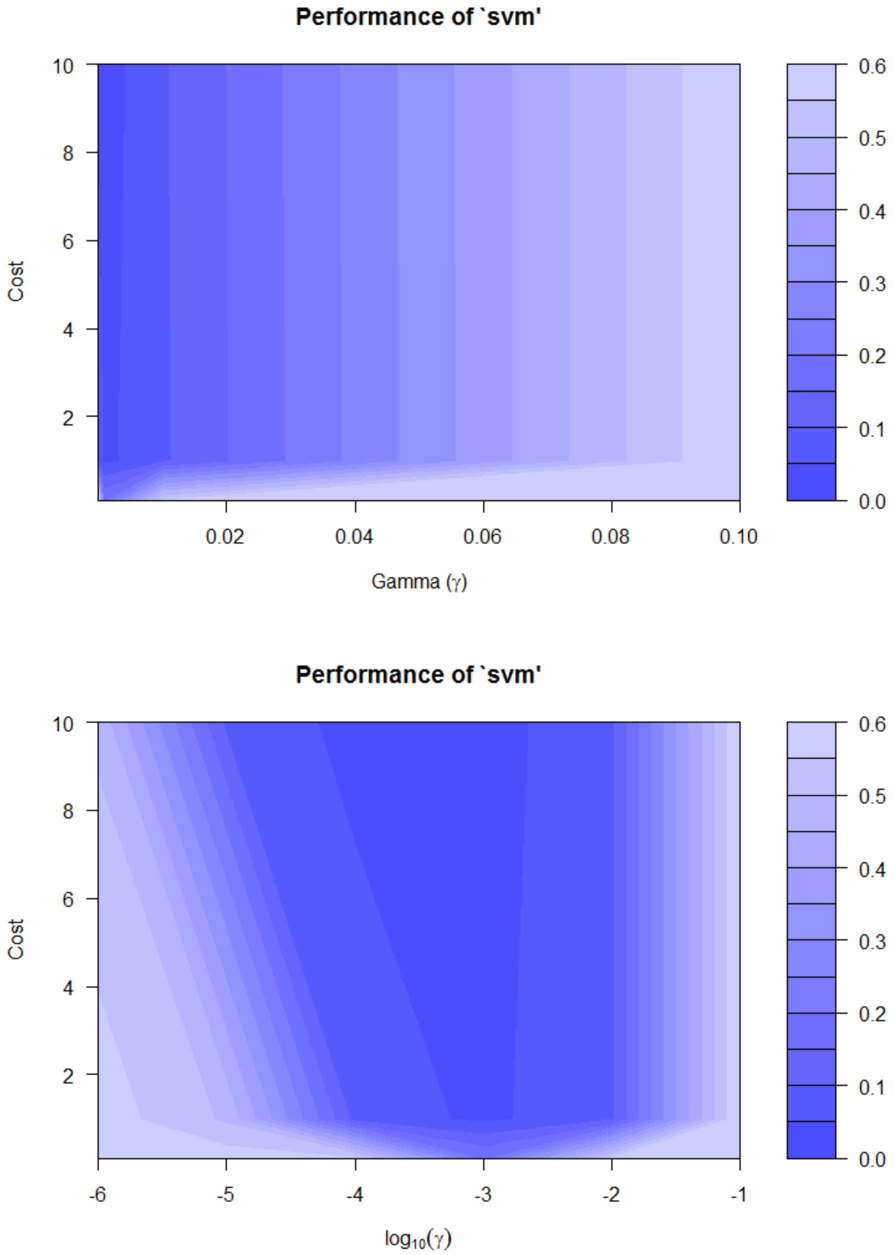


Figure 2.21: Contour plots illustrating error landscape resulting from a hyper parameter grid search for the land-use classification problem

and 10% labelled, 90% unlabelled data. The average misclassification rates are reported in Table 2.9.

SVM Classifier	Training/Test Split	Average Misclass. Rate	LDA error rates
	50/50	0.039	0.009
Site	25/75	0.083	0.034
	10/90	0.195	0.127
	50/50	0.009	0.0104
Land-use	25/75	0.008	0.052
	10/90	0.009	0.138
	50/50	<0.001	0.003
Site*Land-use	25/75	<0.001	0.028
	10/90	<0.001	0.200

Table 2.9: Performance of SVM Classifiers evaluated with average misclassification rates and LDA error rates for various training and test data splits for each class separation problem.

The optimal value for the cost hyper-parameter was $C = 10$ for all classification problems. The optimal gamma value was found to be $\gamma = 0.001$ for the land-use and land-use*site classification problems, but $\gamma = 0.0001$ for the site classification problem. In the site problem, for a 50/50 training and test data split the misclassification rate yielded is 0.0393. As the training sets get smaller the misclassification rates increase, and this is true for all the SVM classifiers. Overall, misclassification rates are very low for the SVM classifiers. For example, the SVM classifiers built for the interaction class discrimination misclassify less than 0.1% of the test data for all training/test splits. The misclassification rates are most variable for the SVM classifiers built for the discrimination of spectra by sampling site. However, with only a 25% training data set the associated SVM classifier still only misclassified 8.3% of the test set. The SVMs built to discriminate the spectra by their land-uses performed very well across all training and data splits- correctly classifying over 99% of the test data in all cases. On the whole, the SVM performance in comparison with that of the linear discriminant analyses of the previous section has been far superior in the class separation problems involving land-use and the interaction class. However, the LDA corresponding with discriminating the sampling site of spectra achieved better classification than the SVM method.

In SVM classification problems, it is common to provide SVM classification plots to illustrate the non-linear decision boundaries and class regions. However, given the high dimensional nature of the MIR data and the high performance of the SVM classifiers, pairwise plots of variables were uninformative. This is a major downfall for the application of SVMs. Although SVMs have been found to achieve high performance in discriminating between the different classes of spectra, it is of no utility since the method has been unable to give any indication of where the differences lie. Whilst this black box nature of SVMs allows for easy implementation, it fails to identify areas of interest along the MIR spectra. However, the application of SVMs has given sufficient evidence to confirm that differences between groups in all classes of the spectra do exist.

2.2.4 Summary

This chapter investigated the applicability of standard multivariate techniques for the purpose of explaining the variation in the spectra across the mid-infrared range and for the classification of groups of soil spectra.

Principal component analysis successfully reduced the MIR dataset from 921 wavenumber variables to just eight principal components. This PCA achieved a massive reduction in dimensionality whilst preserving 97.5% of the total variability found in the spectra. Following examination of the scores plots there were no substantive differences between any of the classes and thus PCA only managed to function as a dimension reduction technique in this case. However, on application of PCA to subsets of the data there were some class distinctions detected. These PCAs summarised the original data in just six or seven principal components whilst explaining upwards of 97% of the total variability. It was found that the PCAs on spectra of individual sampling sites revealed some land-use class separation in the scores plots and inspection of the associated loadings plots managed to give some direction towards potentially responsible regions of the spectra. Within these regions, signals were related to components of the soil which could be responsible for driving the differences between classes (i.e. site, land-use, site*land-use). In all cases the clusters in the scores plots were not linearly separable and in some cases the differences were slight. However, given the initially high degree of dimensionality and the density of the spectra it should be expected that any differences between classes would be subtle. Furthermore, since principal component analysis extracts the most descriptive information then the variation captured does not necessarily translate to variation in the data responsible for differences in spectra of different classes. Whilst the multivariate PCA has been very successful as a dimension reduction technique, the interpretations of the loadings plots in order to identify regions of distinction between classes may be fallible. The interpretations are highly subjective and require the input of a soil science expert.

With additional class information supplied *a priori* , LDA was extremely effective in predicting future class membership of unseen data. This was demonstrated for various training and test data splits for the three class separation problems yielding very low misclassification rates across the board. Just as the loadings were interpreted in the principal component analyses, the coefficients in the linear discriminant functions were examined to identify wavenumbers with the greatest influence on class discrimination. The influencing wavenumbers as identified by a straightforward LDA were compared with those similarly acquired from linear discriminant analyses following a stepwise approach. The original LDA identified only spectral bands in the central region of the mid-infrared range (2750-2230cm^{-1}). It would be unusual for this region of the spectra to influence differences in class discrimination since the absorptions in this range are not dominated by either mineral or organic signals.

In study of the stepwise LDAs, it may be suggested that the original LDA approach is naïve. With the stepwise linear discriminant analyses, the strongest influencing wavenumbers for determining class separations are found scattered across the entire mid-infrared range. The wavenumbers identified by the stepwise procedure were more expected of this analysis identifying regions of higher variability with interesting features (peaks/troughs) that differ from their surrounding data. With a wide array of chemical moieties found in soil, it is likely that a greater number of spectral bands relating to these properties will have an influence on discriminating between the classes of different land-uses, sample sites and the interactions between. Additionally, with different chemical moeities or functional groups associated with different frequencies, it is expected that regions of interest would be scattered across the full range of the spectra.

Despite the high performance of LDA, the interpretation of the results must be taken with caution. By design, both LDA approaches identify specific wavenumbers which are deemed to be important in the separation of spectra of different classes. It is more likely that the components of soil subjected to spectroscopy are defined by multiple wavenumbers or a range of wavenumbers, and that these wavenumber ranges may be partly shared by other soil constituents. Thus, with the identification of a singular wavenumber variable it may be hard to relate to a specific soil component.

Support vector machine methods achieved excellent classification performance. SVMs successfully mapped the MIR data to a higher dimension in which optimal separating hyperplanes could be constructed to separate the data into the appropriate classes. Achieving very low misclassification rates, SVMs outperformed the linear discriminant analyses on the most part but visualisations identifying the most influential regions of the MIR spectra for class discrimination could not be produced. This is a major downfall of the SVM approach in the high-dimensional setting as it was unable to give any indication of where class differences lay. However, the application of SVMs has provided additional evidence that there do exist regions of the mid-infrared range whereby groups of spectra belonging to different land-uses, sites and interaction class differ from one another.

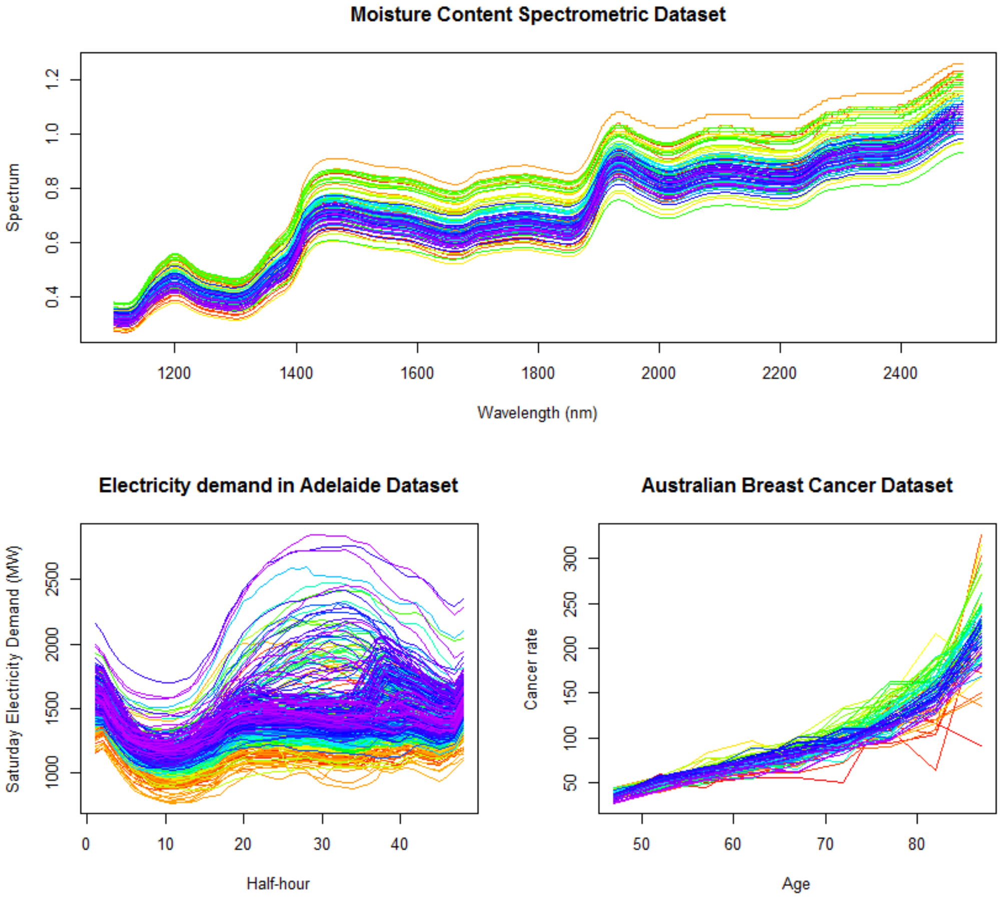
Chapter 3

Functional Data Analysis with MIR soil spectra

First named by Ramsay and Dalzell (1991), Functional Data Analysis (FDA) has received a lot of attention in recent years. This is due to modern technology that now has the capabilities to generate unprecedented amounts of high-dimensional data in many scientific disciplines. The data being observed are not the standard multivariate observations of classical statistics but are observations recorded as curves (functions) or images along a continuous domain. This domain for functional data is usually time but can be anything continuous such as distance, space, frequency and age. The key assumption in FDA is that the underlying process which generates the data is smooth. However, the data are still only observed at discrete points which are subject to measurement error. Thus, an integral part of functional data analysis is the smoothing so that adequate functional data representation can be achieved. Furthermore, FDA makes no assumption of independence between adjacent observations- an assumption often violated in multivariate analysis.

In functional data analysis, the idea is to view each curve or replication as the single observation or entity of interest. To get a feel for functional data, Figure 3.1 displays three different functional datasets from various fields. The top panel represents a dataset of near-infrared reflectance spectra of 100 wheat samples measured in 2nm intervals from 1100 to 2500nm with the associated response variable being the samples' moisture content (Kalivas, 1997). The bottom left panel gives half-hourly electricity demands for Saturdays in Adelaide from 1997-2007 where Magnano et al., 2008 performed analyses to test whether or not, under different temperature scenarios, there would be enough capacity to satisfy electricity demands in the future. The bottom-right panel gives the age-specific cancer rates for Australian females with data provided from the Australian Institute of Health and Welfare (AIHW), an organization that that provides anonymised health and welfare data to national and regional government and community organizations (Erbas et al., 2007). The crude age-specific mortality rates are used which are defined as the number of deaths in a particular age group during the year by the corresponding population in that age group at 30 June of the same year. The rate is expressed per 100,000 people. These plots demonstrate that functional data is often complex with a large number of related quantities not easily described by mathematical formulae. There is also variation between replications which may be hard to explain.

It is becoming more and more common to work with large datasets, and with a wide range of potential applications under the functional data framework considerable efforts have been made in adapting classical statistical methods (Tarrio-Saavedra et al., 2010). For example, principal component


Moisture Content Spectrometric Dataset

Electricity demand in Adelaide Dataset

Australian Breast Cancer Dataset

Figure 3.1: Examples of functional data from various fields of science. *Note.* Original data from Kalivas (1997), Magnano et al (2008) and Australian Institute of Health and Welfare (AIHW) website at http://www.aihw.gov.au/cancer/data/index.cfm.

analysis for functional data is studied by Locantore et al. (1999), regression models with functional covariates are analysed by Cardot et al. (1999) and functional data classification is another important field (Ferraty and Vieu, 2003). Functional data classification includes both supervised and unsupervised classification, and Support Vector Machines have even been adapted (Rossi and Villa, 2006). Some of these methods are straightforward extensions of existing techniques, while others are more complex. Later in this chapter, some of these functional equivalents are introduced.

One of the biggest challenges that comes with FDA is the estimation of functional data from the (potentially) noisy discrete observations. In any functional data analysis, the raw data are also unlikely to be regular in nature and thus an essential first step is the estimation of smooth functions from the observed data. Once these functions are obtained, the original discrete data can be discarded and only the functions are used in subsequent analyses. There are a variety of smoothing methods available but Ramsay and Silverman (1997) advocate that B-splines are most commonly used in functional data analysis than any other smoothing approach due to their high degree of flexibility and computational efficiency.

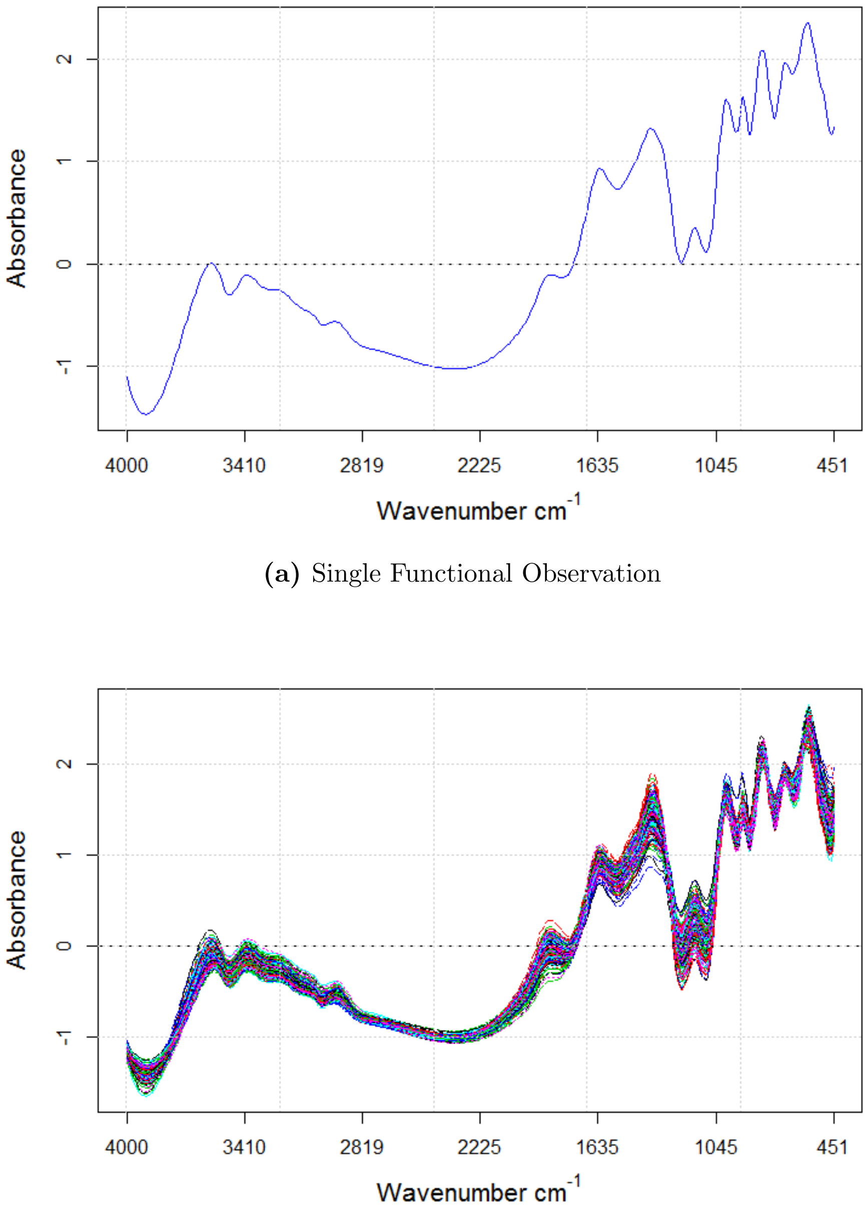
In this section an introduction to the concept of functional data is provided alongside the general underlying theory of some popular FDA statistical methods. In further sections the application of FDA methods to soil spectroscopy is explored. The methods explained in the following sections are based on Ramsay and Silverman (2005). For a more detailed reading, *Functional Data Analysis* by Ramsay and Silverman (2005) gives a manageable overview of the foundations and possible applications of FDA, and their earlier *Applied Functional Data Analysis* (2002) provides many further examples of functional data applied in R statistical software, most often with continuums of age or time.

3.1 Functional Data Theory

According to Ferraty and Vieu (2006), a random variable X is called a functional variable if it takes values in an infinite dimensional space. A sample $\mathbf{X} = X_1, \ldots, X_n$ is called functional data when the i^{th} observation is a real function $\mathbf{X_i}(t), t \in J, i = 1, \ldots, n$, and hence, each $\mathbf{X_i}(t)$ is a point in some function H. In order to avoid confusion, a single functional datum meaning a single observed function is referred to as a replicate. A functional dataset is thus a random sample of replications (Figure 3.2). The argument t is the continuum along which each of these functions or replications is measured at discrete points. Although t is often represented as time, this continuum can be any such continuous domain.

With such continuums, observations are often dependent on adjacent observations and this is commonly true of temporal data bringing about temporal correlation. This correlation between adjacent observations causes problems with classical multivariate analyses violating the assumptions of independence. However in FDA, no such assumption is made and each observed curve is thought of as a single observation rather than a collection of individual observations. Regarding the data in this way also makes it easier to observe common long-term patterns.

Over the last two to three decades, methods of FDA have been applied to functional datasets within medicine, econometrics, biostatistics, environmetrics, geophysics and chemometrics. More generally, examples of functional data can be found in multiple time series analysis where each observation is a complete time series. A number of excellent illustrations of the applications of FDA can be found in Ramsay and Silverman's book (2005)



(a) Single Functional Observation

(b) Functional Dataset

Figure 3.2: An example of functional data generated from the mid-infrared soil spectral data. The top plot gives a single functional observation originating from a single sample. The bottom plot gives all 240 smoothed spectra as the entire functional dataset. Smoothing has been achieved by a generalised cross-validation approach to selecting the number of basis functions.

where data on gait, handwriting, yearly weather data and the growth of children are explored.

3.1.1 Modelling Functional Data

Despite the i^{th} observation in a functional data analysis being a real function, functional data are most often observed and recorded discretely. Typically, there will be a sample of n independent replications, i.e. curves, and a record of replication $X_i(t), i = 1, \ldots, n$ consisting of n_i pairs $\{t_{ij}, n_{ij}\}$, $j = 1, \ldots, n_i$, where t_{ij} denotes the argument, and y_{ij} the observed functional value. These argument values may vary between replications and do not necessarily have to be equally spaced. There must be caution taken when interpreting independence under the functional framework- it should be noted there is independence assumed between replications (curves) but not over the continuum. The number of observations, n_i, are also permitted to differ between replications, but the argument values should fall within the same range of values of interest, i.e. $t_{ij} \in J$ for all i, j.

Like most observed data, usually functional observations are observed with some form of error or noise. Let $\mathbf{y_i} = (y_{i1}, \ldots, y_{in_i})^T, \mathbf{t_i} = (t_{i1}, \ldots, t_{in_i})^T$, and $\epsilon_\mathbf{i} = (\epsilon_{i1}, \ldots, \epsilon_{in_i})^T$. Then a functional data model can be defined as follows:

$$\mathbf{y_i} = X_i(\mathbf{t_i}) + \epsilon_\mathbf{i}$$

where ϵ_i is assumed to be the error term with $\epsilon_\mathbf{i} \sim N(0, \sum_i)$ and $\sum_i = diag(\sigma_{i1}^2, \ldots, \sigma_{in_i}^2)$.

3.1.2 Smoothing and the Basis Function Approach

In representing functional data as smooth functions a flexible method is needed that can track local curvature and minimize the short-term deviations due to observational errors, such as measurement errors or inherent system noise (Ullah and Finch, 2013). One such smoothing procedure is the basis function approach. According to Ramsay and Silverman (2005) a basis function structure is a set of known functions, $\phi_k(t)$, that are mathematically independent of one another combined and used to estimate a function, $x(t)$. Each functional datum is represented by a finite basis such that an explicit form for the function is obtained as follows:

$$y(t) = x(t) + \epsilon(t) = \textstyle\sum_{k=1}^{K} c_k \phi_k(t) + \epsilon(t),$$

The term c_k represents the basis coefficients, t is time (or any other continuum) and K is the number of basis functions. In this way, any function can be approximated by a linear combination of these basis functions.

There is a large number of basis types to choose from and no basis is considered optimal. However, basis functions should be chosen to represent the characteristics of the functional data. For instance, Fourier basis functions are a good choice for data with a periodic nature and spline basis functions are more suitable for non-periodic data. It is preferable that the basis functions and functions to be approximated should have similar characteristics in order to make it more straightforward to attain a sufficient approximation using a relatively small K (Ramsay and Silverman, 2005).

Two of the most popular choices are Fourier basis systems and B-spline bases. B-splines are additionally advantageous since they can model sharp changes in the underlying function as well as its smooth variation. They also have fast computation times favouring their extensive usage within functional data analysis due to the frequent high dimensionality of datasets. The closer the features of the basis functions are to those of the data the better the estimation of the function $x(t)$ will be.

The B-Spline Basis System

Splines are the long flexible strips of wood that shipbuilders bend and hold at control points and their elasticity allows the curved shape of the ships hull to be formed. A spline smoother is analogous to this definition as it consists of linear fits bent at particular control points called knots (Schoenberg, 1964). Splines are non-parametric or semi-parametric techniques for fitting curves to data, and the term was first coined by Schoenberg (1946) describing the process of fitting a smooth function to data.

While there are many other smoothing methods available, B-splines have been especially popular due to their easy implementation and their greater flexibility with respect to other spline methods. However, it was not until the 1970s that B-splines rose in popularity and this was due to growing computing power and stable boundary estimates. The boundary problem was solved by De Boor (1972) who detailed a mathematically stable formula for calculating B-splines using the concept of divided differences and recursively higher-order splines.

The interval along which basis functions are calculated can be divided into L subintervals by the values $\tau_l(l = 1, \ldots, L-1)$ called breakpoints or knots. The B-spline is a piecewise polynomial function of order m over each interval, which is smoothly connected at these breakpoints. The order m is the number of parameters used to define a function, and is defined to be one more than the degree of a polynomial- where degree represents the highest degree in a polynomial. For example, a cubic spline has a degree of 3 and order 4.

