TECHNISCHE UNIVERSITÄT DRESDEN

Fakultät Umweltwissenschaften

Quantitative Mapping of Soil Property Based on Laboratory and Airborne Hyperspectral Data Using Machine Learning

Dissertation zur Erlangung des akademischen Grades

Doktoringenieur (Dr.-Ing.)

vorgelegt von

MSc. Lanfa Liu

Geboren am: 16. August 1989 in: Shandong, China

Gutachter:

Herr Prof. Dr. Manfred F. Buchroithner

Technische Universität Dresden

Herr Prof. Dr. Elmar Csaplovics

Technische Universität Dresden

Herr Ao. Prof. Dr. Hermann Häusler

Universität Wien

Dresden, ... (Tag der Verteidigung)

Erklärung des Promovenden

Die Übereinstimmung dieses Exemplars mit dem Original der Dissertation zum Thema:

„ **Quantitative Mapping of Soil Property Based on Laboratory and Airborne Hyperspectral Data Using Machine Learning** “

wird hiermit bestätigt.

....

Ort, Datum Unterschrift

Abstract

Soil visible and near-infrared spectroscopy provides a non-destructive, rapid and low-cost approach to quantify various soil physical and chemical properties based on their reflectance in the spectral range of 400–2500 nm. With an increasing number of large-scale soil spectral libraries established across the world and new space-borne hyperspectral sensors, there is a need to explore methods to extract informative features from reflectance spectra and produce accurate soil spectroscopic models using machine learning.

Features generated from regional or large-scale soil spectral data play a key role in the quantitative spectroscopic model for soil properties. The Land Use/Land Cover Area Frame Survey (LUCAS) soil library was used to explore PLS-derived components and fractal features generated from soil spectra in this study. The gradient-boosting method performed well when coupled with extracted features on the estimation of several soil properties. Transfer learning based on convolutional neural networks (CNNs) was proposed to make the model developed from laboratory data transferable for airborne hyperspectral data. The soil clay map was successfully derived using HyMap imagery and the fine-tuned CNN model developed from LUCAS mineral soils, as deep learning has the potential to learn transferable features that generalise from the source domain to target domain. The external environmental factors like the presence of vegetation restrain the application of imaging spectroscopy. The reflectance data can be transformed into a vegetation suppressed domain with a force invariance approach, the performance of which was evaluated in an agricultural area using CASI airborne hyperspectral data. However, the relationship between vegetation and acquired spectra is complicated, and more efforts should put on removing the effects of external factors to make the model transferable from one sensor to another.

Kurzfassung

VIS- und NIR-Spektroskopie liefern einen zerstörungsfreien, schnellen und kostengünstigen Ansatz zur Quantifizierung verschiedener bodenphysikalischer und chemischer Eigenschaften auf der Grundlage ihrer Reflexion im Spektralbereich von 400-2500 nm. Mit einer weltweit zunehmenden Zahl großskaliger Bodenspektralbibliotheken und neuen weltraumgestützten Hyperspektralsensoren müssen Methoden erforscht werden, um mithilfe maschineller Lernverfahren informative Merkmale aus Reflexionsspektren zu extrahieren und genaue bodenspektroskopische Modelle zu erstellen.

Merkmale, die aus regionalen oder großräumigen Bodenspektraldaten erzeugt werden, spielen eine Schlüsselrolle im quantitativen spektroskopischen Modell für Bodeneigenschaften. Die Flächenstichprobenerhebung zur Bodennutzung und Bodenbedeckung (LUCAS) wurde verwendet, um PLS-abgeleitete Komponenten und fraktale Merkmale zu erforschen, die aus Bodenspektren in dieser Studie erzeugt wurden. Die Gradientenverstärkungsmethode zeigte gute Ergebnisse, wenn sie mit extrahierten Merkmalen bei der Schätzung mehrerer Bodeneigenschaften kombiniert wurde. Damit das aus den Labordaten entwickelte Modell auf die luftgestützten hyperspektralen Daten übertragbar ist, wurde vorgeschlagen, basierend auf Convolutional Neural Networks (CNNs), ein Transfer Learning zu entwickeln. Die Boden-Ton-Karte wurde mithilfe von HyMap-Bildern und dem aus den LUCAS-Mineralböden verfeinderten CNN-Modell erfolgreich abgeleitet, da Deep Learning das Potenzial hat, übertragbare Merkmale zu lernen, die von der Quelldomäne zur Zieldomäne verallgemeinern. Die äußeren Umweltfaktoren, wie das Vorhandensein von Vegetation, schränken jedoch die Anwendung der Hyperspektralspektroskopie ein. Die Reflexionsdaten können in einen vegetationsunterdrückten Bereich mit einem Force-Invarianz-Ansatz transformiert werden, dessen Leistung in einem Landwirtschaftsgebiet mittels CASI-Luft-Hyperspektraldaten ausgewertet wurde. Allerdings ist die Beziehung zwischen Vegetation und erfassten Spektren sehr kompliziert, und es sollten mehr Anstrengungen unternommen werden, um die