A B-spline, $\phi_K(t)$, of order m and knot sequence, τ, is given by

$$\phi_K(t) = B_K(t, \tau), K = 1, \ldots, m + L - 1$$

where K refers to the number of the largest knot at or to the immediate left of value t. They are created using the following recursive formula as developed by de Boor (1972):

$$N_{i,1}(t) = \begin{cases} 1, k_i \leq t \leq k_{i+1} \\ \\ 0, \text{otherwise} \end{cases}$$

$$N_{i,\delta}(t) = \frac{t - k_i}{k_{i+\delta-1} - k_i} N_{i,\delta-1}(t) + \frac{k_{i+\delta} - t}{t_{i+\delta} - t_{i+1}} N_{i+1,\delta-1}(t)$$

where $N_{i,\delta}(t)$ is the basis function evaluated at t_i, $k_1, \ldots, k_n$ are the knots; and δ is the order of the basis function being calculated. The recursive nature of the formula is illustrated in Table 3.1. The basis functions are recursively calculated by first calculating the lower degree splines. $N_{i,\delta}(t)$ is greater than zero only at δ knots; it is zero everywhere else.

Table 3.1: The recursive nature of de Boor's (1972) formula

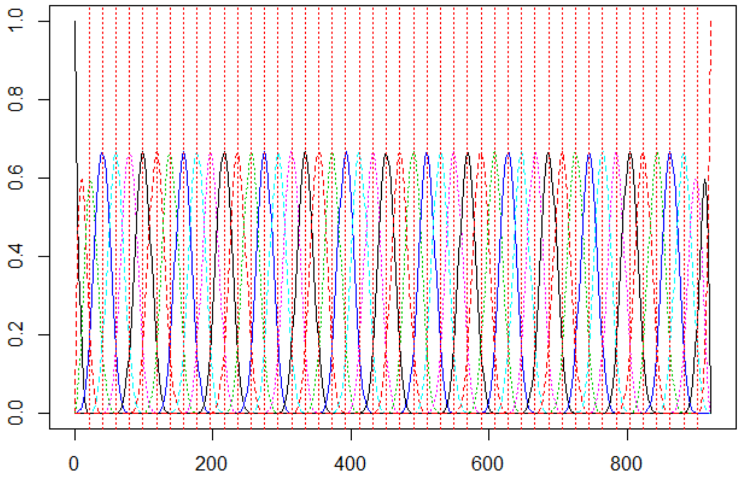
	Order 1	Order 2	$\ldots$	Order δ
k_{i-1}	0	0	$\cdots$	0
k_i	$N_{i,1}(t) = 1$	$N_{i,2}(t)$	$\cdots$	$N_{i,\delta}(t)$
k_{i+1}	0	$N_{i+1,2}(t)$	$\cdots$	$N_{i+1,\delta}(t)$
$\vdots$		0	$\ddots$	$\vdots$
$k_{i+\delta}$			0	$N_{i+\delta,\delta}(t)$
$k_{i+\delta+1}$				0

Placing Breakpoints

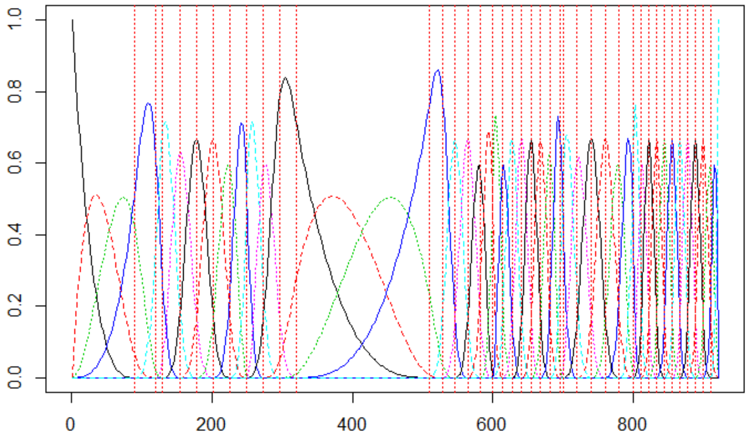
The primary way of achieving more flexibility in a spline is through increasing the number of breakpoints defined. However, there are often regions of a function with more complicated variation that needs to be captured. In this scenario more breakpoints are often loaded in these areas. Conversely, in regions where the function is only slightly non-linear or there are few local features, fewer breakpoints may be desired. Equal spacing of breakpoints or knots is used as a default in the majority of applications. This is common practice when there are lots of sample points per function and if the samples are observed at approximately equal points. However, in the case of irregularly spaced data it may be more sensible to manually design the placement of knots. Figure 3.3 demonstrates a set of fifty B-spline functions with equally spaced knots and another set of fifty B-spline basis functions irregularly spaced. In both cases, the more basis functions that are involved then the more complex the fitted functions can be. The degree to which the data are smoothed, rather than exactly reproduced or interpolated, is determined by the number K of basis functions (Ramsay and Silverman, 1997).

Choosing K

Selecting the number K of basis functions is an important question in the basis expansion problem. If a large enough K is chosen then the data may fit well, but any additive noise may also be fitted. If K is chosen to be too small then structure in the data important to the analysis may be removed. Generally the value of K is chosen so that the plotted functional object resembles the original data with some smoothing that eliminates the most obvious noise (Ramsay & Silverman, 2005; Garca-Portugus et al., 2014).



(a) Equally spaced basis functions



(b) Irregularly spaced basis functions

Figure 3.3: Equi-spaced vs Irregularly Spaced Basis Functions. The top plot illustrates 50 equally spaced B-spline basis functions, and the bottom plot gives 50 irregularly spaced B-spline basis functions for t=921

Figure 3.4 presents an example of smoothing by increasing the number of basis functions on the *mcycle* data from the MASS package of R. The data are a series of head acceleration measurements in a simulated motorcycle accident used to test crash helmets. A B-spline basis was used and as the number of basis functions increases a better fit of the data is observed but eventually it overfits to the noise. As the number of basis functions decreases more smoothing occurs.

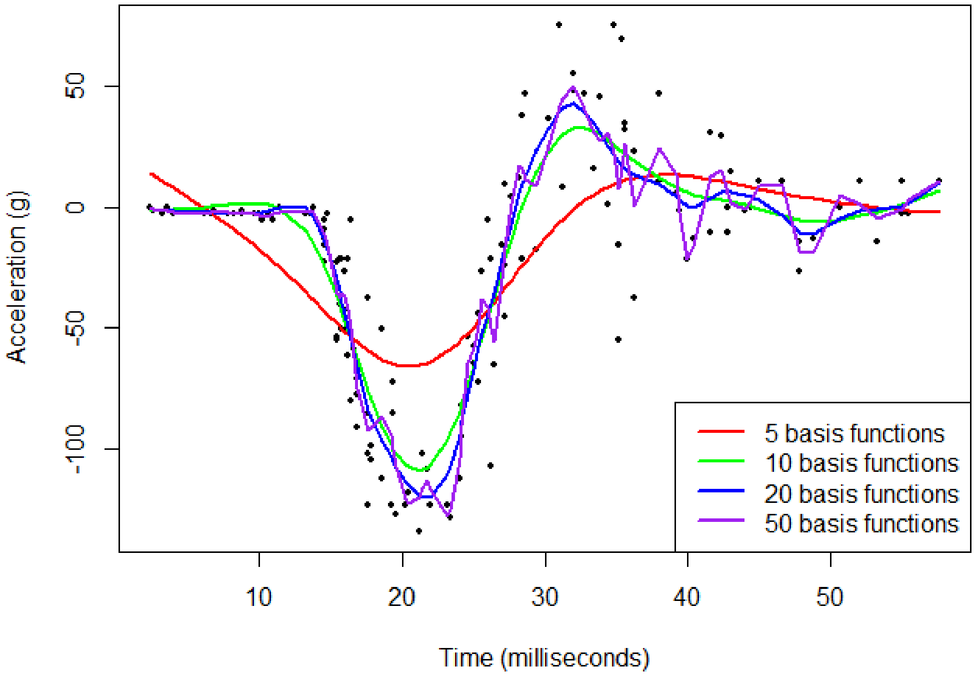


Figure 3.4: The effect of varying the number of basis elements on the smoothing of functional data. The data originates from a motorcycle accident dataset (Silverman, 1985). The data consists of measurements of head acceleration of a motorcycle rider as a function of time in the first moments after an impact.

Green and Silverman (1994) consider two approaches when it comes to selecting smoothing parameters. The first is to choose the smoothing parameter entirely subjectively. By varying the number of basis functions there are features of the data that arise at the different levels of smoothness, and the particular value for which the fit looks best may be chosen. The second approach uses an automatic method such as cross validation. In the functional data context where large datasets with many replications are commonplace, an automatic method seems the more sensible approach. However, if an automatic choice does not provide an adequate level of smoothing it can still serve as a starting point for subsequent subjective fine-tuning (Silverman, 1985).

Cross-Validation (CV) and Generalized Cross Validation are two widely used automatic selection methods for smoothing parameters. The main idea behind cross validation is to set aside a subset of the data, called the validation sample, and to fit and assess a model based on the remaining data against the validation set. By this method the smoothing parameter which gives the best fit is identified. The generalized cross-validation (GCV) method, as developed by Craven and Wahba (1978), can be described as:

$$GCV = \frac{1}{n}\sum_{i=1}^{n}\Big(\frac{y_i - \hat{f}_i}{1 - \frac{1}{n}tr(H)}\Big)^2$$

where, $\mathbf{H}$ is a hat matrix with $\mathbf{H} = \mathbf{X}(\mathbf{X^T X})^{\mathbf{-1}}\mathbf{X^T}$ and where $\mathbf{X}$ is a matrix of explanatory variables and containing the basis of smooths. The term $\hat{f}$ is the estimate of y_i from fitting all of the data.

The cross-validation (CV) approach is based on minimizing the mean squared error (MSE). However, GCV has the advantage over CV with less of a tendency to undersmooth the data.

3.1.3 Functional Descriptive Statistics

The theory from this point on only deals with functional data as smooth curves. Once the functional form has been achieved, the original discrete data can be discarded. The next step, and often the first step of any statistical analysis, is the exploratory analysis of the functional data for summarizing and visualizing the important characteristics. With a functional nature to data, the associated descriptive statistics must also be functional. Much work has been done to create functional equivalents of various statistical analyses and similarly, exploratory plots are also being developed as a tool in the initial analysis of functional data. For example, Sun and Genton (2011) have proposed a functional boxplot in order to visualize summary statistics of functional data as well as identifying outliers.

Mean and Variance Functions

One of the more basic of functional data analysis problems is the estimation of the mean function. Since functional data analysis sees each curve as a distinct datum, the mean function is calculated by averaging the functions point-wise across the replications. The mean function is defined as $\upsilon_x(t) = E(X(t)), \forall t \in T$, and the sample mean curve can be defined as

$$\bar{X}(t) = \frac{1}{N}(X_1(t) + \ldots + X_N(t)), \forall t \in T$$

where N is the number of curves or replications and $X_i(t)$ is the i^{th} function evaluated at time t. The estimation of the functional variance is also very similar to the classical variance. It is defined as $\sigma_x^2(t) = E[X(t) - E(X(t))^2], \forall t \in T$, and the sample variance curve is:

$$Var_x(t) = \frac{1}{N} \textstyle\sum_{i=1}^{N} [X_i(t) - \bar{X}(t)]^2$$

The corresponding correlation function can also be defined as:

$$Corr_x(t_1,t_2) = \frac{Cov_x(t_1,t_2)}{\sqrt{Var_x(t_1)Var_x(t_2)}}$$

Figure 3.5 illustrates the concept of the functional mean applied on the Canadian Weather dataset from Ramsay and Silverman (2005). The functional mean is calculated for four weather stations represented using a Fourier Basis expansion with K=50.

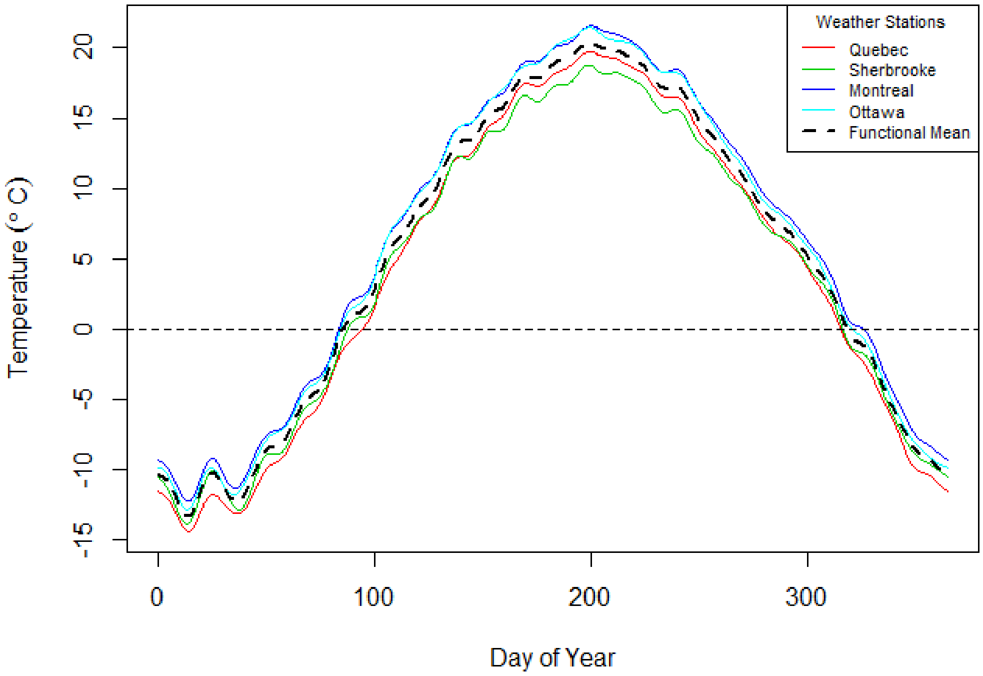


Figure 3.5: Plot of temperature from four selected weather stations with a functional mean curve fitted. Data from the Canadian weather dataset commonly featured in FDA literature (Ramsay and Silverman, 2005)

3.2 Popular Functional Data Methods

After discarding the original discrete data points, the smooth functions can be used in subsequent functional data analyses. Many of the methods used in classical statistics have counterparts available within a functional data context and some of these functional equivalents are introduced in the following sections.

3.2.1 Functional Principal Component Analysis (FPCA)

The most common statistical technique applied to functional data is the dimension reduction method of Functional Principal Component Analysis (FPCA). This technique is used to explore and distinguish any components of variations in the data. The motivation is similar to that of multivariate principal component analysis (PCA) that the directions of high variance will contain more information than the directions of low variance.

In the functional framework, each functional principal component (FPC) is defined by a principal component weight function, $\xi(t)$ and is defined over the functional data argument range of t (Ramsay and Silverman, 2002). The PC scores are given by the values z_i where

$$z_i = \int \xi(t) Y_i(t) dt \tag{3.1}$$

The objective in FPCA is to find the weight function $\xi_1(t)$ that maximises the variance of the PC scores z_i subject to the constraint

$$\int \xi(t)^2 dt = 1. \tag{3.2}$$

From this point the remaining principal components are defined similarly but with additional constraints. For example, the principal component function $\xi_2(t)$ maximises the variance of the principal component scores whilst satisfying the constraint (3.2) and the additional constraint,

$$\int \xi_2(t)\xi_1(t)dt = 0. \tag{3.3}$$

Furthermore, in general, for the jth component the additional constraints can be defined as follows:

$$\int \xi_j(t)\xi_1(t)dt = \int \xi_j(t)\xi_2(t)dt = \ldots = \int \xi_j(t)\xi_{j-1}(t)dt = 0. \tag{3.4}$$

This constraint ensures that all of the estimated functional principal components are mutually orthogonal.

3.2.2 Functional Hypothesis Testing

Ramsay and Graves (2009) state that common statistical tests tend to address questions in determining whether two or more groups of functions are statistically distinct, and whether statistically significant relationships among functional random variables exist. In a functional context, hypothesis tests may be used to investigate whether the shape of mean functions differ for different groups of functions. One such hypothesis test is the Functional ANOVA (FANOVA).

Permutation tests for functional hypothesis tests can be used to determine if there are any statistically significant differences between different groups. In this section, Functional F-tests and permutation t-tests are introduced. Functional F-tests are used to test if there are any statistically significant relationships between functional variables, and permutation t-tests can be used to test if there are any statistically significant differences between specific groups of functions.

Functional ANOVA

This functional version of the one-way ANOVA is achieved when a functional response variable is predicted using the conventional design matrix Z (Ramsay & Silverman, 2005). Suppose there are k independent groups with functional samples $y_{i1}(t), \ldots, y_{in_i}(t), i = 1, \ldots, k$. The FANOVA model takes the following form:

$$y_{ij}(t) = \mu(t) + \alpha_j(t) + \epsilon_{ij}(t), \quad j = 1, \ldots, n_i, \quad i = 1, \ldots, k \tag{3.5}$$

where $y_{ij}(t)$ are the observed functional data, $\mu(t)$ is a mean function, $\alpha_j(t)$ are the group effects and $\epsilon_{ij}(t)$ are independent $N(0, \sigma^2)$ errors. Alternatively Model 3.5 can be rewritten as:

$$y_{ij}(t) = \mu_i(t) + \epsilon_{ij}(t), \quad j = 1, \ldots, n_i; i = 1, \ldots, k \tag{3.6}$$

where the aim of the FANOVA is to test if the mean functions vary among k groups over a continuum t:

$$\begin{gathered} H_0 : \mu_1(t) = \cdots = \mu_k(t), \text{ for all } t \in T, \\ H_A : \mu_i(t) \neq \mu_j(t), \text{ for at least one } i \neq j \text{ and } t \in T. \end{gathered}$$

Permutation Tests

By using permutation tests for functional hypothesis tests, statistically significant differences between groups may be determined. Permutation tests date all the way back to R.A. Fisher's *The Design of Experiments* published in 1935 and involve the process of permuting observed data in order to generate a reference distribution of a test statistic for the purposes of testing a hypothesis. They operate under the belief that if the null hypothesis is true, then the arrangement of the observed data is purely due to chance, and thus all of the possible permutations are equally likely. Permutation tests all follow the same general order:

1. A test statistic is calculated using the observed data
2. All possible permutations are enumerated
3. A new test statistic is calculated for each permutation of the data
4. A hypothesis test p-value is then calculated as the proportion of the permutation distribution with values as extreme or more extreme than the observed test statistic.

Although these steps are straightforward, the challenge lies in selecting an appropriate test statistic and determining how to permute the data correctly. To determine if there are any statistically significant differences between groups permutation tests can be used for functional hypothesis testing. Functional F-tests can be used to test for the presence of any statistically significant relationship between functional variables. Permutation t-tests however can be used to test if there are any statistically significant differences between groups of functions.

Functional F-tests

Based on the functional ANOVA, permutation F-statistics can be used to assess if there are any significant differences between groups of curves. For a set of N curves represented by the smooth functions $g_i(t)$, Ramsay and Silverman (1997) define the functional equivalent of the univariate F-test statistic as:

$$F(t) = \frac{Var[\hat{g}(t)])}{\frac{1}{n}\sum(g_i(t) - g(t))^2} \tag{3.7}$$

where $i = 1, \ldots, N$ and $\hat{g}$ are the predicted values from a fitted FANOVA model. This equation (3.7) gives a function built from a series of point estimates at each t. However, to formally test the null hypothesis that there is no relationship between the functional variables a single test statistic is required, along with a p-value indicating the probability of observing a result as extreme, or more extreme, in the case that the null hypothesis is true. The maximum of the test statistic function, $F(t)$, is used as the test statistic and a distribution of the test statistic under the null hypothesis can be obtained by calculating the test-statistic several times, each time using random permutations of curves. The p-value corresponding to this test is the proportion of instances where the maximum value of the permutation F-statistic function is greater than the maximum of the observed F-statistic function. A pointwise curve can be plotted alongside the observed test statistic curve to provide an indication to the regions whereby the groups are less distinctive.

If N individuals are assumed to form K distinct groups then the null and alternative hypotheses can be defined as:

H_0: There is no difference between the K groups

H_A: There is some difference between at least two of the K groups

Functional t-tests

Similarly to the functional F-test, a permutation t-test can be used to assess if there are any statistically significant differences between groups of functions. Assuming there are two distinct groups of curves, $g_1(t)$ and $g_2(t)$, with N_1 curves in group 1 and N_2 curves in group 2, then a t-test statistic function can be defined as:

$$T(t) = \frac{\mid \bar{g}_1(t) - \bar{g}_2(t) \mid}{\sqrt{\frac{1}{n_1} Var[g_1(t)] \frac{1}{n_2} Var[g_2(t)]}}$$

and the null and alternative hypotheses tested are as follows:

H_0: There is no difference between the groups 1 mean and the group 2 mean

H_A: There exists some form of difference between group means in question

The maximum of the observed t-statistic function can be used as the test statistic and can be compared to a relevant null distribution which is calculated from a set of permutations. Similarly to the functional F-test, this test is based on the idea that under the null hypothesis, for any given t, the pairing of the value of the i^{th} curve in the k^{th} group, $g_{ik}(t)$, and the group number k are entirely random. It is important to note that an assumption of the permutation t-test is that all groups of curves must have the same variability.

A similar procedure to that outlined for the functional F-test can be used to estimate a distribution of the test statistic under the null hypothesis. Again, a p-value is computed by calculating the proportion of instances where the maximum value of the permutation t-statistic function is greater than the maximum observed t-statistic function.

3.3 Applications to the MIR Data

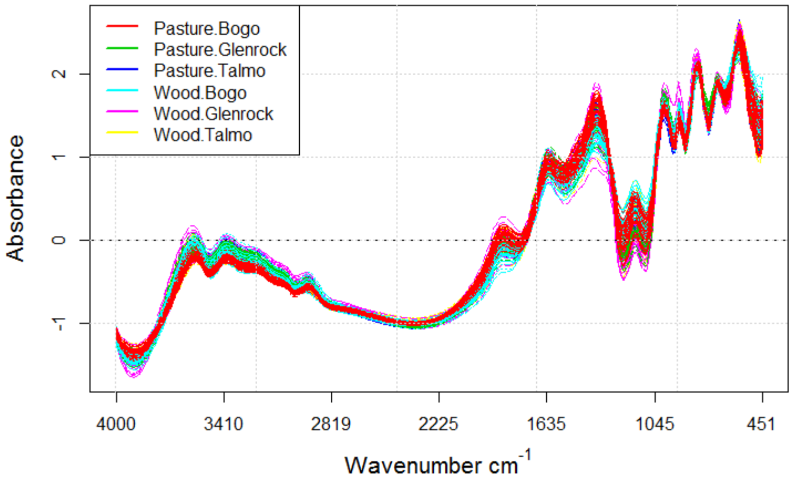
In the analysis of spectrometric data as functional data, each spectrum is a function that maps wavenumbers of the illuminating light to the corresponding absorbances (the responses) of the sample (Rossi and Villa, 2006). However, for functional data representation to be achieved, smoothing must first be applied to the discrete MIR soil spectra data.

3.3.1 Functional Exploration of the MIR Data

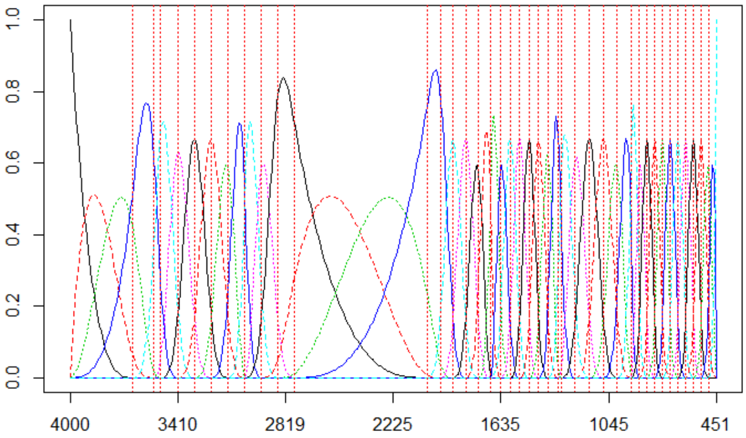
To better understand how to approach the problem of smoothing, the original exploratory plot of the spectra in Figure 2.6 of Chapter 2 was revisited in order to inform on the selection of the type of basis and for the placement of basis functions. Different modes of variation are considered in a functional paradigm including within-curve variation and curve-to-curve variation. The plot of the raw spectral signatures showed that the curves tend to follow each other quite tightly with not much variability between them. Across the entire range there is a great deal of fluctuation with a high number of peaks and valleys in the data for all replications. With these spectra there are no irregularities in terms of missing data or misalignment of curves. However, the spectra as a whole exhibit many areas of pronounced curvature interspersed with areas of relative inactivity. For instance, the majority of the curves in the latter regions of the mid-infrared range from 1700-450 cm^{-1} demonstrate a high degree of curvature with lots of local features. In contrast, the middle regions of the spectra from 3400-3200 cm^{-1} and 2800-2400 cm^{-1} appear more linear with a lack of any discernible features.

Given that the soil spectra data are non-recurrent and there is a high degree of local features, a B-spline basis approach to smoothing was adopted. Originally, GCV was used to choose the appropriate smoothing parameters, and these were computed using the *min.basis* function in R within the *fda.usc* package (CRAN, 2015). However, by this approach the spectral data appeared almost unchanged so a subjective approach to smoothing was adopted instead. It was decided to irregularly space basis functions due to the high degree of varying curvature with periods of inactivity across the spectra. It also makes sense to place more knots in areas known to contain high curvature and fewer knots in areas with less curvature. Discussions with soil scientists also informed the smoothing process. The quantity of smoothing and placement of breakpoints was kept the same for all curves to ensure fair comparisons were made. In total, forty-five irregularly spaced basis functions appeared to represent the spectra best. This was based on the plotted functional object resembling the original data with the applied smoothing eliminating the most obvious noise. Figure 3.6 shows the resulting smoothed spectra (top plot) and a plot of the associated basis functions (bottom plot). It is these functions (curves) in Figure 3.6 (a) which are used in all subsequent functional analyses. The discrete data from which these curves were estimated are no longer considered.

There are several aspects of Figure 3.6 (a) which are worth commenting on here. First of all, the most apparent feature is that there is a huge degree of overlap in the spectra with mixing across the entire range. However, looking at specific regions it is possible to see distinct separations. For example, approximately between 3600 and 2900cm^{-1} , the pasture samples of Bogo in red appear clustered lower in the absorbance range than the rest of the curves whilst the woodland samples of Bogo appear fairly central. All samples appear very tightly gathered in the middle range of the MIR spectra between



(a) The functional representation of the MIR spectra using $k = 45$ basis functions with irregularly spaced knots. These smoothed spectral curves are coloured by the Site*Land-use interaction.



(b) A plot of the $k = 45$ basis functions with irregularly spaced knots.

Figure 3.6: Functional Data Representation of the MIR soil spectra and the associated Basis Functions

2900-2000cm^{-1}. At approximately 1900cm^{-1} there is again clear clustering of the Bogo pasture samples separating from the rest and in the remainder of the curve there is a great deal of mixing and high degree of between-curve variability. It is also possible to view the curves as subsets of stand-alone farmland sites and land-uses. However, these plots also exhibited a high degree of mixing across the entire spectra and it was difficult to identify any specific regions where there were large differences between these groups of spectra.

It is well known that the boxplot is a graphical method for displaying the five common descriptive statistics namely the median, the first and third quartiles, and the maximum and minimum observations. Boxplots may also indicate which observations, if any, can be considered to be outliers. For functional data, Lopez-Pintado and Romo (2009) introduced the notion of band depth which allows for the ordering of a sample of curves from the centre outward. This allows quantiles, centrality and outliers to be defined in the functional context. More recently, Sun and Genton (2011) proposed a natural extension to the classical boxplot using the ranking of curves to compute functional boxplots. These functional boxplots are used to visualize summary statistics of functional data as well as having the ability to identify outliers. Figure 3.7 displays a functional boxplot allowing for an examination of variation between spectra.

Previously, the variability of curves across the mid-infrared range was examined and there was a high degree of variability exhibited with certain regions showing more extensive curvature. Here, the functional boxplot allows for the between-curve variation to be examined. In the classical boxplot, the

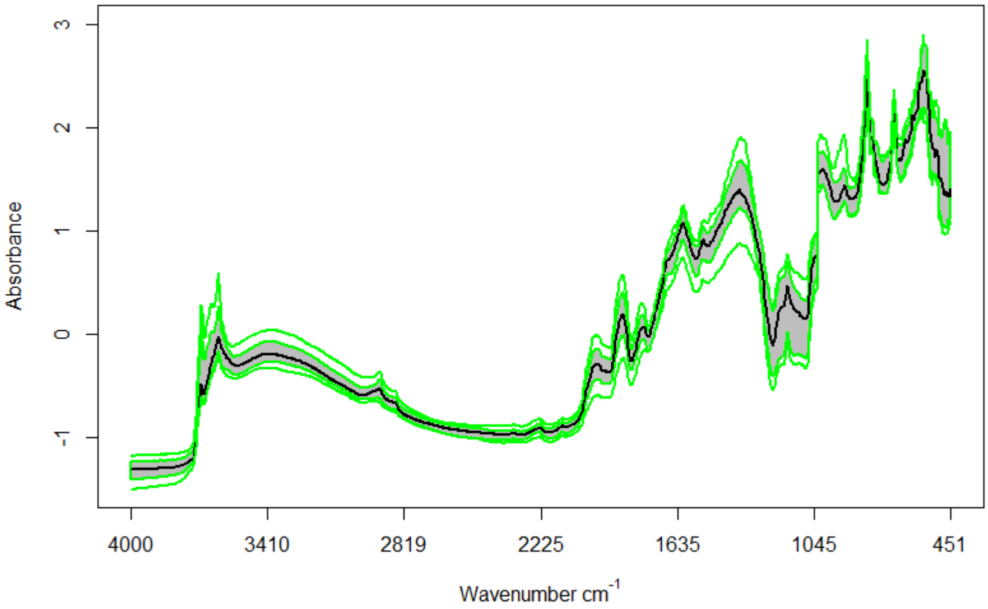


Figure 3.7: Functional Boxplot of the functional data representation of the MIR soil spectra.

box itself represents the central 50% of the data but in the functional setting this is not as easily defined due to the crossing of curves. Liu and Singh (1999) conceptualised the central region of functional data to which the grey portion of the functional boxplot corresponds. Examining the grey central region across the mid-infrared range, it can be seen that variability changes across the spectra. Between 2850cm^{-1} and 2200cm^{-1} the variability is low with the grey central region narrow. However, there are areas of high variability found across the entire range (e.g. 3500-3300cm^{-1}, 1600-1000cm^{-1}, 600-450cm^{-1}). The black curve in the grey 'box' indicates the median or most central curve. The median curve is a robust statistic to measure centrality and thus gives an indication of the general spectral signature. The outer green lines correspond to the whiskers of the boxplot and in this case no outlying curves have been identified.

3.3.2 Functional Hypothesis Testing

Another way to look at the data is to consult the functional means in Figure 3.8. The functional group means and the overall mean function (black) are used to informally assess if the curves of different groups are distinct from one another. Again, this plot shows a large amount of overlap in the groups with the lines of each spectra crossing frequently. However, this plot highlights that each group mean differs from the overall mean within some interval or multiple intervals across the spectrum. Areas to note include the 3600-3000cm^{-1}, 1950-1700cm^{-1}, 1400-1300cm^{-1} and 1200-1000 cm^{-1} intervals. In all these regions, the functional group means are a discernible deviation from the overall functional mean in black. These regions indicate sources of variation which could be responsible for differences between the groups of spectra. The functional means demonstrate that there is separability between the groups in terms of mean levels, although it is in the form of small deviances at specific intervals rather than across the entire spectrum.

Observing the functional means for each of the groups gives the reader an informal impression of how similar or different the current groups are. However, more formal techniques can be implemented to determine if the mean levels truly differ and if groups can be said to be statistically distinct. Thus, functional F-tests were carried out and permuted functional t-tests were used in order to test for group effects. An important assumption of permutation t-tests is that all groups of curves should have the same variability and this was found to hold by a visual inspection of the curves.

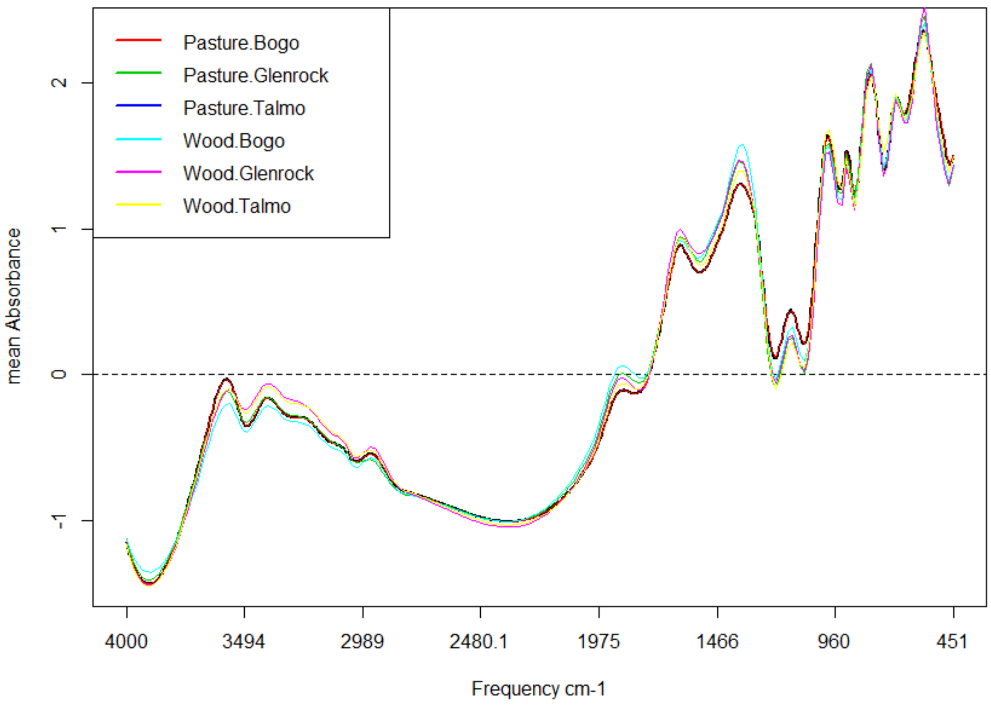


Figure 3.8: Overall mean function curve and the group mean functions for each site*land-use interaction

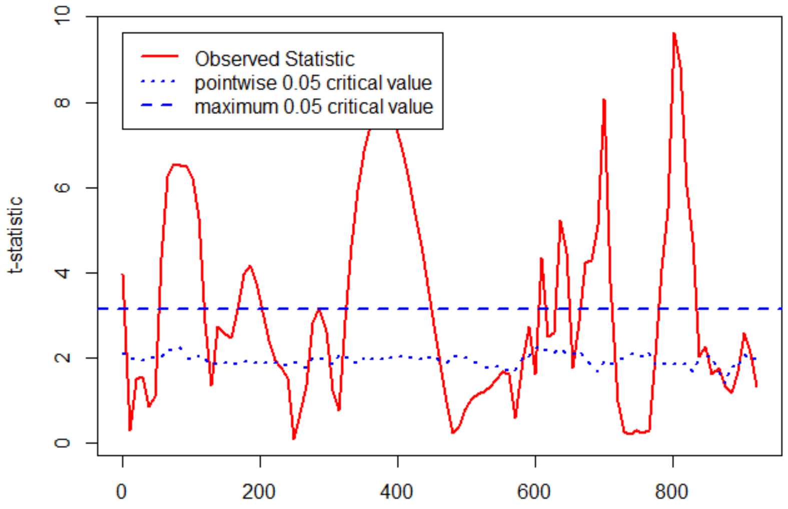
Functional F-tests

Functional F-tests were used in order to investigate if there exists any differences between any of the groups. Regardless if the samples were grouped by site, land-use or the land-use*site interaction, in all cases the p-values of the functional F-tests were highly significant (<0.001) and therefore implied that there are clear differences between at least some of the groupings in terms of their mean function. Given the straying of the group mean functions away from the overall mean function in Figure 3.8, this is unsurprising. In several areas it is clear to see that there is at least one group where the mean level is quite different.

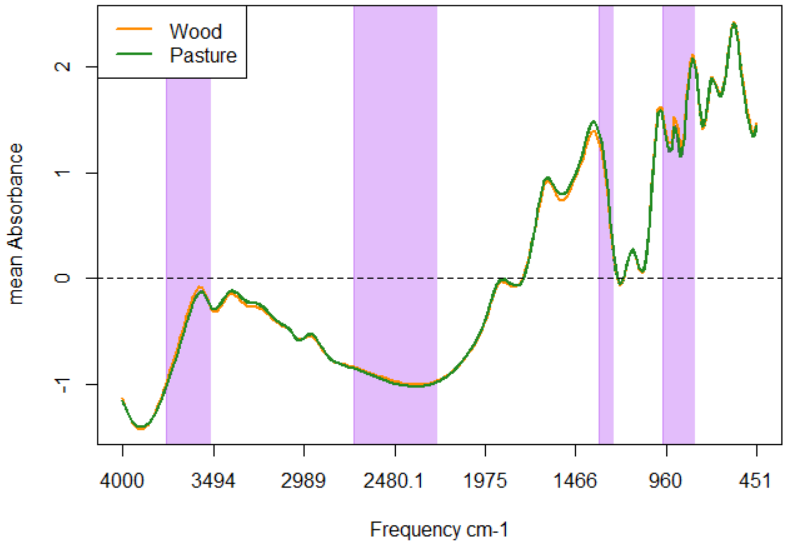
Functional t-tests

In order to find out more specifically where the differences lie, functional t-tests were carried out for each grouping variable. With each pair of groups a p-value was calculated corresponding to the test of the null hypothesis that there is no difference between the mean functions of the two groups in question, against the alternative hypothesis that there exists some form of difference. It was found that each pair of groups in all cases were found to be statistically distinct reporting highly significant p-values (<0.001). It should be noted that to account for the multiple comparisons between variables Bonferroni corrections were applied to the quantiles of the null distributions that the observed t-statistics were compared to.

Although there are only two land-use groups and the functional F-test was enough to confirm a statistically significant difference between their mean functions, a functional t-test was still applied as the resulting output allows the regions of difference to be displayed. Figure 3.9a gives a plot of the observed t-statistic over the entire spectrum range, where the red curve represents the value of the observed test statistic at all values in the range. The dotted blue line represents the pointwise critical values and the higher dashed blue line gives the maximum critical value, both at the 5% level. This test and indeed all the functional t-tests are based on a null distribution which has been constructed using 100 random permutations of the curve labels. Since the red line breaches the maximum critical value level at several intervals along the range it is clear that there is sufficient evidence to reject the null hypothesis, and conclude that there are statistically significant differences between land-use groups. The p-value corresponding to this test is less than 0.001 which further indicates a highly significant result.



(a) Observed functional *t*-statistic over the mid-infrared range. Statistically significant differences between curves of different land-use groups are observed where the observed statistic breaches the 0.05 critical level



(b) Plot of the functional mean curves for woodland and pasture spectra with statistically distinct regions of the spectra highlighted by purple overlays

Figure 3.9: Plot of the observed functional *t*-statistic testing for differences between land-use and the associated wavenumber regions where differences lie in the MIR range

Furthermore, by consulting the test statistic across the range of the spectra it is possible to observe the specific regions of interest whereby the mean functions of the land-use groups are more statistically significantly different. These regions are highlighted by the light purple overlays.

The F-tests found that some unspecified statistically significant differences existed between the grouping variables of site and for the interaction variables. In the further application of functional t-tests, it was found that all pairs of groups were statistically distinct from one another and in all cases highly significant p-values of less than 0.001 were reported. As with the land-use grouping variable, associated plots indicating the regions of specific differences in the mean functions of the groups were produced. Figures 3.10-3.12 display the regions of interest for these three pairwise site comparisons.

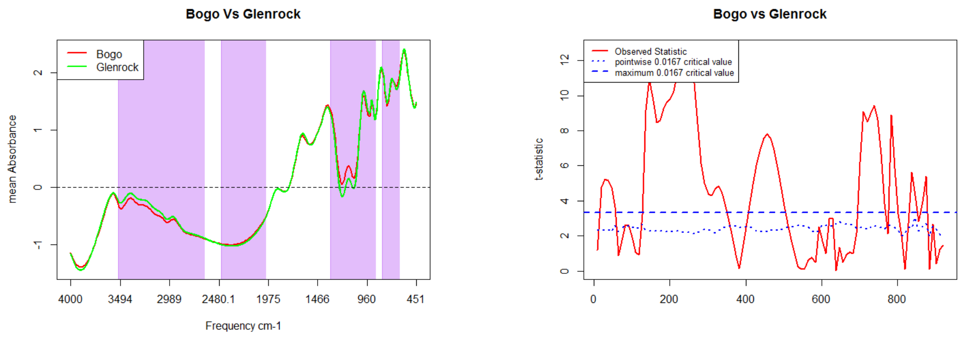


Figure 3.10: Plot of the functional means of Bogo and Glenrock spectra with statistically distinct regions highlighted by purple overlays (left). Plot of the observed functional t-statistic testing for differences between Bogo and Glenrock (right)

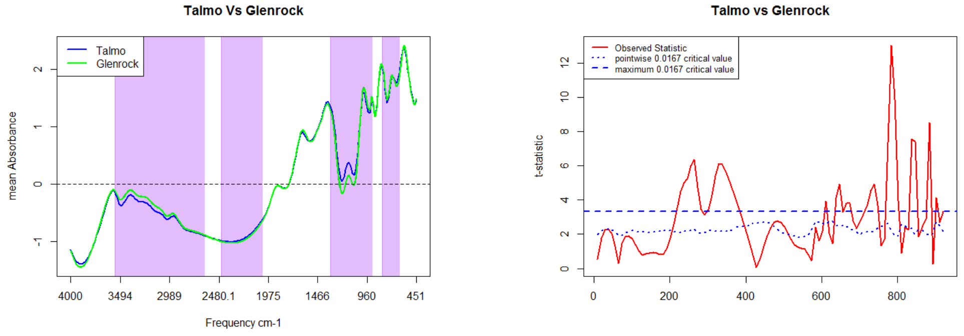


Figure 3.11: Plot of the functional means of Talmo and Glenrock spectra with statistically distinct regions highlighted by purple overlays (left). Plot of the observed functional t-statistic testing for differences between Talmo and Glenrock (right)

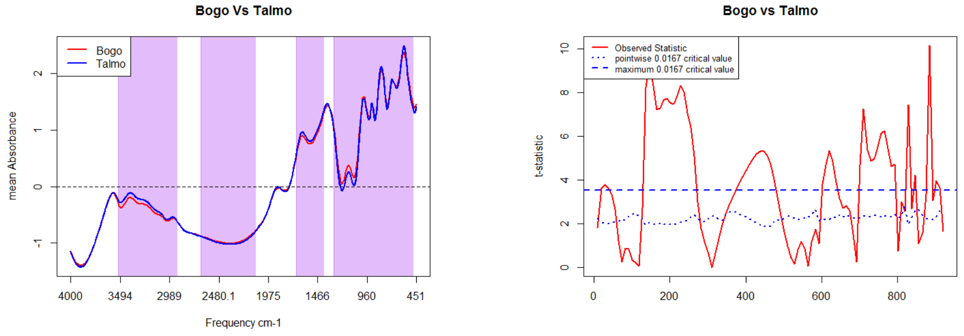


Figure 3.12: Plot of the functional means of Bogo and Talmo spectra with statistically distinct regions highlighted by purple overlays (left). Plot of the observed functional t-statistic testing for differences between Bogo and Talmo (right)

3.3.3 Functional Principal Component Analysis

In this section, Functional PCA is applied to the MIR soil sample data in order to study the variation in the 240 fitted smooth spectra. FPCA characterizes the modes of variability by decomposing functional observations into population level basis functions and subject-specific scores (Ramsay & Silverman, 2005). These basis functions have a clear interpretation that is analogous to that of PCA with the first basis function explaining the largest direction of variation, and each subsequent basis function describing less.

After applying FPCA to the smooth curves developed in Section 3.3.1, it was found that over 90% of the variation between individually fitted curves was expressed in the first five functional principal components. However, it was decided to retain FPCs in line with the selection criteria set out for the multivariate PCA. To recap, principal components were retained to satisfy a cumulative 90% of the total variance explained and trailing PCs were additionally retained if they satisfied Kaisers criterion and explained at least a further 0.01% of the variance.

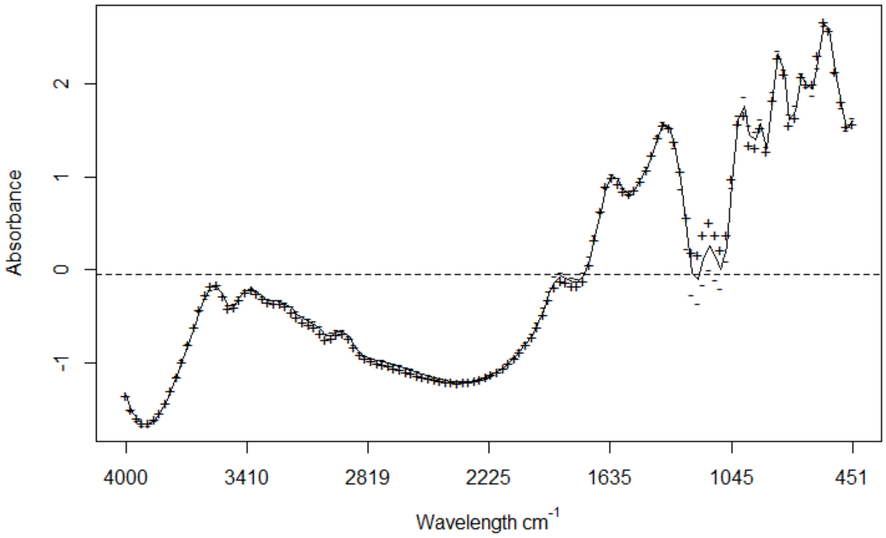
With these criteria satisfied, the essential modes of variation between the fitted curves were extracted by FPCA and a subsequent Varimax rotation was applied to the eigen-functions. The results of the functional principal component analysis are given in Table 3.2. The purpose of using the rotated principal components was to improve interpretability of the principal component plots.

Functional PC	Proportion variance explained	Cumulative variance explained
FPC 1	28.39%	28.39%
FPC 2	19.84%	48.22%
FPC 3	13.75%	61.98%
FPC 4	13.30%	75.28%
FPC 5	8.80%	84.08%
FPC 6	4.51%	88.59%
FPC 7	7.58%	96.17%

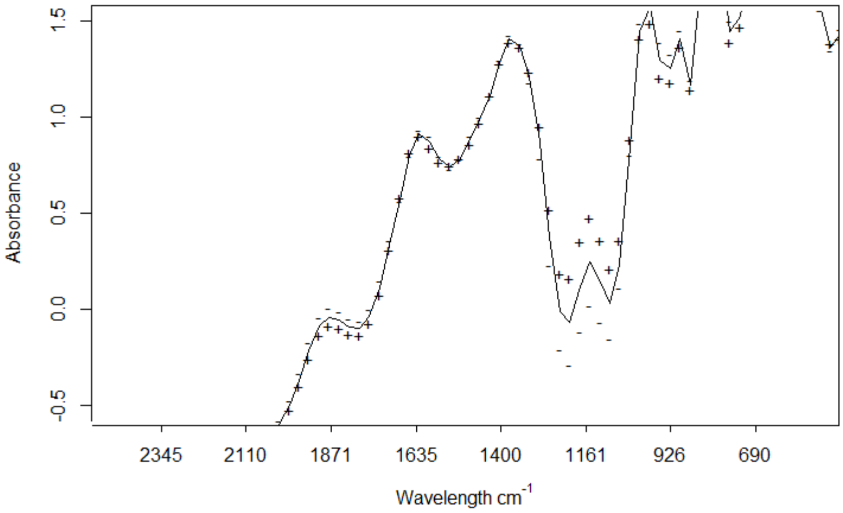
Table 3.2: Proportion of variability explained by FPCA

In total, more than 96% of the total variation between individual curves was explained by the first seven FPC curves chosen for retention. The first functional principal component explained 28.39% of the total variation in the spectra with the second and third FPCs explaining 19.84% and 13.75% respectively. FPCA has successfully managed to extract a limited number of FPCs that describe the patterns associated with the largest proportions of the variation in the individual fitted curves. However, the scores plots and FPC curves can reveal the more interesting aspects of the data and show where the differences between groups lie. These are indicated by perturbations of the mean function in the principal component function plots.

The first principal component function (Figure 3.13), accounting for 28.4% of the variation, reveals that the greatest between-curve variation occurs in the 1300-1000cm^{-1} wavenumber range. There also appears to be some widening in the confidence bands between 1880-1760cm^{-1}. Considering the scores plots, it seems that these sources of variation in the first principal component could be driven by the origin of the soil samples in terms of both site and



(a) The principal component function plot for FPC1 over the entire mid-infrared range

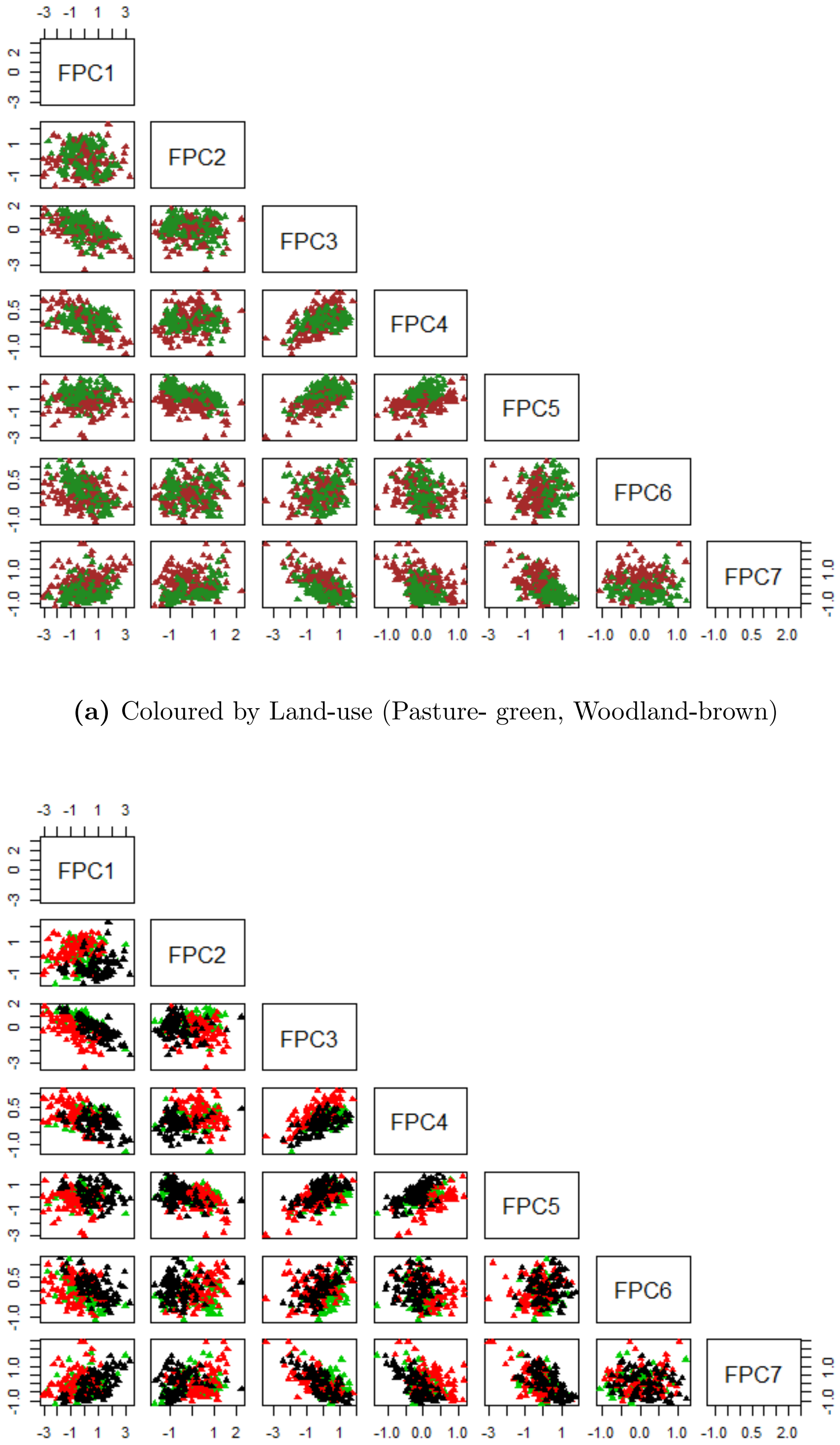


(b) A close-up of the 2460-530cm-1 range

Figure 3.13: The Principal Component Function plot for FPC1 represented as perturbations of the mean function

land-use. The scores plots (Figure 3.14) corresponding to the first FPC also exhibit some clustering. For example, there appears to be a general cluster of pasture scores with two clear split groups of woodland sites in the scores plot between the first and seventh FPC. This is perhaps indicative that there could be subgroups within woodland sites which differ in their chemical moieties and mineralogy in the 1300-1000cm^{-1} and 1880-1760 cm^{-1} regions. The scores plots coloured by site also show similar clustering in the scores plots corresponding to FPC1. Here, it seems that Bogo (black) and Talmo (red) differ most in the ranges identified above. However, this is subjective with the graphical presentation perhaps limiting here. In any case, it seems that there is a visible difference between sites in the 1300-1000cm^{-1} and 1880-1760cm^{-1} wavenumber ranges. The scores plots labelled by the site and land-use interaction have been omitted since with six interactions it is difficult to visually assess the presence of groupings. However, the results of the functional hypothesis testing did indicate the presence of distinct groupings between the interactions, as well as between land-use and sampling site classes.

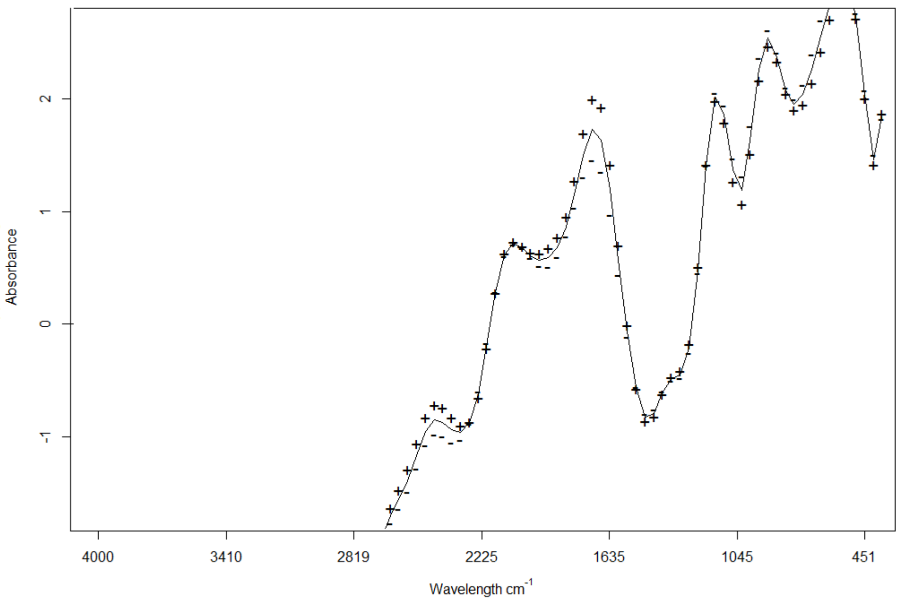
The second principal component function (Figure 3.15a), accounting for 19.8% of the variance, identifies important variation at approximately 1900cm^{-1} and more noticeably at 1450cm^{-1}. There are differences exhibited in the scores plots of the second functional principal component with respect to land-uses (Figure 3.14a), where there is clear clustering of woodland sites in the negative aspect of the scores plot between FPC2 and FPC7. Similar clustering of woodland scores can be viewed across all the scores plots of FPC2 indicative of soil components relating to wavenumbers around 1900cm^{-1} and 1450cm^{-1} influencing land-use differences. Interpretation of the scores plots coloured by site is much more difficult and no obvious clustering of scores is apparent.



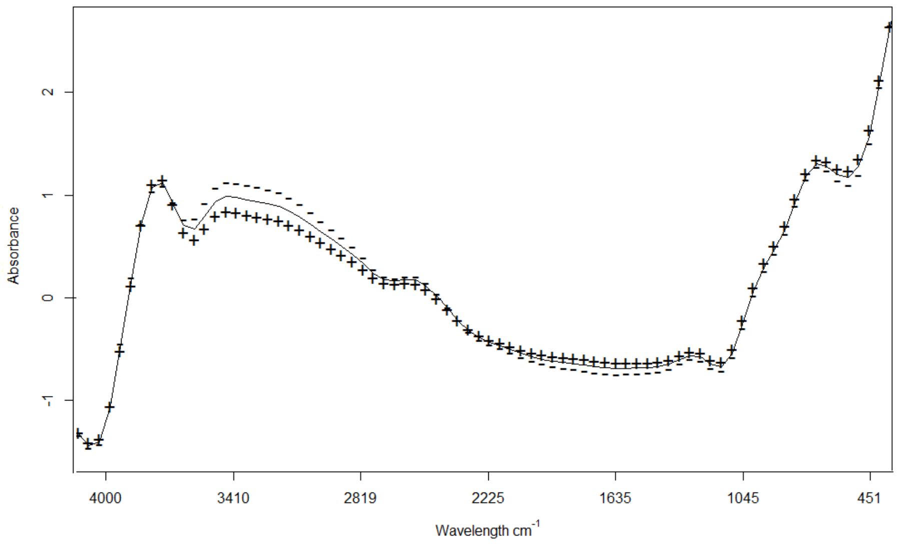
(a) Coloured by Land-use (Pasture- green, Woodland-brown)

(b) Coloured by Site (Bogo: Red, Talmo: Blue, Glenrock: Green)

Figure 3.14: Functional Scores Plots used to identify clusters of data grouped by site and land-use



(a) The Principal Component Function plot for FPC2 represented as perturbations of the mean function



(b) The Principal Component Function plot for FPC3 represented as perturbations of the mean function

Figure 3.15: FPC Function Plots for FPC2 and FPC3

The third FPC (Figure 3.15b) contributes to explaining the variation in the 3550-3150cm^{-1} region accounting for 13.8% of the variation. By examination of the scores plots it appears that this wavenumber range has less of an impact on land-use differentiation in comparison with both FPC1 and FPC2. However, there does appear to be a strong indication of differentiation between sites. In particular, the scores corresponding with Bogo appear to cluster in the positive aspect of all scores plots featuring FPC3. The remaining principal component functions (FPCs 4-7) account for the remaining variation as summarized in Table 3.2, but their function plots appear to demonstrate little effect on the mean function and thus are not presented here.

3.3.4 Summary

In Chapter 2, it was found that the multivariate PCA successfully reduced the MIR spectral dataset to just eight principal components explaining 97.5% of the total variability. Similarly, the functional PCA reduced the spectra to just seven FPC curves and over 96% of the total variability was preserved. However, whilst only vague class clustering was observed in the scores plots of the multivariate PCA, there was greater evidence of class separation in the scores of the functional PCA. Furthermore, of the seven FPC curves extracted, the first three exhibited perturbations from the mean function. These FPC curves were interpreted alongside their respective scores and it is thought that the regions identified are responsible for the class separations. Figure 3.16 illustrates the regions where these differences were found via a functional approach.

There are 3-4 main areas of interest identified by the FPCA. FPC1 identifies the 1880-1750 and 1300-1000cm^{-1} ranges accounting for 28.4% of the

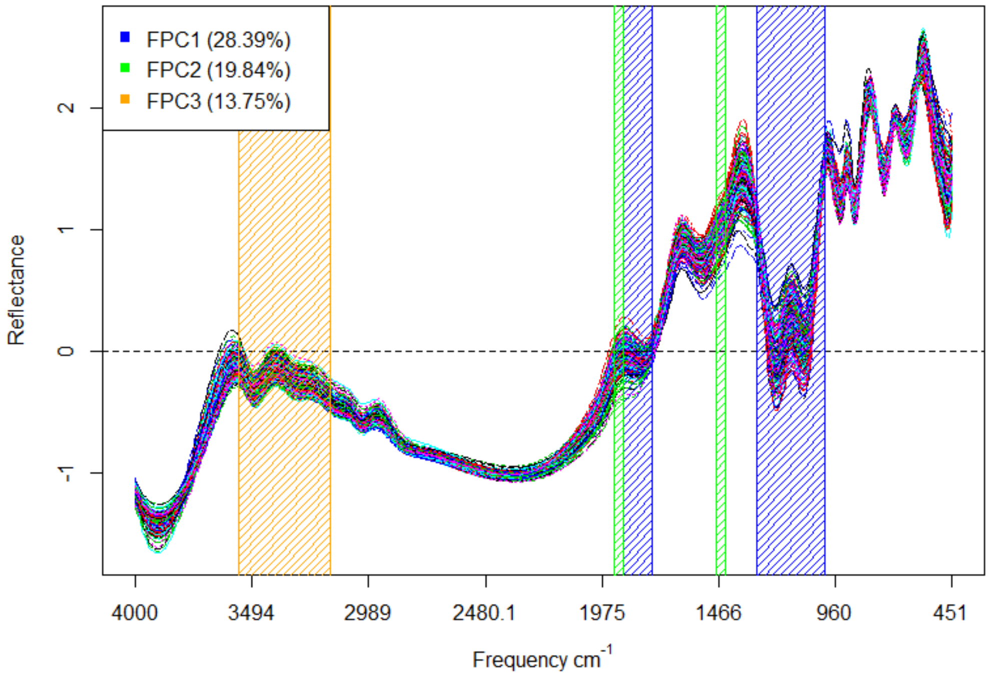


Figure 3.16: Wavenumber ranges identified as the modes of functional variation by the first three FPCs from an FPCA on the MIR data

total variation and consultation of the scores plots for this FPC seems to suggest that these regions contain data responsible for differences between soil sample spectra of different sites and land-uses. FPC2 accounts for 19.8% of the total variability in the spectra and identifies two very narrow bands of wavenumbers in which spectra appear to differ. These narrow bands at approximately 1900 and 1450cm^{-1} are identified as potentially containing information which may drive differences in soil samples of different land-uses. The third principal component accounts for 13.75% of the total variability and identifies with a larger region of wavenumbers at the beginning of the wavenumber range from 3550-3150cm^{-1}. From the examination of the scores plots, these wavenumbers are thought to explain differences between soil samples of both different land-uses and sample sites. Interestingly, these same general regions were informally identified by a visual inspection based on the group functional means.

In contrast with multivariate methods, the functional PCA had the ability to identify important wavenumbers spanning regions. Linear discriminant analysis is too precise for the problem as it identifies distinct bands, and the multivariate PCA requires a more subjective interpretation of principal component loadings. The absorption regions of soil constituents are known to span ranges of the mid-infrared spectra and so the functional approach is found to be more coherent. Thus, by considering entire intervals of the MIR range, aspects of the soil composition are less likely to be overlooked. The functional approach also takes into consideration the between-curve variability which is key to identifying differences between spectra of different classes (i.e. site and land-use).

The application of functional F-testing also allowed statistically significant differences between all grouping variables to be confirmed. In addition, functional t-tests further indicated statistically significant differences between the mean functions of all levels of each grouping variable. The functional t-tests additionally enable the interpretation of p-values along the range of the spectra, and thus regions of the spectra driving differences between groups could be established for all pairs. There were 3-4 main regions where groups are found to be statistically distinct. These regions were found to approximately similar to those identified by FPCA in Figure 3.16.

Chapter 4

Functional Regression Analysis

The focus up until now has been on exploring the variability of a functional data object derived from the mid-infrared soil spectra. However, this has not investigated how much of the variation can be explained by what is actually in the soil. Classical statistical methods such as linear regression and analysis of variance investigate the way in which variability in the observed data can be explained by predictor variables. Similarly, extensions of these classical methods in the form of functional linear regression can be employed to relate physical wet chemistry attributes of the soil to the mid-infrared soil spectra.

The last decade or so has seen methods of partial least squares regression (PLSR) being developed for the rapid and cost-effective prediction of soil properties based on near-infrared and mid-infrared spectra (Janik and Rawson, 2009). One of the central aims of this study is the presentation of functional regression as a possible alternative to PLSR for the prediction of soil properties. Mid-infrared technology in conjunction with PLSR has previously been demonstrated to provide a much cheaper method of characterizing the constituents in soil. The main interest of this chapter is to investigate whether functional regression can equally provide a cost-effective approach and successfully predict components in the soil. Additionally, any benefits of adopting a functional approach are explored.

In the first half of this chapter, a brief review of the different classes of functional regression model are introduced, and regression models based on derived inputs (including PLSR) are also discussed. The second half of this chapter applies these methods and compares the functional approach to the soil science industry standard of PLSR.

4.1 Functional Linear Regression

The models considered for analysis fall generally within the class of functional regression models (Ramsay and Silverman, 1997) whereby the regression model is said to be *functional* when at least one of the involved variables, either predictor or response, is functional (Febrero-Bande and Oviedo de la Fuente, 2012). There are three broad subcategories that can be defined:

1. Function-on-Function regression (both predictors and responses are functions)

2. Function-on-Scalar regression (responses are functions and predictors are scalars)

3. Scalar-on-Function regression (responses are scalars and predictors are functions)

A large literature exists for both function-on-scalar and the more widely used scalar-on-function regression. In contrast, the literature addressing function-on-function regression is quite sparse (Meyer, 2014). However, due to its inapplicability to the data and research aims, function-on-function regression theory is omitted for this thesis. It should be noted that the majority of the theory in this chapter follows that of Ramsay and Silverman (2005), and if the reader desires a fuller description of the theory they should consult Chapters 12-17 of their book *Functional Data Analysis.* Ramsay and Silverman (2002) also provide additional illustrations of a wide range of applications across all functional regression types in *Applied Functional Data Analysis: Methods and Case Studies.*

4.2 Scalar-on-Function Regression

In this section, the functional linear model with scalar response and functional predictor is considered. However, first the classical multivariate general model should be introduced. Interest lies in explaining the variability observed in quantity Y by a number of other quantities or covariates, $\mathbf{x} = (x_1, x_2, \ldots, x_p)$. In classical linear regression, models are often of the form

$$y_i = \sum_{j=0}^{p} x_{ij}\beta_j + \epsilon_i, \quad i = 1, \ldots, N \tag{4.1}$$

where β_j are regression parameters, x_{ij} are the covariates and ϵ_i are the error terms. The purpose of the error term is to allow for sources of extraneous variation such as measurement error to be accounted for in the model. Note, when $j = 0$, β_0 is a constant intercept term incorporated in the model. There are several approaches which have been developed to estimate the parameters and the simplest is based on minimizing the sum of residual squares and is commonly known as the ordinary least squares (OLS) method. The classical model in Equation 4.1 can be rewritten in vector form as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{4.2}$$

where $\mathbf{y}$ is typically a vector of observations, $\boldsymbol{\beta}$ is a vector of parameters, $\mathbf{X}$ is a matrix of regressors and $\boldsymbol{\epsilon}$ is an error vector with mean zero. For a scalar-on-function regression model, the vector of covariate observations x_{ij} =$(x_{i1}, \ldots, x_{ip})$ in Equation 4.1 is replaced by a function $x_i(t)$, and the model can be defined with an integral and intercept as

$$y_i = \beta_0 + \int x_i(t)\beta(t)dt + \epsilon_i, \quad i = 1, \ldots, n. \tag{4.3}$$

where the $\beta(t)$ are known as the coefficient functions and β_0 represents some constant parameter.

The coefficient functions determine the effect of $X_i(t)$ on Y_i. This means that in regions where $\beta(t) = 0$, any changes in $X_i(t)$ have no effect on the response and changes in $X_i(t)$ have a greater effect on the response in regions where $|\beta(t)|$ is large (James and Zhu, 2009). This is important since it means that the interpretation of the relationship between predictors and the response can be challenging, especially when presented with complicated shapes of $\beta(t)$. In addition, estimates of $\beta(t)$ can be accompanied by confidence intervals which are used for determining the significance of effects.

4.3 Function-on-Scalar Regression

Under the framework of a function-on-scalar regression, there are two general formats for predicting a functional response. There is the prediction of the functional response with values $x(t)$ by the standard design matrix which is known as functional analysis of variance, or there is functional multiple regression by a set of scalar covariates. As in scalar-on-function regression, the main change from a classical regression is that the regression coefficients now become coefficient functions with values $\beta_j(t)$.

The theory for functional analysis of variance (FANOVA) was covered in Chapter 3. More generally, a functional response can be predicted using a set of scalar variables rather than just 0s and 1s representing group membership. In some cases, a model may involve both types of predictors. However, the most common function-on-scalar model and the one applied in this thesis follows:

$$y_i(t) = \beta_0(t) + \sum x_{ij}\beta_j(t) + \epsilon_i(t), \quad i = 1, \ldots, n \tag{4.4}$$

where the $\beta_j(t)$ are the effects associated with scalar covariates and the $\epsilon_i(t)$ are residual functions (curves). As with scalar-on-function regression, the coefficients $\beta_j(t)$ are interpreted analogously to the coefficients of a standard multiple linear regression. That is, $\beta_j(t)$ can be interpreted as the expected change in the response for a single unit change in the predictor. However, differing from multiple linear regression, the coefficient functions are defined over a continuum t and are assumed smooth in t.The intercept coefficient, $\beta_0(t)$, represents the effects on the response not accounted for by the covariates.

4.4 Model Diagnostics and Assessment of Fit

Diagnostic procedures are just as important for functional regression as they are in traditional models. Once a regression model has been constructed, it is necessary to assess the adequacy of the model and confirm that the model fits the data well. The validity of the inference depends on the model assumptions holding, and classical regression diagnostics are based on residuals (Anscombe and Tukey, 1963). Despite this, residual diagnostics specific to functional linear regression is a largely unexplored area. However, common approaches to residual diagnostics are still applied and Ramsay & Silverman (2009) advise checking regression assumptions via residual versus fitted values plots and normal probability plots. The most commonly used residual plot is the scatterplot of residuals versus fitted values, and it serves the purpose of assessing whether the residuals depend in any way on the fitted response. If there are any deviations from a random scattering of points then this is indicative of a lack of fit.

In order to validate the performance of a regression function, it is appropriate to use a goodness-of-fit measure, meaning how well the model fits the data. The conventional coefficient of determination is still used within functional regression as a performance measure. It is denoted R^2 and defined as

$$R^2 = 1 - \frac{SSE}{SSY}.$$

SSE is the sum of squares of the residuals, and SSY is the total sum of squares where

$$SSE = \sum_t (y_t - \hat{y}_t)^2,$$

$$SSY = \sum_t (y_t - \bar{y}_t)^2,$$

and where

$$\bar{y} = \frac{1}{N}\sum_{t=1}^{N} y_t$$

is the mean of the observed data. In regression, R^2 can be said to explain how well the model approximates the data and the closer an R^2 approaches the value one the better the model fits the data. These goodness-of-fit statistics can be calculated using functional analogues of the sums of squares, and their inference holds in the functional regression case. In the functional case, there is a dependence of these quantities on the continuum t that makes the calculation of R^2 different. In the functional framework, the sums of squared errors function is calculated by

$$SSE(t) = \sum_i \left[y_i(t) - Z_i\beta(t)\right]^2,$$

where Z_i is a row vector of covariate values. Similarly, the sum of squares for regression is calculated by

$$SSY(t) = \sum_i \left[y_i(t) - \bar{X}(t)\right]^2$$

where $\bar{X}(t)$ is the overall mean function. The functional analogue of the R^2 can then be calculated as:

$$R^2 = 1 - \frac{SSE(t)}{SSY(t)}.$$

However, diagnostic checks should be considered cautiously and a model does not have to be perfect to be worthwhile. As George Box states, ‘all models are wrong, but some are useful’ (Box, 1979). Thus, the purpose of residual diagnostics is to check that the model is not grossly wrong (Faraway, 2006).

4.5 Regression via Derived Inputs

There can be situations where there are a large number of inputs which are also highly correlated. In these situations it can be of interest to produce a smaller number of linear combinations of the original variables and then use these new combinations as the inputs in a regression (Hastie et al., 2009). One such method of derived inputs is Principal Components Regression. Essentially, principal components regression uses the principal components retained from a principal components analysis in the construction of a regression model. Similarly, partial least squares regression (PLSR) defines linear combinations or components to be used in regression.

Partial Least Squares Regression

Similarly to PCA, Partial Least Squares Regression (PLSR) seeks to select components that maximise the covariance between the response and selected components (De Jong, 1993). For p predictor variables $x_1, x_2, \ldots, x_p$, each PLS component is a weighted linear combination of the p covariates. These new linear combinations are called eigenspectra or PLS loadings (Haaland and Thomas, 1988). In PCA, the extraction of components is independent of the response. However, PLS selects components by taking into account their relationships with the response and the goal is to identify as few of these components which describe most of the inherent variable information in the response. In partial least squares regression, the extraction of the components operates under the same constraints as PCA (Tu et al., 2011). These include the sum of squared weights equalling one, and the orthogonality of PLS components. Following extraction, the PLS components are ranked in order of importance according to the amount of variance in the response that they explain. For example, the first PLS component has the largest covariance with the response and the second PLS component has the second largest covariance with the response.

The key difference between PCA and PLSR is that PLSR seeks the directions that have high variance and high correlation with the response, whereas PCA only considers the variability in the predictors (Stone, 1990; Frank, 1993). For further reading, Hastie et al. (2009) set out an algorithm for the partial least squares regression procedure and this can be found in Section 3.5 of their book, *The Elements of Statistical Learning.*

4.5.1 Performance Measures

The performance of a PLSR model can be assessed by the coefficient of determination (R^2) between predicted and measured observations and the root mean square error of prediction (RMSEP). A good model should have a relatively high R^2 and a low RMSEP. In this study, the optimum number of PLS components were chosen according to the lowest RMSEP achieved via a leave-one-out cross-validated approach. RMSEP is calculated as:

$$RMSEP = \sqrt{\frac{\sum_{i=n}^{n}(y_i - y_{predicted})^2}{n}}$$

where n is the number of points used in the calculation, y_i is the observed value and $y_{predicted}$ are the predicted values.

4.6 Applied Functional Regression

In this section, methods of Functional Linear Regression are used to relate the functional data objects formed from the MIR spectra data to the wet chemistry variables in Table 4.1. These wet chemistry variables were recorded by standard wet chemistry laboratory methods and when it came to the regression analyses some of these variables were transformed before the fitting procedure. These transformations are also outlined in Table 4.1, and they were performed in order to make each of their distributions approximately symmetric. This table also presents the range of values for which these wet chemistry values were observed.

Wet Chemistry Variables	Units	Transformations	Min.	Max
pH	pH	-	4.64	6.35
Moisture	%	-	6.74	38.67
Carbon	mg/g	Square-Root	3.21	8.92
Nitrogen	mg/g	Square-Root	0.57	2.07
Total Dissolved Nitrogen	mg/g	-	4.53	76.90
Dissolved Organic Nitrogen	mg/g	-	1.00	44.10
Amino-N	mg/g	Log	0.06	2.62
NH4-N	mg/g	Log	-1.90	4.33
NO3.N	mg/g	Log	-4.6050	1.5830
Biomass N	mg/g	-	2.00	115.00
Total Dissolved Carbon	mg/g	-	38.49	276.40
Microbial Carbon	mg/g	-	3.98	984.20
Inorganic Phosphorus	mg/g	-	2.40	129.70
Organic Phosphorus	mg/g	-	21.20	347.30

Table 4.1: The wet chemistry variables, their transformations, units and approximate ranges

All functional regression fitting and statistical analyses including confidence intervals for regression coefficients and goodness-of-fit statistics were implemented using the fda package (Ramsay and Silverman, 2012). Partial least squares regression was performed using the pls package (Mevik, 2007).

4.6.1 Function-on-Scalar Regression

This section explores function-on-scalar regression whereby the MIR spectra is collectively predicted by all the scalar covariates in Table 4.1 together.

Following Equation 4.4, the function-on-scalar regression model was of the following form:

$$y_i(t) = \beta_0 + \beta_1(t)\text{pH} + \ldots \beta_{14}(\text{t})\text{Organic Phosphorus} + \epsilon(\text{t})$$

The main focus of any functional regression is the estimation of the regression coefficients, and the coefficient functions associated with the prediction of the MIR spectra are displayed in Figure 4.1 (top).

The intercept function is taken to represent the overall mean trend over all covariates and thus generally follows the shape of the original smooth spectra in Figure 4.1 (bottom). However, sizeable differences can be observed in approximate wavenumber regions 3700-3000cm^{-1} and 1300-1000cm^{-1}. These regions are where the coefficient functions appear to demonstrate some form of influence on the response. However, at this scale and plotted alongside the intercept function it is difficult to get a feel for the size of the effects of each coefficient function on the MIR spectra. Thus, Figure 4.2 presents the same coefficient functions, zoomed in and plotted without the intercept function. It is clear to see that each covariate appears to have a different effect on the response and the strength of these effects varies across the wavenumber range. However, with some coefficient functions maintaining values of zero across the entire range it appears that some of the wet chemistry variables have no influence on the MIR spectra. There are also some covariates which have greater effects across different regions of the wavenumber range.

For example, it appears that pH (red) has a minimal effect on the spectra in the middle region of the range but at other regions both negative and positive effects may be observed. It can be observed that Nitrogen (light blue)

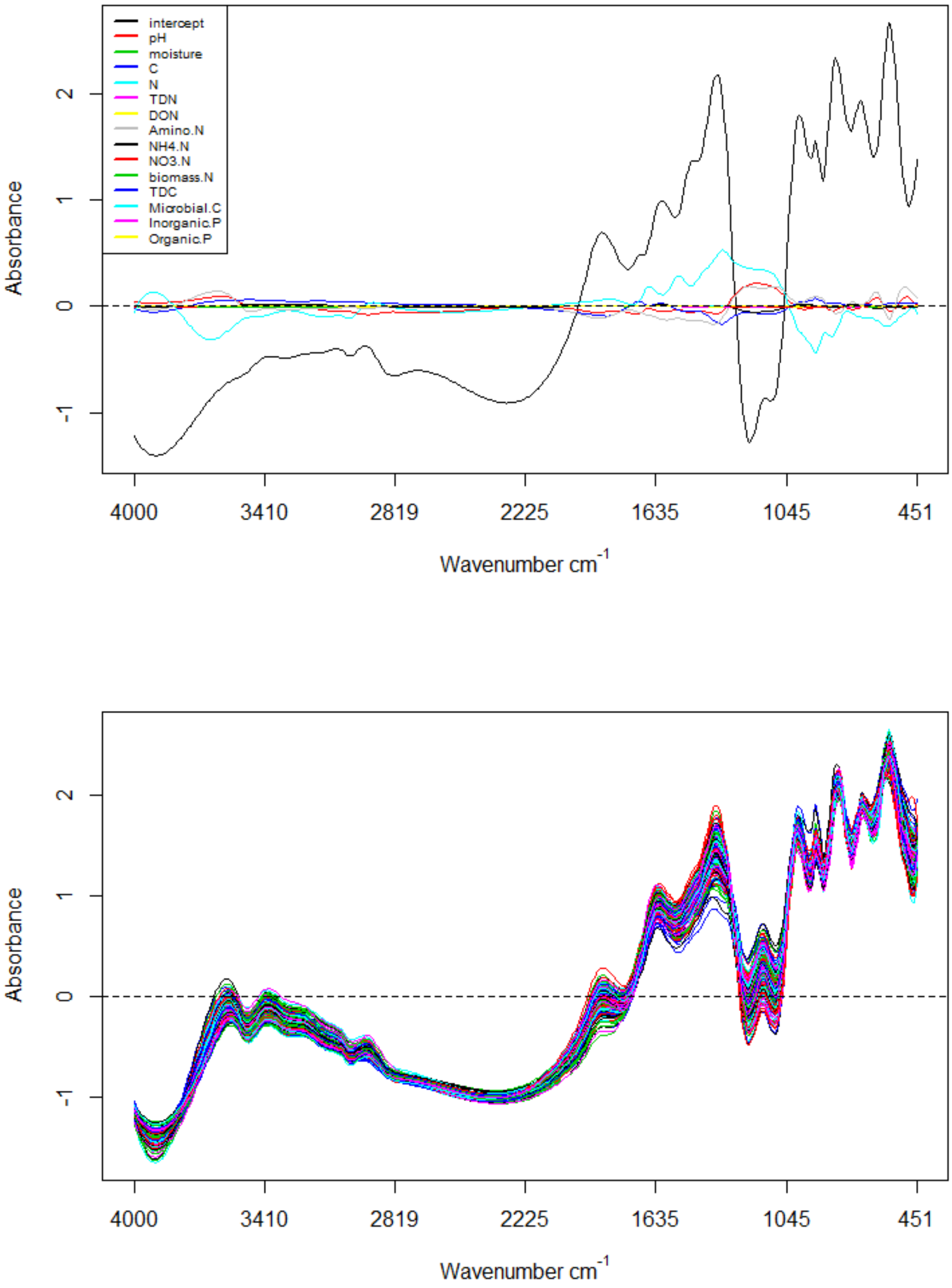


Figure 4.1: Function-on-scalar coefficient function estimates (top) and the smoothed spectra (bottom)

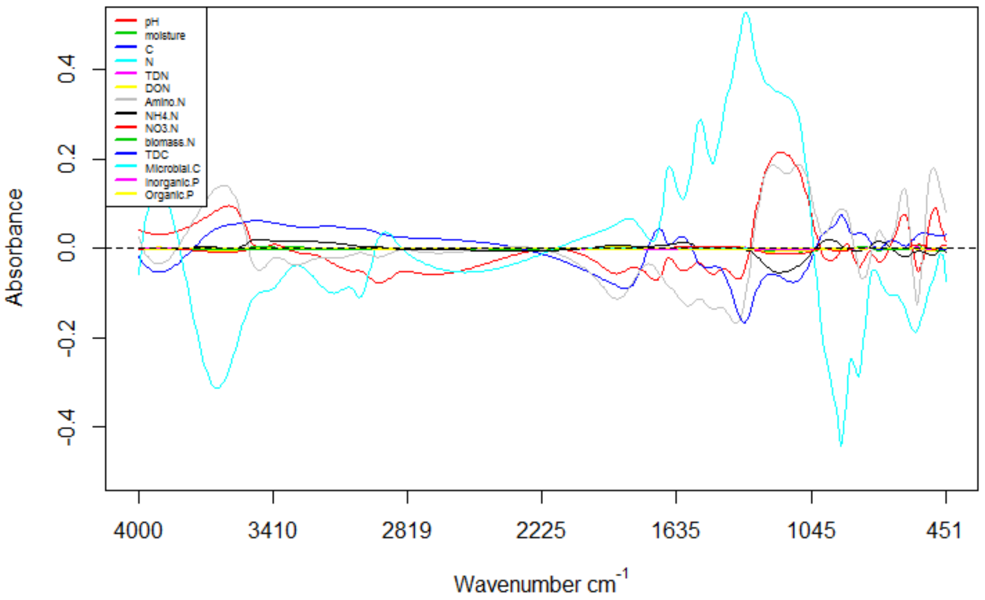


Figure 4.2: Function-on-scalar coefficient function estimates plotted without the intercept function

has the greatest influence on the response than any other covariate with its maximum value of approximately 0.45 at a wavenumber around 1000cm^{-1}. Although other covariates such as pH, Carbon, Amino-N and NH4-N have some form of influence in the surrounding regions, it appears that Nitrogen has the widest range of influence with greatest effect on the response from approximately 1550-900cm^{-1}. Nitrogen also exhibits effects of a large magnitude at around 3600-3500cm^{-1}. Most of the covariate effects appear to have greatest magnitude at either end of the spectra, and this coincides with the regions whereby organic and mineral signals may be found. The shapes of the $\beta(t)$ in Figure 4.2 are quite complicated and thus interpretation of the predictor-response relationship is difficult. If any of the $\beta(t)$ had been constant for any given non-zero region then the effect of the associated covariate on the response would be considered constant within that wavenumber region. Similarly, if the coefficient function had been exactly linear for any given range then the change in the effect of the covariate would be constant over that region. The coefficient function plots allow the magnitude of each covariate effect to be investigated across the mid-infrared range. However, confidence intervals of the coefficient functions are able to confirm the significance of these effects.

Upon study of the covariate effects with associated confidence intervals it was found that Total Dissolved Nitrogen (TDN), Moisture, Dissolved Organic Nitrogen (DON), Biomass-N, Total Dissolved Carbon (TDC), Microbial Carbon, Inorganic Phosphorus and Organic Phosphorus have had no influence on the functional MIR spectra. This is demonstrated by their coefficient function estimate either not diverging at all from the zero line or with zero being contained in their confidence interval in the minimal instances that it does diverge. For all other covariates, zero was not contained in their respective confidence intervals for at least some interval along the range. This allowed for conclusions to be made about where along the wavenumber range a covariate played a strong role in prediction. The most influencing covariates were pH, Carbon (Square-Root), Nitrogen (Square-Root) and Amino-N. Most of their influences on the response can be indicated to come from either end of the spectra with the greatest influence coming from the 1500-500cm^{-1} region. Figure 4.3 gives the coefficient function estimate with confidence intervals associated with Nitrogen. Amongst all influential variables, the confidence intervals were all quite variable with respect to their width but Nitrogen in particular exhibited the widest intervals. Again, Nitrogen exhibits the greatest influence on the response in the 1500-500cm^{-1} region with additional significant influence around 3600cm^{-1}. Nitrogen also

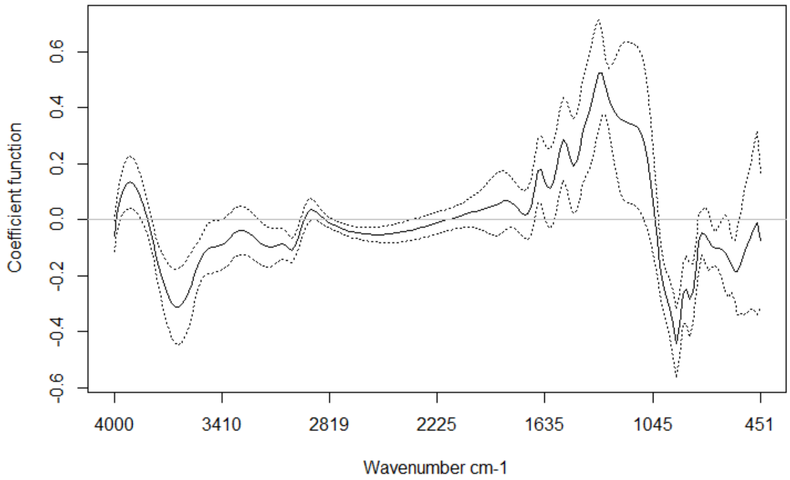
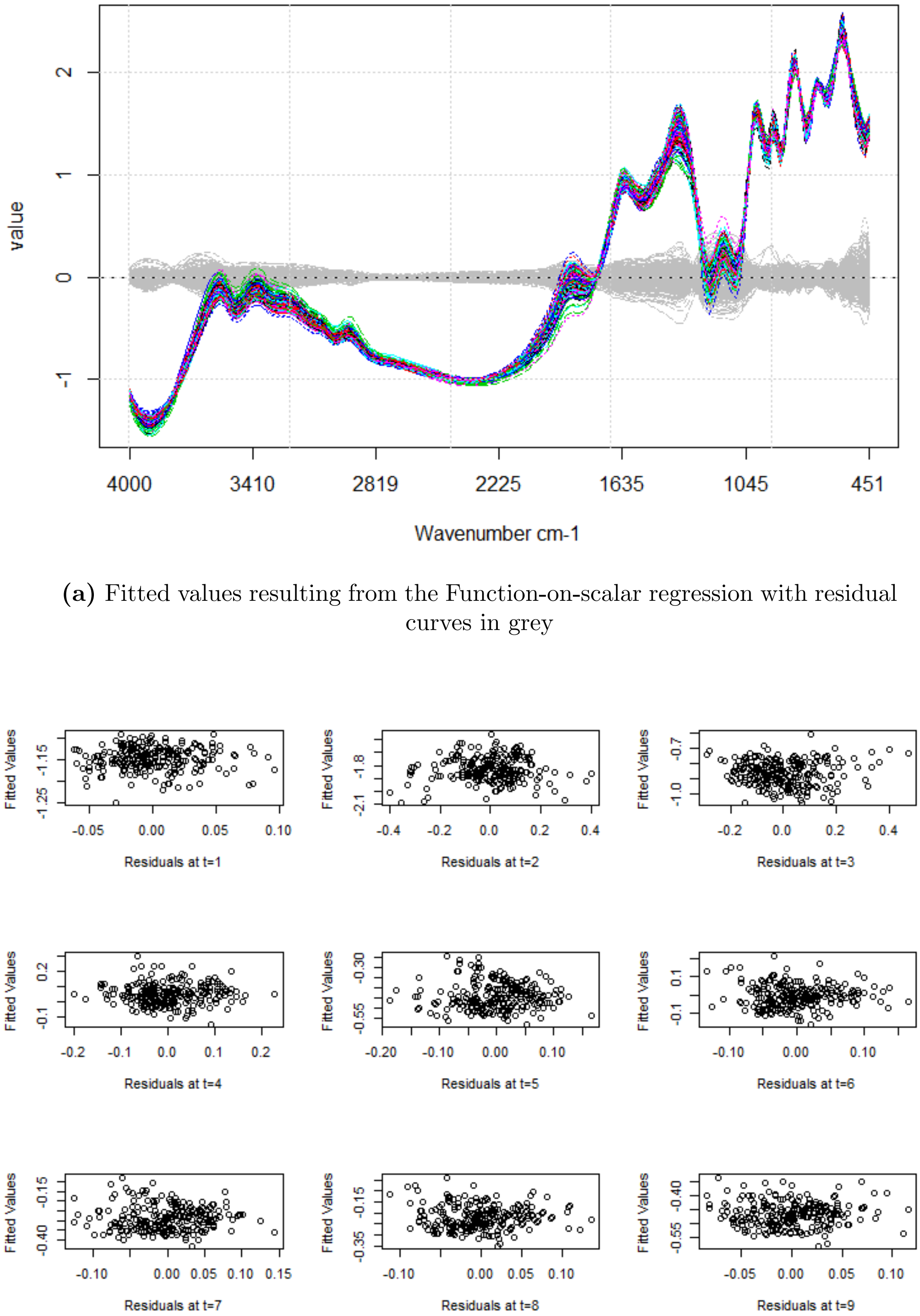


Figure 4.3: Function-on-scalar coefficient function estimate with 95% confidence interval for Nitrogen

provides a perfect example of the kind of complicated shapes which can be exhibited in a coefficient function plot.

Suitable diagnostic checks on the residuals were performed as recommended by Ramsay & Silverman (2002) to make sure there were no residual patterns remaining. Figure 4.4a gives the fitted values from the functional regression with the residual curves (grey) over the mid-infrared range. The residual checks did not reveal anything unusual here but it is worth noting that the residuals appear to vary across the mid-infrared range with the largest values noted around 1500-1000cm^{-1}. The residual functions are evaluated at a fine grid of points with residuals plotted against the predictor value for each t. Figure 4.4b gives the first nine of such plots. These residual checks did not reveal anything unusual and no curvature or lack of model fit was detected.

Whilst nothing irregular was detected there may still be potential to simplify the model. However, it is difficult to simplify the current model as there



(a) Fitted values resulting from the Function-on-scalar regression with residual curves in grey

(b) Fine Grid of Residuals from the function-on-scalar regression analysis for argument values t=1, . . . , 9.

Figure 4.4: Functional Residual Diagnostic Plots

is a lack of variable selection methods within function-on-scalar regression (Chen and Ogden, 2016). However, since there were a considerable number of covariates which had no effect on the response, as indicated by the coefficient function plots and associated confidence intervals, an alternative model was fitted without these particular covariates. The new model only included an intercept term along with the scalar covariates; pH, Carbon (Square-root), Nitrogen (Square-root), Amino-N (log), NH4-N (log) and NO3-N (log). The results of this regression were near identical to the full model and all covariates appeared to have the same effects as found by the original model.

Function-on-Scalar Regression using Principal Components

Complicated by the dimensionality of the response and the correlation structure of the residuals, variable selection is difficult in the functional setting (Chen and Ogden, 2016). Few approaches considering variable selection in the context of functional regression have been proposed in current literature. In an attempt at modelling the response using only the most essential predictors, Principal Component Regression was investigated.

By the same retention criteria of Chapter 2, a PCA on the fourteen wet chemistry variables successfully reduced the data to just four principal components which cumulatively accounted for 99.3% of the variability in the original data. These principal components were retained and had a Varimax rotation applied to improve their interpretability. The loadings plots for these principal components are given in Figure 4.5. These loadings plots indicate which variables are dominant within each principal component. For example, 76.4% of the total variability in the data is explained by the first principal component (PC1), and this is dominated by the 12th variable which

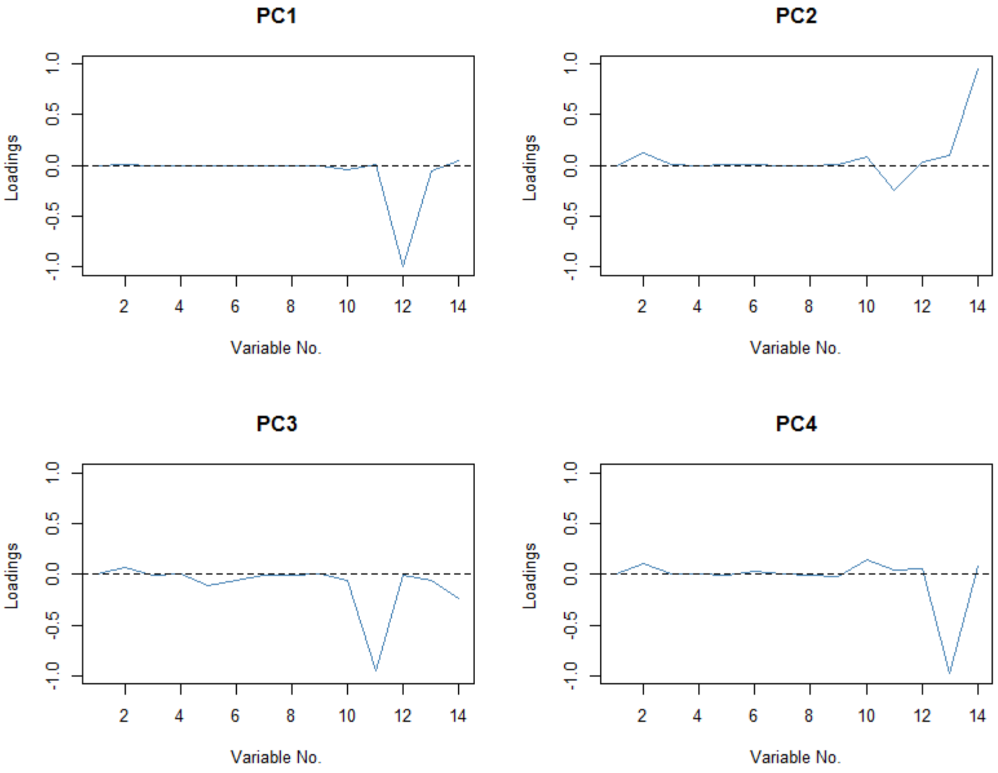


Figure 4.5: Loadings for the principal components retained from a PCA on the wet chemistry dataset

corresponds to Microbial Carbon (Table 1.2). Similarly, Organic Phosphorus, Total Dissolved Carbon and Inorganic Phosphorus dominate the remaining principal components which respectively account for 16.3%, 5.4% and 0.01% of the total variability. Subsequently, these principal components were extracted and used in a function-on-scalar regression to predict the functional spectral response, $y_i(t)$:

$$y_i(t) = \beta_0 + \beta_1(t)\text{PC1} + \ldots \beta_4(\text{t})\text{PC4} + \epsilon(\text{t})$$

Coefficient function plots were similarly produced and these are displayed in Figure 4.6. By using the principal components, it is clear to observe that all principal component coefficient functions have some form of influence on the spectral response with values falling above and below the zero line and the confidence intervals confirming the presence of a covariate effect. However, consulting the values of the y-axis, the magnitude of these covariate effects is very small. The same diagnostic checks were performed as before and the assumptions appeared reasonable. It should be noted that the principal components analysis has not picked up Nitrogen as an important predictor, where in the original function-on-scalar regression Nitrogen was found to be the most influential. However, one reason for this could be that PCA only considers the variability in the explanatory variables before they are fed into the regression.

Alternatively, the loadings from the PCA in Figure 4.5 could be used to identify the wet chemistry variables which contain most of the variability in the data, and these variables could be used in a functional regression. The loadings had indicated that these variables were Microbial Carbon, Total Dissolved Carbon, Inorganic Phosphorus and Organic Phosphorus. A function-on-scalar regression was explored for the prediction of the spectra using just these variables. However, the coefficient function plots achieved very similar shapes carrying the same interpretations.

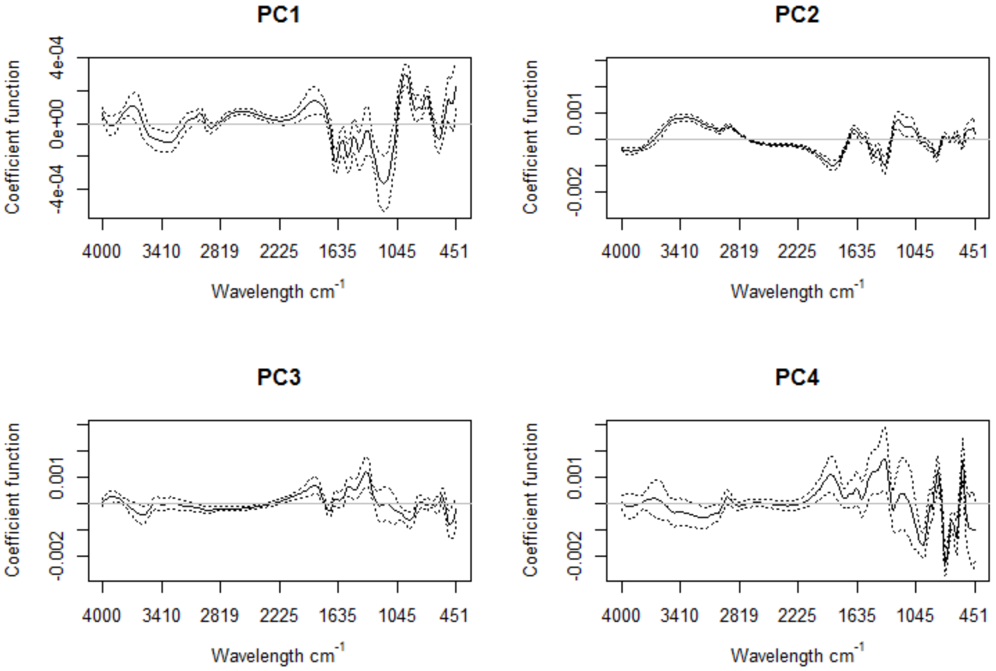


Figure 4.6: Coefficient function estimates with associated 95% confidence intervals from a function-on-scalar regression on the retained principal components from a PCA on the wet chemistry dataset

Function-on-Scalar Regression Summary

In summary, function-on-scalar regression has been successfully applied to these type of data and it has been shown that a model can be established between the functional MIR response and scalar wet chemistry covariates. Additionally, it has been shown that covariates have different effects on a response across the wavenumber range both in terms of magnitude and polarity. However, this section has been primarily exploratory to introduce functional regression and it's associated output. There is limited use in predicting an MIR curve from a series of wet chemistry variables. The primary aims of the study have much more interest in the scalar-on-function regression whereby the reverse is considered. In this form, the components in the soil may be predicted by the MIR spectra. In this way, soil samples can be analysed via

MIR spectra to identify soils which contain valuable chemical compositions. For example, a soil with greater carbon content may be of more value with respect for agricultural needs.

4.6.2 Scalar-on-Function Regression

The problem of predicting continuous scalar outcomes from functional predictors has received high levels of interest in recent years, driven partly by the collection of increasingly larger complex datasets and an increase in computational power (Goldsmith and Scheipl, 2014). Despite the increased interest and development in this area of statistics, functional regression is not widely applied in soil science for the prediction of physical soil properties. This section describes a scalar-on-function regression approach for the prediction of the fourteen wet chemistry variables introduced in the previous sections. The value of this type of functional regression within soil science is the ability to use the cheap method of MIR spectrscopy to estimate the chemical composition of a soil; avoiding rigorous, time consuming and expensive wet chemistry techniques.

As with classical regression analyses, the fit of scalar-on-function models can be summarized in terms of R^2 statistics (Ramsay & Silverman, 2005). For each scalar-on-function regression, the coefficients of determination (R^2) were obtained and these are presented in Table 4.2. The largest coefficients of determination are reported for the models associated with Carbon, Nitrogen and Organic Phosphorus with R^2 values of 0.625, 0.611 and 0.550 respectively. These values indicate that their functional linear relationships with the spectra are moderately strong. The relationships which appear poorest are with the models predicting Amino-N and NH4-N compounds, giving R^2

values of just 0.106 and 0.065.

Wet Chemistry Variable	R^2
pH	0.381
Moisture	0.258
Carbon (C)	0.625
Nitrogen (N)	0.611
Total Dissolved Nitrogen	0.153
Dissolved Organic Nitrogen (DON)	0.150
Amino-N	0.106
NH4-N	0.065
NO3-N	0.122
Biomass-N	0.275
Total Dissolved Carbon (TDC)	0.117
Microbial Carbon	0.246
Inorganic Phosphorus	0.335
Organic Phosphorus	0.550

Table 4.2: Soil chemical, physical and mineralogical properties, with indicative coefficients of determination (R^2) or capability predicted using MIR

Fitted value plots were produced following each regression and a couple examples are explored here. The fitted value plot for the functional linear model predicting Carbon is given in Figure 4.7. There appears to be considerable agreement between the two methods with a moderately strong positive linear correlation. With an R^2 of 0.625, the functional regression model predicting soil carbon exhibited the strongest relationship to the MIR spectra. However, the majority of the other soil quantities exhibited only weak linear correlations with the spectra indicating poor predictive performance.

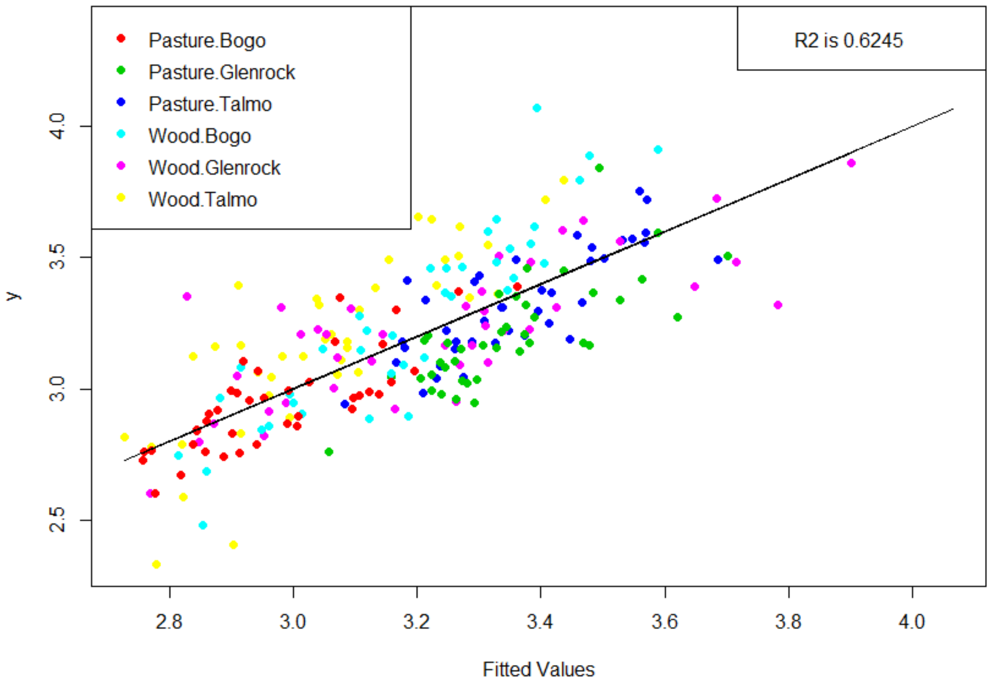


Figure 4.7: Fitted Value Plot associated with the scalar-on-function regression predicting Carbon, coloured by site and land-use ID

Additionally, the fitted values plots indicate group memberships (i.e. site and land-use) and this allowed for the investigation of possible groupings and the informal assessment of whether certain classes of the spectra should be modelled separately. In the Carbon example of Figure 4.7, the fitted values for the pasture site of Bogo (red) tend to group together having generally lower Carbon values. Values for the pasture sites of Glenrock (green) and Talmo (blue) have congregated around the middle portion of the fitted line. Meanwhile all woodland sites in teal, pink and yellow appear to straddle the fitted line across the entire range. This suggests that there is reason to form functional linear regression models separately according to the type of land-use the soil samples originate from. By splitting the data and performing regression analysis on each land-use grouping separately any improvements in prediction performance and model fit can be assessed.

The prediction of moisture is another example where differences between sampling locations is highlighted. In Figure 4.8, the fitted values associated with the pasture site of Talmo are elevated above the fitted line. For whatever reason, the soil samples taken from this sampling location are far more saturated than the rest of the samples. Thus there is scope to remove this grouping entirely, or alternatively model based on land-use splits of the data.

Studying the residual plots for the moisture regression model also revealed poor fit (Figure 4.9a). The residuals versus fitted values plot exhibits a clustering of points and no random scatter showing that there is some structure in the data which has not been accounted for. Upon exclusion of the Talmo pasture data, the residuals behave satisfactorily (Figure 4.9b). The fitted values plot for this subset, Figure 4.8 (bottom), also indicates a weak positive linear relationship between the lab-measured and the predicted values. Similar residual analysis was conducted for all regression models, however no irregularities were discovered.

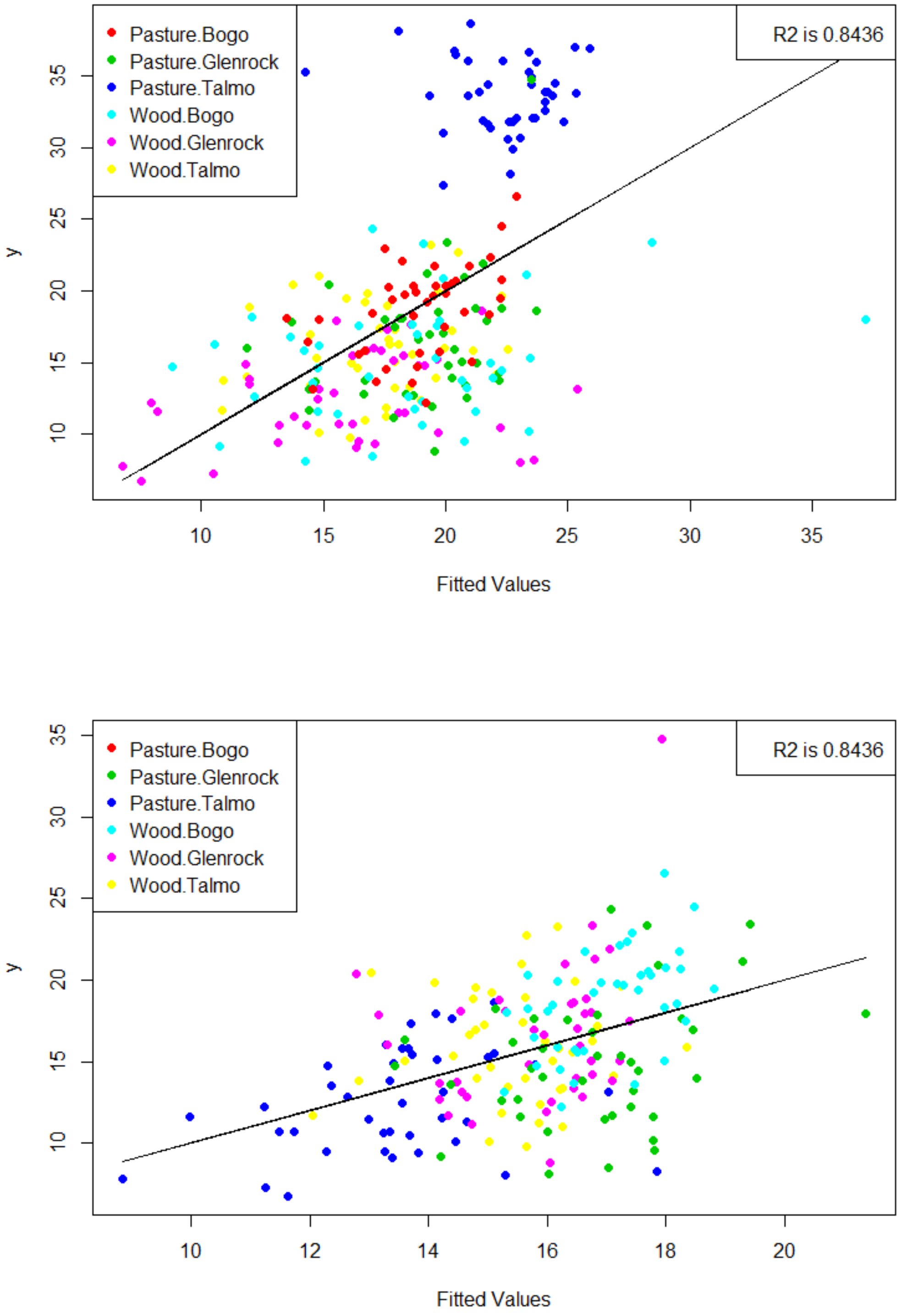
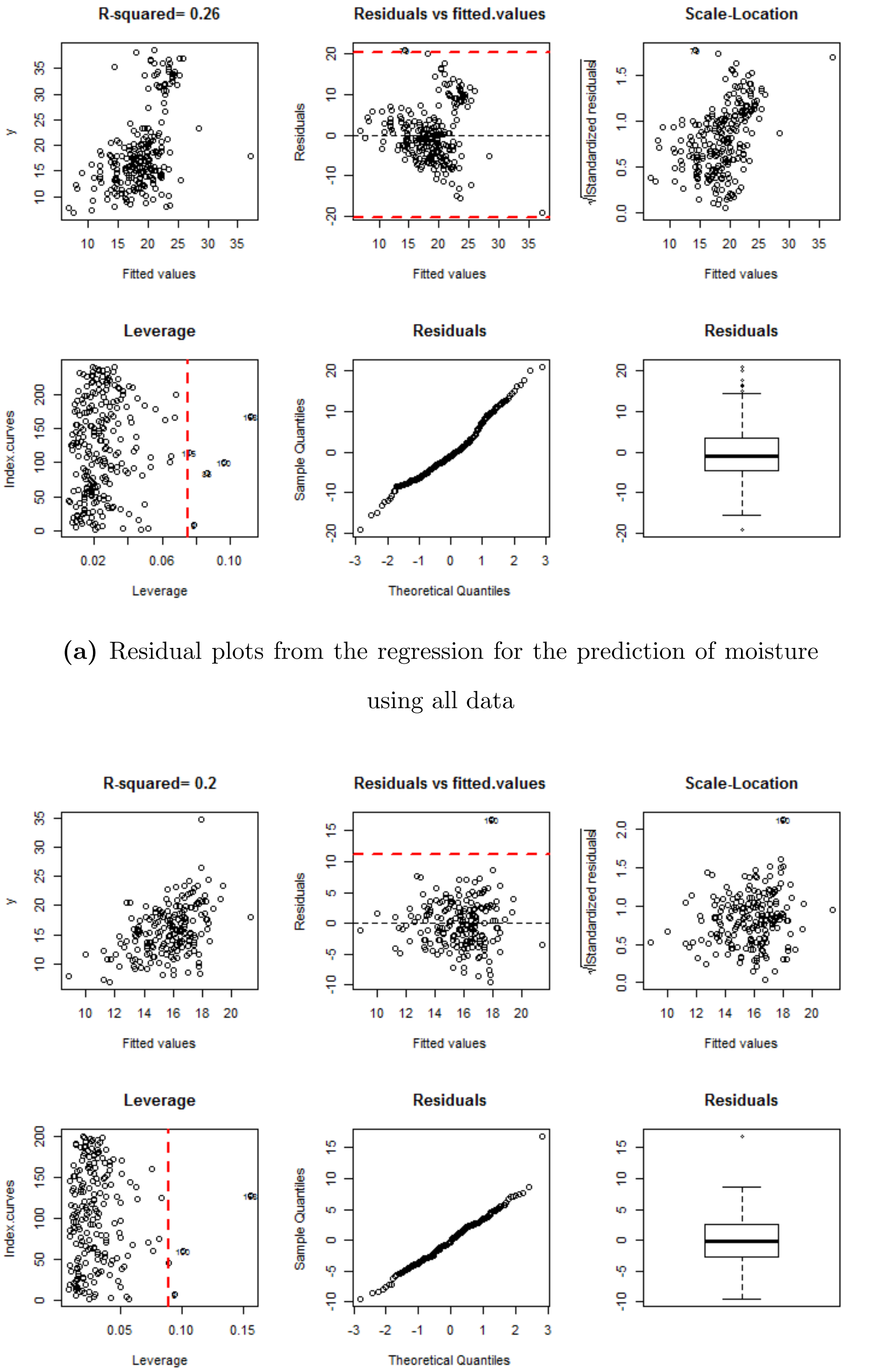


Figure 4.8: Fitted value plots for the prediction of moisture with data on all sites (top) and with the exclusion of Talmo pasture data (bottom)



(a) Residual plots from the regression for the prediction of moisture using all data

(b) Residual plots from the regression for the prediction of moisture with the exclusion of the Talmo pasture data

Figure 4.9: Residual Diagnostic Plots with and without the inclusion of Talmo Pasture Data

4.6.3 Scalar-on-Function Regression by land-use

The fitted values plots gave some evidence for the rationale behind constructing separate regression models based on different land-use types. Additionally, the functional F-tests in Chapter 3 also indicated that there were significant differences between woodland and pasture groups of spectra.

In line with the previous section, Table 4.3 presents the R^2 values for regression models based on the three subsets of the data investigated. Following the separation of land-use types it was found that higher R^2 values were achievable. The R^2 values highlighted in bold indicate that in comparison with the original regression, the overall fit has been improved. The values not highlighted in bold exhibited poorer fit than the original regression model.

Wet Chemistry Variable	R^2	R^2 (woodland)	R^2 (pasture)
pH	0.381	**0.535**	**0.459**
Moisture	0.258	0.200	**0.394**
Sqrt Carbon (C)	0.625	**0.686**	**0.720**
Sqrt Nitrogen (N)	0.611	**0.658**	0.538
Total Dissolved Nitrogen	0.153	**0.187**	0.028
Dissolved Organic Nitrogen (DON)	0.150	**0.215**	0.079
LogAmino-N	0.106	**0.227**	0.079
Log NH4-N	0.065	**0.191**	**0.119**
Log NO3.N	0.122	0.075	**0.136**
Biomass N	0.275	**0.363**	0.190
Total Dissolved Carbon (TDC)	0.117	**0.326**	0.084
Microbial Carbon	0.246	**0.330**	0.185
Inorganic Phosphorus	0.335	**0.377**	**0.631**
Organic Phosphorus	0.550	**0.557**	**0.602**

Table 4.3: Soil chemical, physical and mineralogical properties, with indicative coefficients of determination (R^2) or capability predicted using MIR

Woodland Models: For the woodland data specifically, in almost all cases the model fit improved with the exceptions of models predicting N03-N and Moisture. Similarly to the original regressions, the models associated with Carbon, Nitrogen and Organic phosphorus achieved the highest coefficients of determination with R^2 values of 0.686, 0.658 and 0.557 respectively. Interestingly, the most notable improvement was with the model for Total Dissolved Carbon achieving an R^2 value of 0.326 contrasting with a value of 0.117 from the original model. The pH regression also improved to an R^2 of 0.535 in contrast to a previous value of 0.381. This example is presented in the fitted value plots in Figure 4.10.

The top panel gives the original fitted values plot, the centre panel gives the fitted values plot corresponding to the pasture regression model and the lower panel gives the plot corresponding to the woodland specific model. It is observed that higher R^2 values are achieved by splitting the data into these separate models. The original regression achieved an R^2 value of 0.381 and following the separation, R^2 values of 0.535 and 0.459 were realised for woodland and pasture regression models respectively. A stronger positive linear relationship can be observed with the fitted values of the Woodland model in Figure 4.10 (bottom). This is exhibited by a greater density of clustering around the fitted line. In contrast, the spread of points in the Pasture model's fitted value plot is slightly wider. However, despite the improved R^2 values in the land-use specific models, the residual diagnostics revealed that the original model still performed adequately in capturing the relationship between pH values and the spectra.

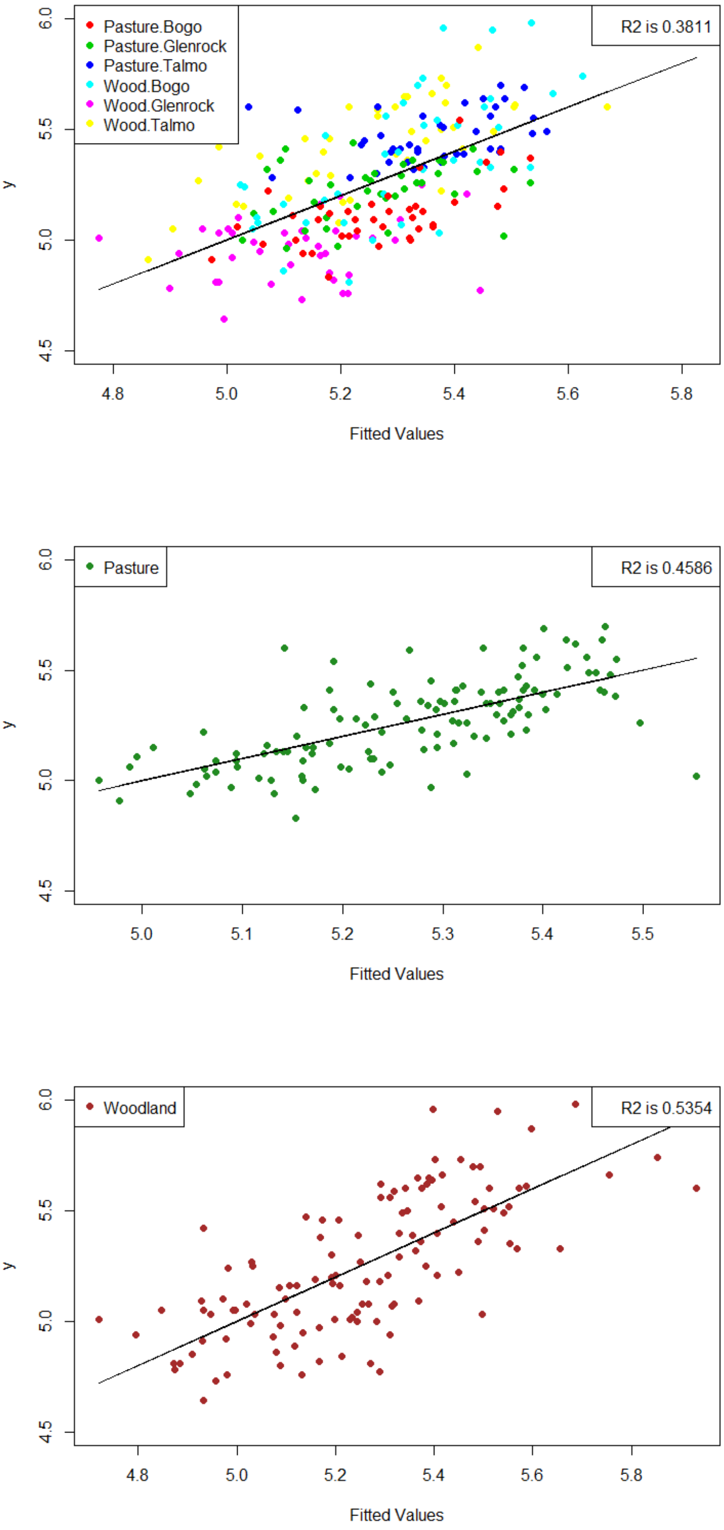


Figure 4.10: Fitted values plots for the regression models predicting pH using all data (top), only pasture data (centre) and only woodland data (bottom).

Pasture Models: Constructing models based on the pasture data alone did not highlight the same overall improvements that were observed with the woodland models. Only half of all pasture models achieved better fit than their original models. However, in the cases that the changes in R^2 values were negative, these changes were mostly negligible and these were found with original models which were already poorly fitting. Following the changes in fit reported in Table 4.3, the overall conclusions about which models achieved the best fit remained the same. Interestingly, performing separate regressions appears to have considerably improved the prediction of Inorganic Phosphorus. In the original regression model for the prediction of Inorganic Phosphorus, an R^2 value of 0.335 was achieved but following the split, R^2 values of 0.377 and 0.631 were realized for the woodland and pasture regressions respectively. Although the fit of the woodland model has not improved by much, the fit of the pasture model has considerably improved.

Similarly, Figure 4.11 gives all the fitted value plots for the regression models involving Inorganic Phosphorus. The fitted values of the pasture model are observed to follow the fitted line more tightly than those of the woodland model. The data also appear to be associated with different ranges and this can be more easily visualized by boxplot comparison in Figure 4.12. This gives further reason to investigate regression models per land-use. As before, all the relevant diagnostic checks were performed and assumptions appeared justified for all models constructed.

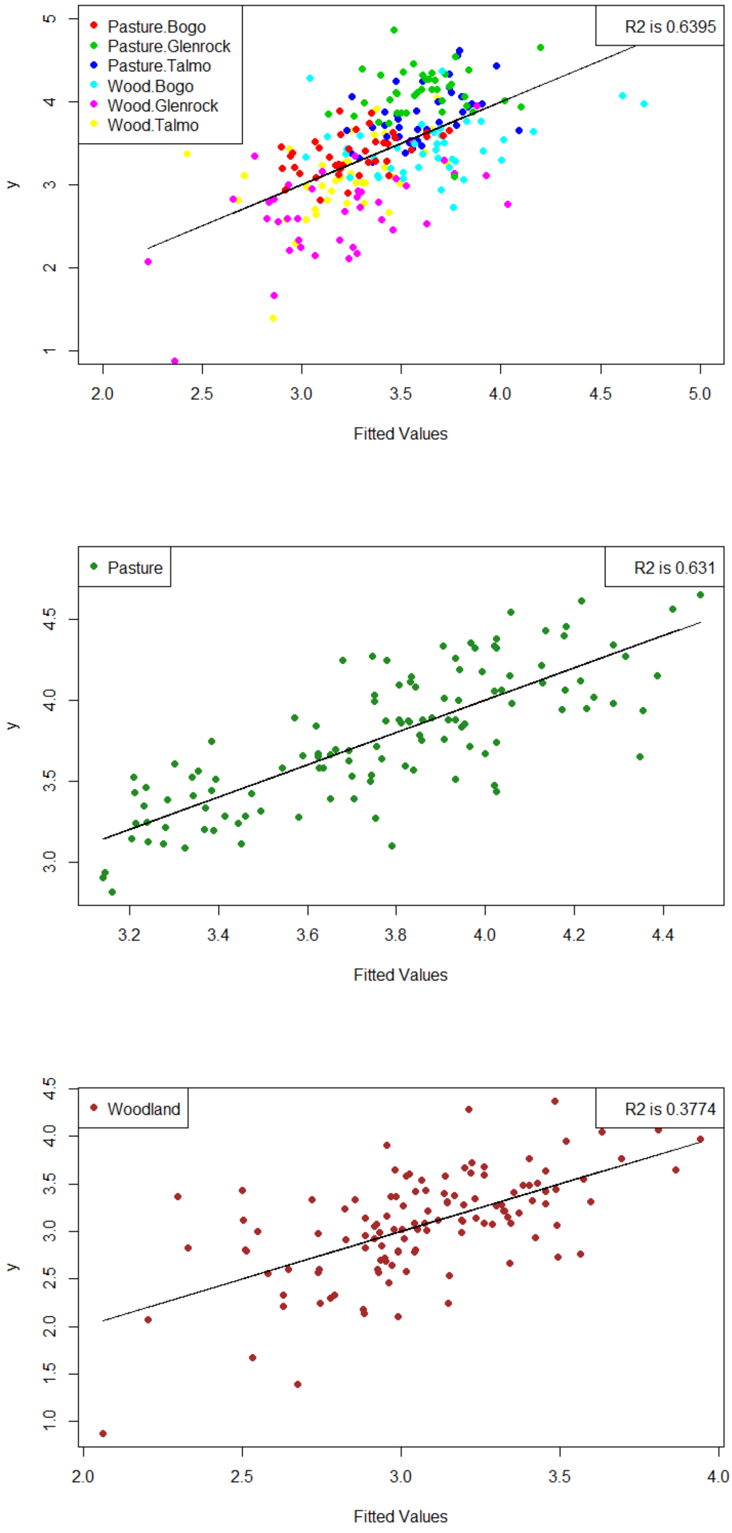


Figure 4.11: Fitted values plots for the regression models predicting Inorganic Phosphorus using all data (top), only pasture data (centre) and only woodland data (bottom).

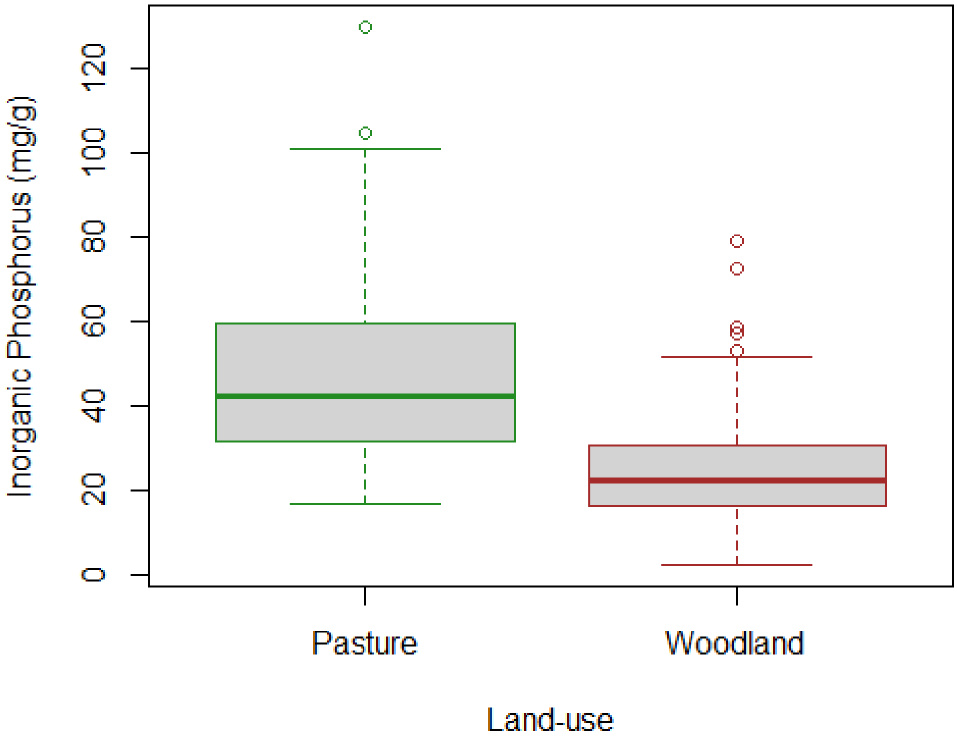


Figure 4.12: Boxplots of Inorganic Phosphorus Wet Chemistry Data illustrating the different ranges observed between land-uses

Interpretation of Coefficient Function Plots

Following the identification of the best fitting models the focus turns to the coefficient functions, $\beta(t)$, and their interpretation. For illustration, the reported interpretations are limited to the top two best fitting models as indicated by the coefficients of determination. These models were developed to predict Carbon and Nitrogen. A summary of the coefficients of determination for these regression models are given in Table 4.4.

Wet Chemistry Variable	R^2	R^2 (woodland model)	R^2 (pasture model)
Square-Root Carbon (C)	0.625	**0.686**	**0.720**
Square-Root Nitrogen (N)	0.611	**0.658**	0.538

Table 4.4: Soil chemical, physical and mineralogical properties, with indicative coefficients of determination (R^2) or capability predicted using MIR

In the scalar-on-function set-up the coefficient plots indicate the effect that the functional MIR spectra has on a particular soil wet chemistry response at any given wavenumber. As a reminder, the natural interpretation of the coefficient function, $\beta(t)$, is that the locations t with largest $|\beta(t)|$ are most influential to the response. That is, regions of the coefficient function plots where $\beta(t) \neq 0$ correspond to the regions of the spectra where a relationship between the MIR spectra, $X(t)$, and corresponding wet chemistry response exist. Conversely, where $\beta(t) = 0$ there is thought to be no indication of a relationship.

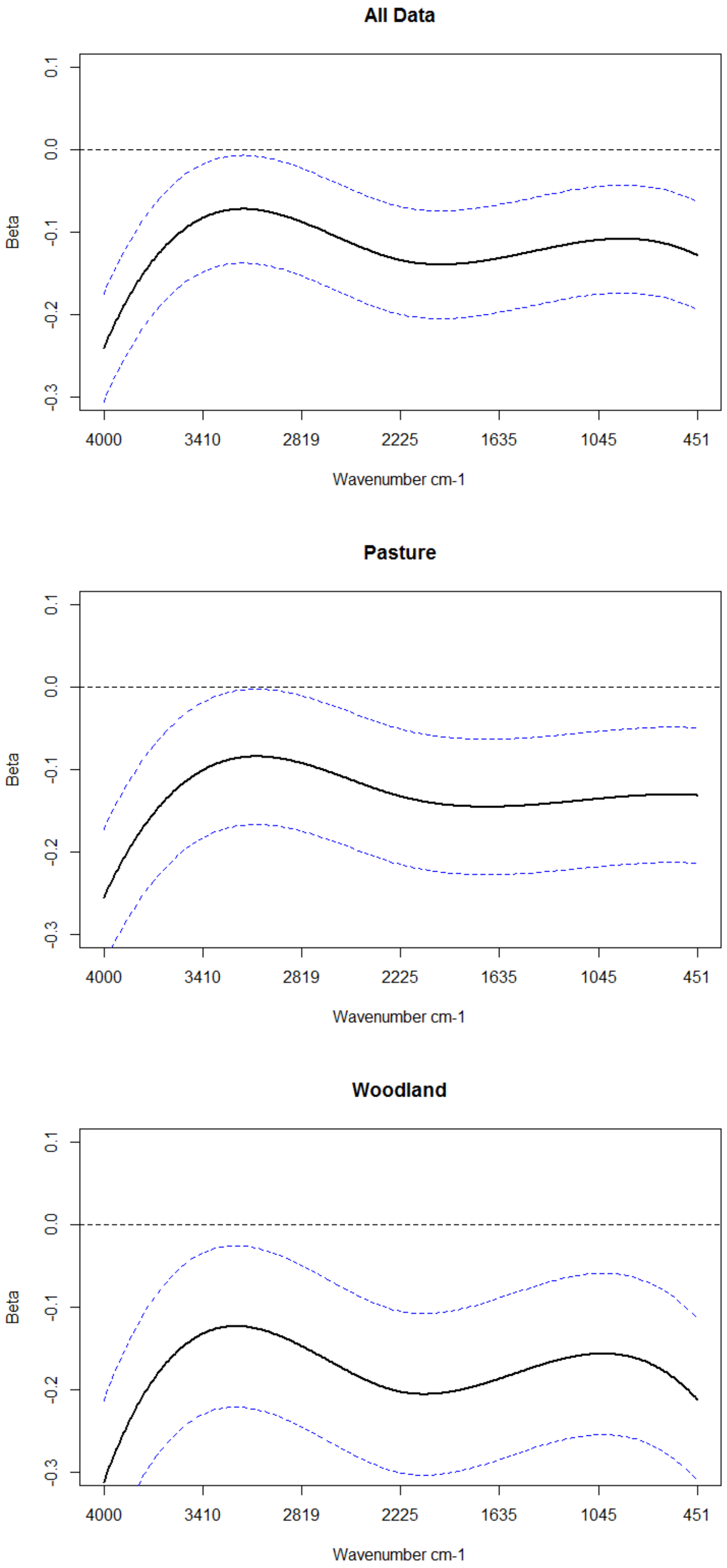
Given that each scalar-on-function regression was performed with only a single functional covariate, there is only one coefficient function plot associated with each regression model. In addition, confidence intervals for the coefficient functions were produced in order to allow conclusions about where along the wavenumber axis a covariate plays a strong role in prediction. Where the confidence intervals contain zero it can be said that the effect of the MIR spectra on the respective wet chemistry response is not statistically significant. In contrast, regions whereby the confidence interval does not contain zero are thought to be important in prediction.

Carbon Model Coefficient Function Estimates:

The coefficient function plots with associated 95% pointwise confidence intervals for the regression models concerned with the prediction of Carbon are given in Figure 4.13. To reduce the chances of obtaining any Type I errors (false-positive results), Bonferroni corrections were made for the calculation of the pointwise intervals. The confidence intervals do not contain the value zero at any point and thus reveal that the functional MIR spectra has a significant influence on Carbon values across the entire spectrum.

The Bonferroni correction is used to reduce the chances of obtaining false-positive results (type I errors) when multiple pair wise tests are performed on a single set of dat

However, the shapes of the coefficient functions indicate where the spectra has the greatest effect on Carbon. The coefficient function plot from the functional regression on the entire dataset (Figure 4.13 (top)) shows that the greatest influence is exhibited at the beginning of the spectra at around 4000cm^{-1}. This influence then diminishes with the weakest effect exhibited around 3000cm^{-1}. The coefficient estimate then drops to around a -0.15 effect on Carbon and remains steadily around this value for the remainder of the range (2000-450cm^{-1}).



All Data

Pasture

Woodland

Figure 4.13: Coefficient Function Estimates with confidence intervals from the full Carbon Regression Model (top), pasture regression model (centre), and woodland regression model (bottom).

The shapes of the coefficient functions for the land-use subsets are very similar and this suggests that similar relationships exist between the spectra and the Carbon responses of either land-use type. However, the confidence interval provided for the woodland estimate appears slightly wider than that of the pasture estimate. This suggests that pasture functional regression models can achieve higher accuracy in prediction. The pasture coefficient function estimate (Figure 4.13 (centre)) also appears the closest in shape to that of the general model. The estimates for the woodland coefficient function appear slightly larger in absolute value and thus have a greater effect on the response.

Nitrogen Model Coefficient Function Estimates:

The coefficient function plots with associated confidence intervals for the regression models concerned with the prediction of Nitrogen are given in Figure 4.14. Each coefficient function exhibits a very similar shape and gives an overall negative effect on the Nitrogen responses at all points along the spectral range. Whilst the construction of separate regression models per land-use did not have much of an effect on the estimates of Carbon, this is not the case for Nitrogen. Study of the confidence intervals indicated that the effects observed in the coefficient function of the original model are not statistically significant. However, the confidence intervals of the pasture and woodland models indicated statistically significant negative effects on Nitrogen across the entire spectral range.

The shapes of the coefficient function estimates remain very similar in the new models. However, confidence intervals have become a lot tighter and zero is not contained within the intervals. This shows that there has been a benefit to constructing models separately by land-use type. The spectra is now shown to have a significant effect on the Nitrogen response in each pasture and woodland model. For both cases however, it appears that the same areas of the spectra are important for the prediction of Nitrogen and have similar influence on the response.

4.6.4 Alternative Functional Data Representation for Scalar-on-Function Regression

In a brief investigation into the representation of the functional data it appears that the level of smoothing has influenced the prediction performance of the functional regression models of the previous section. Originally, the spectral data were smoothed via B-splines with $k = 45$ basis functions. The placement of knots were irregularly spaced corresponding to the areas of highest curvature in the spectra. By varying k and keeping the placement of knots equal across the argument range, various functional data representations of the spectra were obtained. Scalar-on-function models were constructed using these new functional data and their prediction performance is presented in Table 4.5.

Through increasing k larger coefficients of determination can be achieved by the scalar-on-function regression models. For example, the functional regression models predicting Carbon and Nitrogen achieved R^2 values of 0.62 and 0.61 respectively in the original model. By changing the smoothing design and increasing the number of basis functions to $k = 200$ then these R^2 values increase to 0.95 and 0.93 for Carbon and Nitrogen respectively. These improvements in R^2 are achieved for all models through increasing k and using an equal spacing approach for knot placement. However, caution must

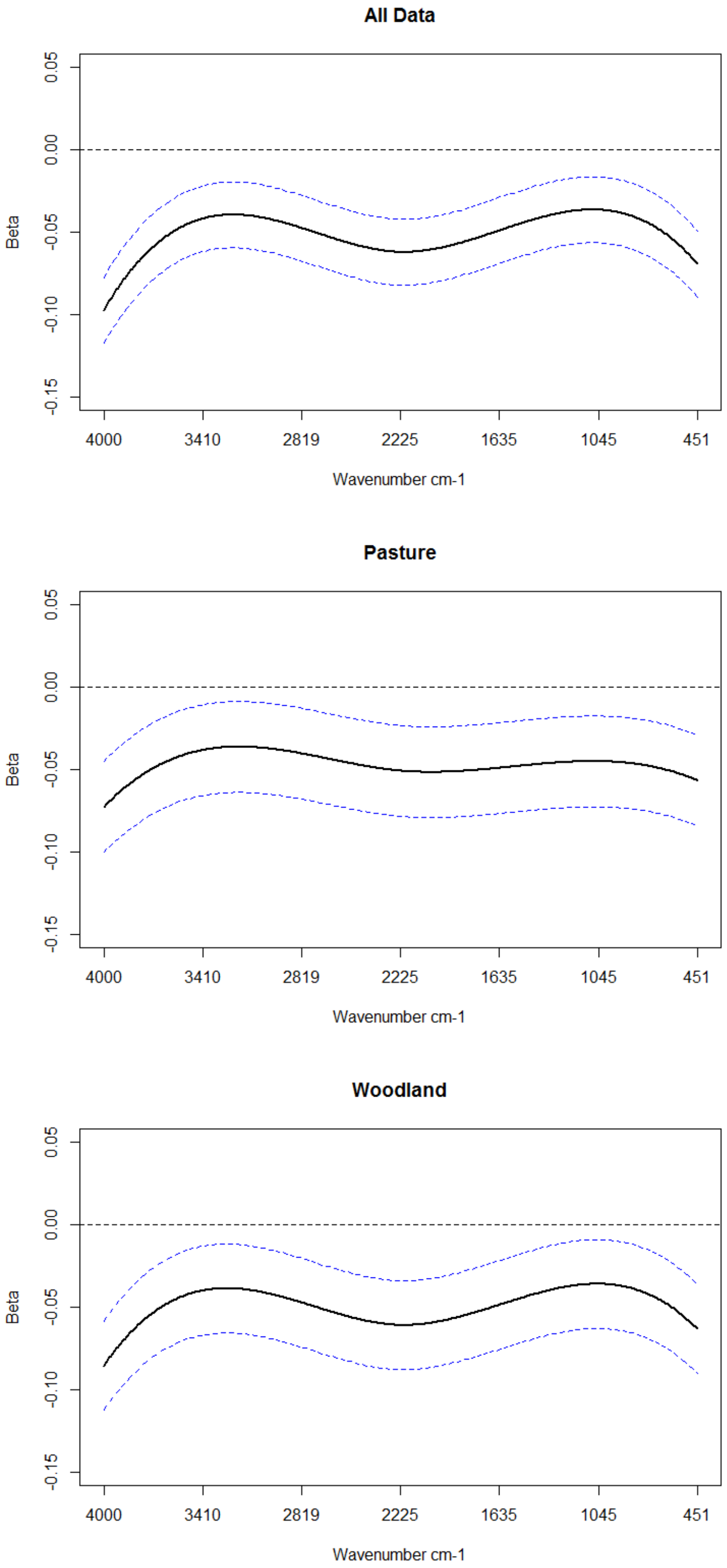


Figure 4.14: Coefficient Function Estimates with confidence intervals from the full Nitrogen Regression Model (top), pasture regression model (centre), and woodland regression model (bottom).

be taken in interpreting these results. By increasing k, the level of smoothing is altered such that more flexibility is allowed and there could be a case of overfitting with a large k. Thus, there needs to be care in choosing k such that the resulting model does not describe random error or noise instead of the underlying relationship.

Model Response	R^2 (original)	R^2 (k=100)	R^2 (k=150)	R^2 (k=200)
pH	0.38	0.73	0.79	0.82
Moisture	0.26	0.43	0.66	0.72
Carbon	0.62	0.80	0.94	0.95
Nitrogen	0.61	0.77	0.91	0.93
TDN	0.15	0.13	0.17	0.38
DON	0.15	0.18	0.31	0.40
Amino-N	0.11	0.15	0.23	0.35
NH4-N	0.07	0.19	0.28	0.37
N03-N	0.12	0.26	0.37	0.40
Biomass-N	0.28	0.31	0.62	0.65
TDC	0.12	0.18	0.28	0.38
Microbial Carbon	0.25	0.37	0.48	0.55
Inorganic Phosphorus	0.34	0.47	0.62	0.64
Organic Phosphorus	0.55	0.67	0.83	0.84

Table 4.5: Coefficients of Determination obtained from Scalar-on-Function Regression Models with different levels of smoothing applied to the functional data object

4.7 Applied Partial Least Squares Regression

It has already been established that Partial Least Squares Regression (PLSR) is the industry standard for the characterization of soil spectra and has been successful in the prediction of a multitude of soil consituents. In this section, prediction models are developed using PLSR to investigate what is currently acheivable without a functional regression approach. PLSR is a method which reduces the number of predictor variables to a smaller set of uncorrelated components and carries out least squares regression on these components. Often, PLSR is used when the predictor variables are highly collinear and this is reason for its application to soil spectral data with highly correlated wavenumber variables.

In examining the prediction performance of regression models, the original dataset is often split into training and test datasets. The training data are used to estimate and fine-tune the parameters in a model, while the test sets are used to validate model performance. Thus, a test set is kept entirely independent from the training set. However, this approach was not adopted in evaluating the performance of the functional regression models of the previous section. This would have involved the construction of many different functional data representations, and comparisons between models with different levels of smoothing applied to their spectral responses would not have been permitted. In this section, the original data were randomly split so that 75% of the total observations were used in all training datasets with the remaining 25% making up the test datasets.

In building a PLS regression model, an important part of the process is selecting the optimal number of components. Initially, PLS models were constructed using the training datasets and allowed as many as 20 components to be considered. The optimal models for each soil property were then determined by choosing the number of components that gave the first local minimum in cross-validated root mean squared error of prediction (RMSEP). Figure 4.15 (left) gives a plot of the RMSEP against the number of components for the general PLS training model concerning the prediction of Carbon.

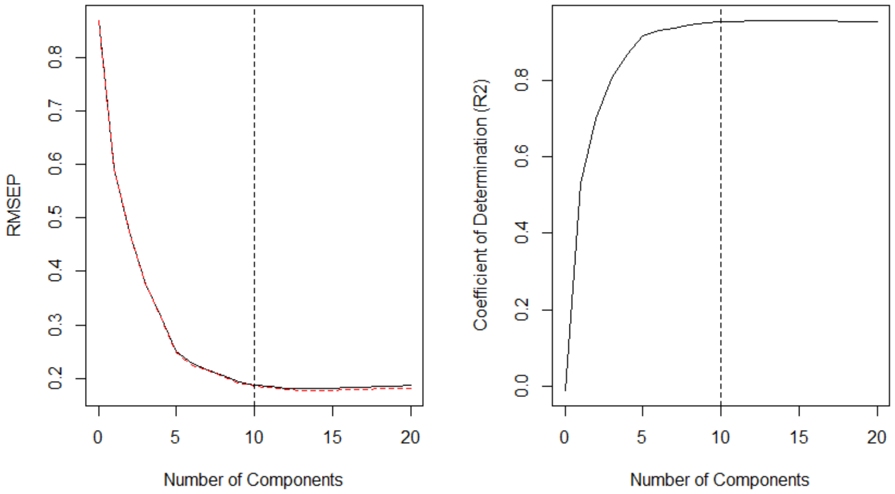


Figure 4.15: Plot of the root mean square error of prediction (left) and coefficient of determination (right) against the number of components retained in a PLS regression model for Carbon.

The first local minimum RMSEP can be observed when 10 components contribute to the PLS model. Thus, only the first 10 components are retained in the prediction model for Carbon. Figure 4.15 (right) demonstrates how the number of components retained affects the R^2 value which indicates goodness of fit. The first two components in the PLS model are able to explain approximately 70% of the relationship between the original predictor space and the response variable. However, the optimal model is chosen with 10 components and achieves an R^2 value of 0.95. Thereafter, any additional components do not add greatly to the R^2. In general, a PLS regression model is thought to provide significant and good predictions when R^2 is greater than 0.5. Thus, the model for the prediction of Carbon appears to be a very good fitting model. Training models were constructed in the same manner for the prediction of all wet chemistry variables. However, all these training models require to have their performance validated using the test sets of previously unseen data.

The prediction equations obtained from the PLS training models were extracted and applied to the validation test sets. The predictive abilities of the models were then assessed through the coefficients of determination for both training and test data. The ${R_c}^2$ values inform on the performance of the PLSR on the training data and the ${R_v}^2$ values indicate how well the models perform on the test data. These coefficients of determination indicated that successful prediction and better fitting models were found with the same core soil property variables as found by the functional linear regression models. These were for Carbon, Nitrogen, Organic Phosphorus and Inorganic Phosphorus. Additionally, PLS regression also found good predictive ability for Moisture and pH. Table 4.6 presents the coefficients of determination obtained. The soil property with best behaviour and best fitting relationship was Nitrogen. This model achieved an R_c^2 value of 0.95 with the training data and excellent validation performance with R_v^2= 0.92. Closely following, the PLS model predicting Carbon achieved coefficients of determination of 0.94 and 0.87 in the calibration (training) and validation subsets respectively. To illustrate the excellent performance of the PLS models in the prediction of Carbon and Nitrogen, Figure 4.16 presents their Observed vs Predicted value plots.

Wet Chemistry Variable	R_c^2	R_v^2
pH	0.80	0.81
Moisture	0.75	0.72
Carbon	0.94	0.87
Nitrogen	0.95	0.92
Inorganic Phosphorus	0.53	0.61
Organic Phosphorus	0.79	0.88

Table 4.6: Coefficients of Determination from PLSR models using calibration (c) and validation (v) data

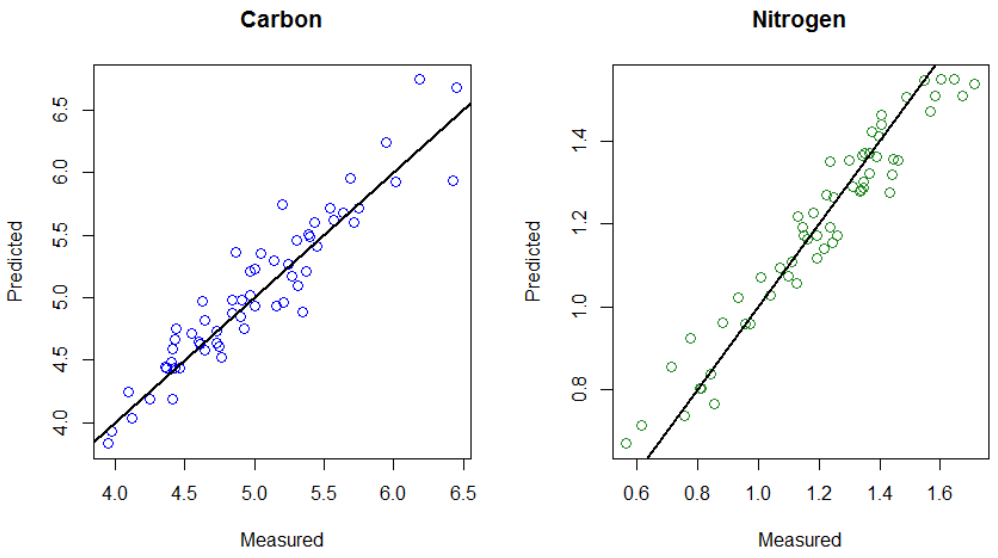


Figure 4.16: Observed vs Predicted Values for Carbon (left) and Nitrogen (right) PLSR models with $y = x$ line

These plots demonstrate the prediction performance of the models constructed using the training data evaluated on the validation subsets for each variable. In both examples, all points appear to follow the target line closely with no indication of curvature or other anomalies. In particular, the points corresponding with Nitrogen estimates appear to follow the line $y = x$ very tightly. This is reflected in the R_v^2 values, with a superior R_v^2= 0.92 achieved in the prediction of Nitrogen compared with R_v^2= 0.87 from the prediction of Carbon. The remaining wet chemistry variables of Table 4.6 all exhibited good behaviour with all R^2 values greater than 0.5. All other PLS models exhibited low values of R_v^2 and poor predictive ability of their respective soil properties.

Figure 4.17 gives the loadings for the first two PLS components from the PLS model predicting Carbon. These PLS components cumulatively represent 66.2% of the total variability and individually explain the most variability of all the PLS components. These components can be given a physically meaningful interpretation by inspecting which variables they weight most heavily. Similarly to PCA, it is possible to interpret the intensity peaks of the PLS loadings in terms of compounds present in the soil.

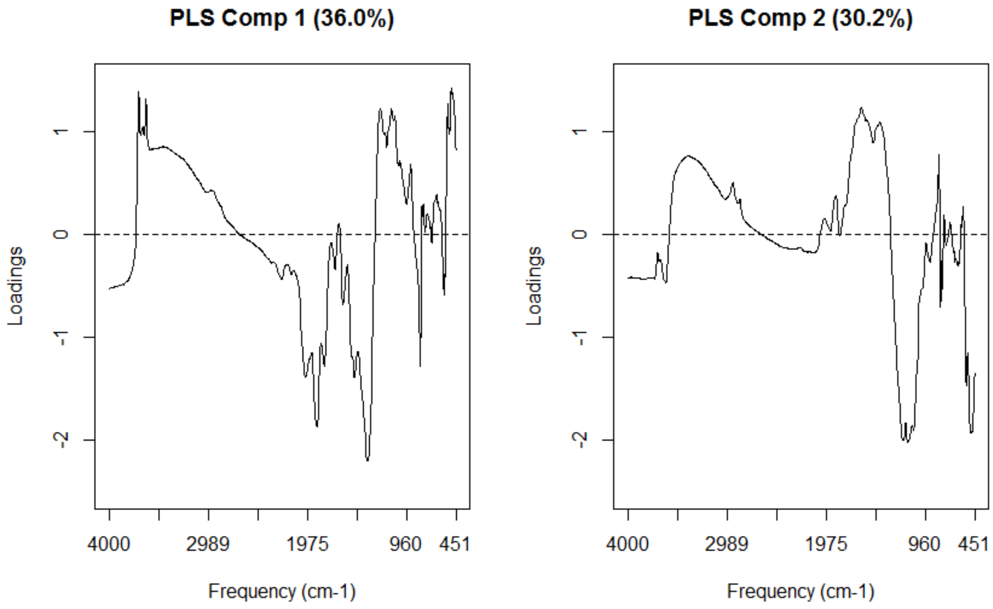


Figure 4.17: Loadings for PLS Component 1 and PLS Component 2 from the PLS regression model for the prediction of Carbon

4.8 Summary

In this chapter, two forms of functional linear regression were successfully applied to relate the MIR spectra to the wet chemistry. Function-on-scalar regression found that the scalar covariates associated with pH, Amino-N, Carbon and Nitrogen had the most influence on the MIR spectra with the remaining covariates having minimal or no effect on the response. Similarly, it was shown that principal components could be incorporated into a function-on-scalar regression model. However, different scalar covariates were identified as having the greatest influence on the spectra by this method. Since PCA only identifies components which account for the total variability in the wet chemistry dataset, this does not necessarily mean that the variables identified by this approach are those that are most strongly related to the response. Following on from function-on-scalar regression, the more interesting application of scalar-on-function regression was applied. The models constructed in this approach achieved best prediction performance for scalar responses characterized by Carbon, Nitrogen, Organic Phosphorus and Inorganic Phosphorus. Furthermore, performing scalar-on-function regression on a land-use basis was found to achieve higher prediction performance. This was especially true for regression models based on pasture data. The influence of the MIR spectra on the prediction of wet chemistry was examined via coefficient function plots for each of the soil variables. It was found that the MIR spectra had different intensities of influence depending on the wavenumber region considered. In a brief investigation into different functional data representations of the soil spectra it was found that the level of smoothing has influenced the prediction performance of the scalar-on-function regression models. It was found that by varying k and standardizing the equal placement of knots, greater coefficients of determination could be achieved.

Furthermore, the results of partial least squares regression mirrored that of the scalar-on-function regression models in terms of the most successful predictions. The best predictive ability was found for Carbon, Nitrogen, Organic Phosphorus and Inorganic Phosphorus. Additionally, PLSR demonstrated successful prediction performance for Moisture and pH. In terms of fit, the PLSR models appeared to outperform the original scalar-on-function models; achieving higher R^2 values across the board. However, with k= 200 the prediction performances of scalar-on-function regression models actually appear to rival that of PLSR. For example, the functional regression model predicting Nitrogen (k= 200) attains an R^2 value of 0.93 and the corresponding model via PLSR attains an R_v^2 of 0.92. By allowing more flexibility in the smooth representation of the functional data, functional regression models can have greater prediction performance.

Chapter 5

Findings and Conclusions

5.1 Introduction

Functional data analysis provides an alternate way of studying continuous data, recognising that it is sometimes more natural, and often fruitful, to view a collection of data points as observed realisations of random functions. The evolution of data collection technologies is allowing vast amounts of data to be recorded and often at a large number of finely spaced observation points. These new data collection capabilities strengthen the justification for taking a functional approach across a wide range of disciplines. Furthermore, functional data analysis provides access to many functional equivalents of methods currently used in chemometrics, with the benefits of no strong assumptions regarding neighbouring observations.

Motivated by problems of MIR diffuse reflectance spectroscopy for the prediction of soil properties, this thesis explored functional data applications to the mid-infrared spectra of soil samples. Traditionally, classical multivariate approaches such as linear discriminant analysis, principal component analysis and partial least squares regression are used to explore the soil spectra.

In particular, these multivariate approaches are used in conjunction with mid-infrared spectroscopy as a cost effective means of characterizing the soil in contrast with expensive laboratory procedures. In soil science, the use of functional data analysis has been under-represented and the focus remains on these standard multivariate approaches. This is despite the fact that the functional data analysis methods appear more theoretically coherent, especially with spectral data being inherently functional in nature. Furthermore, functional data anaysis makes no assumption of independence between adjacent observations- an assumption often violated by the classical approaches.

5.2 Methods Discussion

5.2.1 Multivariate Methods

The thesis began with an investigation into standard multivariate approaches to explore the soil spectra and to give a baseline of what is currently achievable without FDA. Principal Component Analysis successfully reduced the high-dimensional spectra defined at 921 wavenumbers to a set of just eight wavenumber variables (principal components). This is a massive reduction in dimensionality and a high percentage (97.5%) of the total variability in the data was preserved. Additionally, with the combined study of loadings and scores plots, it was possible to identify wavenumbers with spectral differences depending on site/land-use groupings. Principal component analyses on independent groups of spectra found that the clearest separations could be found between land-uses within sampling sites. However, these differences were slight and made for subjective inferences to be made based on loadings plots. Likewise, stepwise linear discriminant analysis identified wavenumbers responsible for class separations in regions not dissimilar to those identified by PCA. The classification rates achieved by the linear discriminant analyses were also very high. Similarly, support vector machines demonstrated high classification rates. However, the technique was very black-box in nature and did not inform graphically, or otherwise, as to the locations along the spectra which may be of interest.

There has been evidence in recent literature of the successful prediction of a wide range of soil quantities via Partial Least Squares Regression. This technique was also applied in this thesis and found particularly successful in the prediction of Carbon, Nitrogen, Organic Phosphorus, Inorganic Phosphorus, Moisture and soil pH. The PLS models predicting these soil properties were evaluated on test datasets via coefficients of determination.

5.2.2 Functional Data Analyses

In place of working with discrete data, the abilities of functional data analyses were showcased from Chapter 3. Taking advantage of the smooth functions underlying the data, various methods of functional data analysis proved successful. Functional Principal Component Analysis was explored and the method reduced the dimensionality of the soil spectra data to just seven functional principal component curves whilst explaining 96.17% of the total variability. Furthermore, functional hypothesis testing revealed that all groups of spectra were statistically distinct and additionally had the ability to inform where along the wavenumber range these differences were most significant.

In Chapter 4 functional linear regression was introduced for the prediction of soil properties as measured by laboratory based methods. Firstly, function-on-scalar regression found that the mid-infrared spectra could be successfully predicted by the wet chemistry variables. In accompaniment, functional principal component regression showed how the wet chemistry variables could be reduced to a selection of principal components for the prediction of the soil spectra. However, these analyses were all secondary interests to the more interesting scalar-on-function regression. The scalar-on-function regression analyses permitted the prediction of a wide range of soil constituents from the functional data representation of the soil spectra. Carbon, Nitrogen, Organic Phosphorus and Inorganic Phosphorus were amongst the most successfully predicted, and the predictive ability of the models was assessed based on R^2 statistics. It was found that better prediction performance could be achieved by constructing models separately for woodland and pasture data. Once effective models had been identified, the coefficient function plots with confidence intervals were produced. Interpretation of these plots permitted conclusions to be made about what areas of the spectra had more or less of an influence on the value of the response predicted.

5.2.3 Multivariate Methods vs. Methods of Functional Data Analysis

Despite the success of the multivariate methods, functional data analysis provides a superior approach for identifying the regions of the mid-infrared spectra which comprise of the essential modes of spectral variation. The absorbance of different soil quantities is not limited to specific spectral wavenumbers and spans absorbance regions of the spectra. For this reason, Linear Discriminant Analysis is not effective in the context of relating regions of the spectra to components in the soil. Linear Discriminant Analysis is found to be too specific to the problem and cannot relate an individual wavenumber to a certain soil property over another due to the overlapping nature of absorption regions of different soil properties. The intensity peaks in the loadings of a principal component analysis are similarly difficult to interpret and any inferences made can be subjective. These methods were contrasted to the results of a functional principal components analysis. Whilst FPCA suffers the same interpretability issues which come with the overlapping absorption regions, FPCA has the ability to direct attention to wavenumber regions of interest. This reduces the problem of interpretability and minimizes the risk of overlooking soil quantities which could contribute significantly to differences observed in the spectra.

In the prediction of wet chemistry from the MIR spectral data, Partial Least Squares Regression proved to be more successful than the scalar-on-function approach with a superior level of performance achieved for all soil quantities considered. However, with further exploration into different levels of smoothing for functional data objects it was found that the functional regression methods could be just as effective as PLSR. With this in mind, functional regression could be considered to be more favourable since it can provide extra information on prediction through coefficient function plots. Presented in the same domain as the spectra, the coefficient function plots allow the user to identify which regions of the spectra are best for predicting an individual soil constituent. This information is lost with a partial least squares regression approach, and thus there is no indication of a spectral range where focus could be concentrated for further study. Despite this, the performance of PLSR was evaluated on training and test data and thus direct comparisons with the scalar-on-function approach is not permitted.

However, the formation of training and test data sets to evaluate the performance of functional regression models is challenging; involving separate functional data representations for both training and test data.

More broadly, functional data analysis has a wealth of functional exploratory methods which can reveal aspects of the data lost in a multivariate framework. Functional data analysis allows for the ordering of data outward from the centre based on a notion of band depth. Through the development of these functional depth measures, functional boxplots are available to be used as an informative exploratory tool. Additionally, functional measures of centrality and assessment of functional outliers are possible. Although not investigated in this thesis, functional data analysis also provides access to the data derivatives. Functions are differentiable, and the derivatives can be a potential source of additional information about the nature of the data that is otherwise locked away in the data.

5.2.4 Shortcomings

Functional data analysis is a very powerful tool and is philosophically the most coherent of approaches for the characterization of soil properties. However, in addition to its advantages it also has shortcomings. One of the prevailing issues associated with functional data is the initial construction of a functional data object. The first step in any FDA is to convert raw discrete values into functions via methods of smoothing. Various smoothing techniques are available and they are applied to emphasize patterns in the data by minimizing short-term deviations due to errors. However, the process of transforming data into a functional form is not an exact science and somewhat of an art form. Given the nature of the spectra, a basis function approach using B-splines was chosen to smooth the data due to their high degree of flexibility and ability to model sharp changes in curvature. Basis functions were also irregularly spaced and knots loaded at areas of high interest. However, a great deal of care must be taken not to over smooth data which would result in the loss of information and cause important features of the true underlying curves to be missed. Under smoothing would also cause noise (random error) to be estimated. In reflection, it is feasible to suggest that the spectra in this study did not require to be smoothed as much as they were. This is suspected due to the strong performance of the multivarate approaches which have already been in place. Furthermore, the data were very finely spaced and appeared fairly smooth to begin with. In FDA, the data of interest may be subject to substantial measurement error or could be so accurate that the associated errors may be ignored. The absorptions in the mid-infrared range were recorded at a large number of finely spaced wavenumber bands and thus the latter scenario is suspected.

Another issue with functional data representation is the decision to use irregularly spaced basis functions. The rationale was coherent in that it is more sensible to manually design the placement of knots so that more knots are placed in areas of the spectra with high curvature and less in areas of relative inactivity. By increasing the number of knots in a particular region this allows for more flexibility. This was guided by soil scientists to help identify where the pertinent areas of the spectra responsible for driving differences between groups of soil spectra were located. However, this method is not generalizable since these recommendations were made based solely on previous studies of the same Australian soil data. Thus, for a wider study of soils a standard approach to selecting the level of smoothing should be developed. This may involve different smoothing approaches for different soils as spectral signatures may differ vastly depending on their origins.

Although the methods of functional data analysis are not more computationally intensive than methods of a multivariate framework; the time taken to create a functional data object and the soil science input required makes functional data analyses a much lengthier process. However, in the development of a more efficient approach to smoothing the data this would not be an issue. This would also not be a problem for data of a more regular nature. For example, regional temperature data could be easily and quickly smoothed using Fourier splines. However, PLSR does appear to be a more robust method suffering from none of these issues.

Another downfall of functional data is the lack of consistent diagnostics. This is an area of functional regression which has been largely ignored in the literature. Classical regression diagnostics are based on residuals (Anscombe and Tukey, 1963) and have an important place in applied statistics for the task of checking model assumptions that underlie statistical analysis. These techniques have largely been limited to classical linear and non-linear regression models, where response and predictor variables are scalars. However, with increasing attention in functional regression analysis it is of interest to develop similar diagnostic procedures for functional regression models. Although there is a general lack of diagnostic procedures, some methods have been developed and for the reader's interest, Chiou and Muller (2007) provide some extensions for functional regression diagnostics.

5.3 Remaining Challenges

Finding the best functional data representation of the data is the backbone of functional data analysis and thus deserves careful consideration. Otherwise, by misrepresenting the underlying true functions a wealth of information may be lost. Given that the utility of functional data analysis is pivotal on achieving suitable functional data representation; further research is needed into how an efficient, reliable and optimal smoothing procedure can be developed for soil spectral data. Rossel et al. (2016) describe a global spectral library for soil vis-NIR spectra which has been in development since 2008. With the collection of such data and similar recording of soil mid-infrared spectra, analysis of spectral signatures across a wide range of soil origins could direct the development of functional data representation within soil science.

Bibliography

Allison, L. E., W. B. Bollen, and C. D. Moodie (1965). *Total Carbon.* Madison, WI: Agron. Monogr.

Anderson, D. and E. Gregorich (1984). Effect of soil erosion on soil quality and productivity. *Soil erosion and degradation*.

Anscombe, F. and J. Tukey (1963). The examination and analysis of residuals. *Technometrics*, 141–159.

Balakrishnama, G. (1998). Linear discriminant analysis - a brief tutorial.

Batten, G. (1998). Plant analysis using near infrared reflectance spectroscopy: the potential and the limitations. *Australian Journal of Experimental Agriculture*.

Batuwita, R. and V. Palade (2013). *Class Imbalance Learning Methods for Support Vector Machines.*

Bellino, A., C. Colombo, P. Iovieno, A. Alfani, G. Palumbo, and D. Baldantoni (2016). Chemometric technique performances in predicting forest soil chemical and biological properties from uv-vis-nir reflectance spectra with small, high dimensional datasets. *iForest - Biogeosciences and Forestry* (1), 101–108.

Box, G. E. P. (1979). Robustness in the strategy of scientific model building. *Robustness in Statistics, Academic Press*, 201–236.

Bricklemyer, R., P. Miller, K. Paustian, T. Keck, G. Nielsen, and J. Antle (2005). Soil organic carbon variability and sampling optimization in montana dryland wheat fields. *Journal of Soil and Water Conservation 60*(1), 42–51.

Cardot, H., F. Ferraty, and P. Sarda (1999). Functional linear model. *Statistics and Probability Letters 45*, 11–22.

Cattell, R. B. (1966). The scree test for the number of factors. *Multivariate Behavioral Research 1*(2), 245–276.

Cattell, R. B. (1978). *The scientific use of factor analysis.* (1st ed.). New York: Plenum Press.

Chen, Y., G. J. and R. Ogden (2016). Variable selection in function-on-scalar regression. *Stat 5*(1), 88–101.

Chiou, J. and H. Muller (2007). Diagnostics for functional regression via residual processes. *Comput. Stat. Data Anal. 51*(10), 4849–4863.

Cobo, J. G., G. Dercon, T. Yekeye, L. Chapungu, C. Kadzere, A. Murwira, R. Delve, and G. Cadisch (2010). Integration of mid-infrared spectroscopy and geostatistics in the assessment of soil spatial variability at landscape level. *Geoderma 158*(34), 398 – 411.

Conforti, M., R. Froio, G. Matteucci, and G. Buttafuoco (2015). Visible and near infrared spectroscopy for predicting texture in forest soil: an application in Southern Italy. *iForest - Biogeosciences and Forestry* (3), 339–347.

Cortes, C. and V. Vapnik (1995). Support-vector networks. *Machine Learning 20*(3), 273–297.

Cozzolino, D. and A. Moron (2006). Potential of near-infrared reflectance spectroscopy and chemometrics to predict soil organic carbon fractions. *Soil and Tillage Research 85*, 78 – 85.

Craven, P. and G. Wahba (1978). Smoothing noisy data with spline functions. *Numerische Mathematik 31*(4), 377–403.

De Boor, C. (1972). On calculating with b-splines. *J. Approximation Theory 6*, 50–62.

De Jong, S. (1993). Simpls: An alternative approach to partial least squares regression. *Chemometrics and Intelligent Laboratory Systems 18*(3), 251 – 263.

De Menezes, A. B., M. T. Prendergast-Miller, A. E. Richardson, P. Toscas, M. Farrell, L. M. Macdonald, G. Baker, T. Wark, and P. H. Thrall (2015). Network analysis reveals that bacteria and fungi form modules that correlate independently with soil parameters. *Environmental Microbiology 17*(8), 2677–2689.

Devos, O., C. Ruckebusch, A. Durand, L. Duponchel, and J. Huvenne (2009). Support vector machines (SVM) in near infrared (NIR) spectroscopy: Focus on parameters optimization and model interpretation. *Chemometrics and Intelligent Laboratory Systems 96*(1), 27 – 33.

Doran, J. and T. Parkin (1994). Defining and assessing soil quality for a sustainable environment. *Soil Science Society of America and American Society of Agronomy*.

Erbas, B., R. Hyndman, and D. Gertig (2007). Forecasting age-specific breast cancer mortality using functional data models. *Statistics in Medicine 26*(2), 458–470.

Etzion, Y., R. Linker, U. Cogan, and I. Shmulevich. (2004). Determination of protein concentration in raw milk by mid-infrared fourier transform infrared/attenuated total reflectance spectroscopy. *J Dairy Sci*.

Faraway, J. (2006). Extending the linear model with R. *Journal of the Royal Statistical Society: Series A (Statistics in Society) 169*(4), 1008–1008.

Febrero-Bande, M. and M. Oviedo de la Fuente (2012). *fda.usc: Functional Data Analysis and Utilities for Statistical Computing (fda.usc).* R package version 0.9.7.

Ferraty, F. and P. Vieu (2003). Curves discrimination: a nonparametric functional approach. *Computational Statistics & Data Analysis 44*(12), 161 – 173. Special Issue in Honour of Stan Azen: a Birthday Celebration.

Ferraty, F. and P. Vieu (2006). *Nonparametric Functional Data Analysis: Theory and Practice (Springer Series in Statistics).* Secaucus, NJ, USA: Springer-Verlag New York, Inc.

Feyziyev, F., M. Babayev, S. Priori, and G. L'Abate (2016). Using visible-near infrared spectroscopy to predict soil properties of Mugan Plain, Azerbaijan. *Open Journal of Soil Science*, 52–58.

Fidencio, P., R. Poppi, J. de Andrade, and H. Cantarella (2002). Determination of organic matter in soil using near-infrared spectroscopy and partial least squares regression. *Communications in Soil Science and Plant Analysis 33*(9-10), 1607–1615.

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics 7*(2), 179–188.

Forrester, S., J. L.J., S.-D. J.M., S. Mason, L. Burkitt, P. Moody, G. C.J.P, and M. McLaughlin (2015). Use of handheld mid-infrared spectroscopy and partial least-squares regression for the prediction of the phosphorus buffering index in Australian soils. *Soil Research 53*, 67–80.

Frank, I.E., F. J. (1993). A statistical view of some chemometrics regression tools. *Technometrics 35*(2), 109–135.

Garca-Portugus, E., W. Gonzlez-Manteiga, and M. Febrero-Bande (2014). A goodness-of-fit test for the functional linear model with scalar response. *Journal of Computational and Graphical Statistics 23*(3), 761–778.

Goldsmith, J. and F. Scheipl (2014). Estimator selection and combination in scalar-on-function regression. *Computational Statistics & Data Analysis 70*, 362 – 372.

Gorecki, T. and M. Luczak (2013). Linear discriminant analysis with a generalization of the Moore-Penrose pseudoinverse. *International Journal of Applied Mathematics and Computer Science 23*(2), 463–471.

Gorsuch, R. L. (1983). *Factor analysis* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

Green, P. and B. Silverman (1994). *Nonparametric regression and generalized linear models: a roughness penalty approach.* Chapman & Hall.

Haaland, D. and E. Thomas (1988). Partial least-squares methods for spectral analyses. *Analytical Chemistry 60*(11), 1202–1208.

Hastie, T. J., R. J. Tibshirani, and J. H. Friedman (2009). *The elements of statistical learning : data mining, inference, and prediction.* Springer series in statistics. New York: Springer.

Heinze, S., M. Vohland, R. G. Joergensen, and B. Ludwig (2013). Usefulness of near-infrared spectroscopy for the prediction of chemical and biological soil properties in different long-term experiments. *Journal of Plant Nutrition and Soil Science 176*(4), 520–528.

Hsu, C., C. C. and C. Lin (2010). A practical guide to support vector classification.

James, G.M., W. J. and J. Zhu (2009). Functional linear regression thats interpretable. *The Annals of Statistics 37*, 2083–2108.

Janik, L.J., F. S. and A. Rawson (2009). The prediction of soil chemical and physical properties from mid-infrared spectroscopy and combined partial least-squares regression and neural networks (PLS-NN) analysis. *Chemometrics and Intelligent Laboratory Systems 97*(2), 179 – 188.

Jolliffe, I. T. (2002). *Principal component analysis* (Second ed.). Springer.

Kaiser, H. F. (1958). The varimax criterion for analytic rotation in factor analysis. *Psychometrika 23*(3), 187–200.

Kaiser, H. F. (1960). The application of electronic computers to factor analysis. *Educational and Psychological Measurement 20*(1), 141–151.

Kalivas, J. (1997). Two data sets of near infrared spectra. *Chemometrics and Intelligent Laboratory Systems 37*(2), 255 – 259.

Kivinen, J., A. J. Smola, and R. C. Williamson (2004). Online learning with kernels. *IEEE Transactions on Signal Processing 52*, 2165–2176.

Larson, W. and F. Pierce (1991). Conservation and enhancement of soil quality. *Evaluation for Sustainable Land Management in the Developing World 2*.

Lim, T.-S., W.-Y. Loh, and Y.-S. Shih (2000). A comparison of prediction accuracy, complexity, and training time of thirty-three old and new classification algorithms. *Machine Learning 40*(3), 203–228.

Liu, R. Y., P. J. M. and K. Singh (1999). Multivariate analysis by data depth: descriptive statistics, graphics and inference. *Annals of Statistics*, 783–858.

Locantore, N., J. S. Marron, D. G. Simpson, N. Tripoli, J. T. Zhang, K. L. Cohen, G. Boente, R. Fraiman, B. Brumback, C. Croux, J. Fan, A. Kneip, J. I. Marden, D. Peña, J. Prieto, J. O. Ramsay, M. J. Valderrama, A. M. Aguilera, N. Locantore, J. S. Marron, D. G. Simpson, N. Tripoli, J. T. Zhang, and K. L. Cohen (1999). Robust principal component analysis for functional data. *Test 8*(1), 1–73.

Lopez-Pintado, S. and J. Romo (2009). On the concept of depth for functional data. *Journal of the American Statistical Association*, 718–734.

Magnano, L., J. Boland, and R. Hyndman (2008). Generation of synthetic sequences of half-hourly temperature. *Environmetrics 19*(8), 818–835.

Martens, H. and T. Naes (1990). Multivariate calibration. *Journal of Chemometrics 4*(6), 441–441.

Martínez, A. M. and A. C. Kak (2001). PCA versus LDA. *IEEE Trans. Pattern Anal. Mach. Intell. 23*(2), 228–233.

McDowell, M., G. Bruland, J. Deenik, S. Grunwald, and N. Knox (2012).

Soil total carbon analysis in Hawaiian soils with visible, near-infrared and mid-infrared diffuse reflectance spectroscopy. *Geoderma 189190*, 312 – 320.

Merry, R. and L. Janik (2001). Mid-infrared spectroscopy for rapid and cheap analysis of soils. *10th Australian Agronomy Conference*.

Mevik, B.H., W. R. (2007). *The pls Package: Principal Component and Partial Least Squares Regression in R.*

Meyer, M. (2014). *Function-on-Function Regression with Public Health Applications. Doctoral Dissertation.* Ph. D. thesis, Harvard University.

Mooney, S., J. Antle, S. Capalbo, and K. Paustian (2004). Influence of project scale and carbon variability on the costs of measuring soil carbon credits. *Environmental Management 33*(1), S252–S263.

Niazi, N. K., B. Singh, and B. Minasny (2015). Mid-infrared spectroscopy and partial least-squares regression to estimate soil arsenic at a highly variable arsenic-contaminated site. *International Journal of Environmental Science and Technology 12*(6), 1965–1974.

Nocita, M., A. Stevens, C. Noon, and B. Van Wesemael (2013). Prediction of soil organic carbon for different levels of soil moisture using Vis-NIR spectroscopy. *Geoderma 199*, 37–42.

Nocita, M., A. Stevens, B. Van Wesemael, M. Aitkenhead, M. Bachmann, B. Barths, E. Ben Dor, D. Brown, M. Clairotte, A. Csorba, P. Dardenne, J. Dematte, V. Genot, C. Guerrera, M. Knadel, L. Montanarella, C. Noon, L. Ramirez-Lopez, J. Robertson, H. Sakai, J. Soriano-Disla, K. Shepherd, B. Stenberg, E. Towett, R. Vargas, and J. Wetterlind (2015). Soil spectroscopy: an alternative to wet chemistry for soil monitoring. *Advances in Agronomy 132*, 139–159.

Pearson, K. (1901). On lines and planes of closest fit to systems of points in space. *Philosophical Magazine*, 559–572.

Ramsay, J.O., H. G. and S. Graves (2009). *Functional Data Analysis in R and Matlab.* New York, NY, USA: Springer Series in Statistics, Springer.

Ramsay, J. O. and C. J. Dalzell (1991). Some tools for functional data analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 539–572.

Ramsay, J. O. and B. W. Silverman (1997). *Functional Data Analysis* (1st ed.). New York, NY, USA: Springer Series in Statistics, Springer.

Ramsay, J. O. and B. W. Silverman (2002). *Applied Functional Data Analysis: Methods and Case Studies* (1st ed.). New York, NY, USA: Springer Series in Statistics, Springer.

Ramsay, J. O. and B. W. Silverman (2005). *Functional Data Analysis* (2nd ed.). New York, NY, USA: Springer Series in Statistics, Springer.

Ramsay, J. O. and B. W. Silverman (2012). *fda: Functional Data Analysis (fda).*

Rao, C. R. (1948). The utilization of multiple measurements in problems of biological classification. *Journal of the Royal Statistical Society. Series B (Methodological) 10*(2), 159–203.

Reeves, J., G. McCarty, and V. Reeves (2001). Mid-infrared diffuse reflectance spectroscopy for the quantitative analysis of agricultural soils. *Journal of Agricultural and Food Chemistry 49*(2), 766–772. PMID: 11262026.

Rossel, R. V., T. Behrens, E. Ben-Dor, D. Brown, J. Dematte, K. Shepherd, Z. Shi, B. Stenberg, A. Stevens, V. Adamchuk, H. Aichi, B. Barthes, H. Bartholomeus, A. Bayer, M. Bernoux, K. Bottcher, L. Brodsky, C. Du, A. Chappell, Y. Fouad, V. Genot, C. Gomez, S. Grunwald, A. Gubler, C. Guerrero, C. Hedley, M. Knadel, H. Morras, M. Nocita, L. Ramirez-Lopez, P. Roudier, E. R. Campos, P. Sanborn, V. Sellitto, K. Sudduth, B. Rawlins, C. Walter, L. Winowiecki, S. Hong, and W. Ji (2016). A global spectral library to characterize the world's soil. *Earth-Science Reviews 155*, 198 – 230.

Rossi, F. and N. Villa (2006). Support vector machine for functional data classification. *Neurocomputing 69*(79), 730–742.

Schoenberg, I. J. (1946). Contributions to the problem of approximation of equidistant data by analytic functions. *Quarterly of Applied Mathematics 4*(1), 45–99.

Schoenberg, I. J. (1964). Spline interpolation and best quadrature formulae. *Bulletin of the American Mathematical Society 70*(1), 143–148.

Shivaswamy, P.K., C. W. J. M. (2007). A support vector approach to censored targets. *CA: IEEE Computer Society*, 655–660.

Silverman, B. (1985). Some aspects of the spline smoothing approach to nonparametric regression curve fitting. *Journal of the Royal Statistical Society*, 1–52.

Soriano-Disla, J. M., L. J. Janik, R. A. V. Rossel, L. M. Macdonald, and M. J. McLaughlin (2014). The performance of visible, near-, and mid-infrared reflectance spectroscopy for prediction of soil physical, chemical, and biological properties. *Applied Spectroscopy Reviews 49*(2), 139–186.

Stenberg, B., R. Viscarra Rossel, A. Mouazen, and J. Wetterlind (2010). Visible and near infrared spectroscopy in soil science. *Advances in Agronomy 107*, 163–215.

Stone, M., B. R. J. (1990). Continuum regression: Cross-validated sequentially constructed prediction embracing ordinary least squares, partial least squares and principal components regression. *Journal of the Royal Statistical Society. Series B (Methodological) 52*(2), 237–269.

Sun, Y. and M. Genton (2011). Functional boxplots. *Journal of Computational and Graphical Statistics 20*(2), 316–334.

Tabachnick, B. G. and L. S. Fidell. (2007). *Using multivariate statistics* (5th ed.). Boston: Pearson/Allyn & Bacon.

Tarrio-Saavedra, J., S. Naya, M. Francisco-Fernandez, R. Artiaga, and J. Lopez-Beceiro (2010). Application of functional anova to the study of thermal stability of micronano silica epoxy composites. *Chemometrics and Intelligent Laboratory Systems 105*(1), 114–124.

Thurstone, L. (1947). *Multiple-factor Analysis: A Development and Expansion of The Vectors of Mind.* The University of Chicago Committee on Publications in Biology and Medicine. University of Chicago Press.

Tu, Y.-K., G. Davey Smith, and M. S. Gilthorpe (2011, 04). A new approach to age-period-cohort analysis using partial least squares regression: The trend in blood pressure in the Glasgow alumni cohort. *PLoS ONE 6*(4), 1–9.

Ullah, S. and C. Finch (2013). Applications of functional data analysis: A systematic review. *BMC Medical Research Methodology 13*(1), 43.

Vapnik, V. N. (1995). *The Nature of Statistical Learning Theory.* New York, NY, USA: Springer-Verlag New York, Inc.

Viscarra-Rossel, R., D. Walvoort, A. McBratney, L. Janik, and J. Skjemstad (2006). Visible, near infrared, mid infrared or combined diffuse reflectance spectroscopy for simultaneous assessment of various soil properties. *Geoderma 131*(12), 59 – 75.

Walkley, A. and I. A. Black (1934). An examination of Degtjareff method for determining soil organic matter and a proposed modification of the chromic acid titration method. *Soil. Sci. 37*, 29–38+.

Wetterlind, J., B. Stenberg, and R. A. V. Rossel (2013). *Soil Analysis Using Visible and Near Infrared Spectroscopy.* Totowa, NJ: Humana Press.

Yang, H., B. Kuang, and A. M. Mouazen (2012). Quantitative analysis of soil nitrogen and carbon at a farm scale using visible and near infrared spectroscopy coupled with wavelength reduction. *European Journal of Soil Science 63*(3), 410–420.

Zornoza, R., C. Guerrero, J. Mataix-Solera, K. Scow, V. Arcenegui, and J. Mataix-Beneyto (2008). Near infrared spectroscopy for determination of various physical, chemical and biochemical properties in mediterranean soils. *Soil Biology and Biochemistry 40*(7), 1923 – 1930.