Auswirkungen externer Faktoren, einschließlich der Vegetation auf Bodenspektren, die unter natürlichen Bedingungen gemessen werden, zu beseitigen.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

1D	One Dimension
AE	Autoencoder
BRDF	Bidirectional Reflectance Distribution Function
CEC	Cation Exchange Capacity
CNN	Convolutional Neural Network
CR	Continuum Removal
DAE	Denoising Autoencoder
DL	Deep Learning
EC	Spectral Mixture Analysis
ELM	Extreme Learning Machine
EnMAP	Environmental Mapping and Analysis Program
EVI	Enhanced Vegetation Index
GBDT	Gradient-Boosted Decision Tree
GLM	Generalized Linear Model
HyspIRI	Hyperspectral Infrared Imager
ISRIC	International Soil Reference and Information Centre
LLE	Local Linear Embedding
LUCAS	Land Use/Land Cover Area Frame Survey
MBL	Memory-Based Learner
MNF	Minimum Noise Fraction
N	Total Nitrogen Content
NDVI	Normalized Difference Vegetation Index
NIPALS	Nonlinear Iterative Partial Least Squares
NIR	Near Infrared

PCA	Principal Component Analysis
PLS	Partial Least Squares
R^2	Coefficient of Determination
RBM	Restricted Boltzmann Machines
RF	Random Forest
RMSE	Root Mean Square Error
RPD	Ratio of Performance to Deviation
SAVI	Soil-Adjusted Vegetation Index
SBL	Spectrum-Based Learner
SHALOM	Space-borne Hyperspectral Applicative Land and Ocean Mission
SMA	Spectral Mixture Analysis
SNV	Standard Normal Variate
SOC	Soil Organic Carbon
SVM	Support Vector Machine
SWIR	Short-wave Infrared
TDS	Total Dissolved Solids
VAE	Variational Autoencoder
VIP	Variable Importance in the Projection

Chapter 1

Introduction

1.1 Motivation

Soil is an essential part of the natural environment. It provides a habitat for a wide range of organisms and is responsible for plant growth, decomposition and microbial biomass recycling. It also plays an important role in addressing climate change. However, there are unprecedented pressures on soil from degradation to pollution. To gain a better understanding of soil, effective methods are in need not only to measure and monitor soil physical and chemical properties but also to characterise their variations at spatial and temporal scales. Traditional laboratory technologies are often time-consuming and expensive, and these soil analyses are usually limited to a few samples and lack information on the spatial variability of soil [1]. Soil spectroscopy, as a fast, cost-effective and environmental-friendly technique, has successfully been utilised to retrieve soil properties.

Soil spectroscopy has been established as an analytical technique for decades as a result of the work by K.H. Norris and co-workers [2]. The measured spectra encode information on the inherent composition of soil, which comprises minerals, organic compounds and water. The encoded information is often represented in the spectra as absorptions at specific wavelengths of electromagnetic radiation, which can be used to describe soil both qualitatively and quantitatively. Soil organic carbon, for example, has absorption features near 600, 1700 and 2300 nm. Water has a strong influence on spectra around 1400 and 1900

nm. Absorption features near 2200-2300 nm characterise clay minerals. To produce general and robust calibration models, soil spectral libraries are established at regional, continental and even global scales. Also, there are several satellite hyperspectral sensors that will be launched in the coming years, which have the potential to acquire soil properties covering a large spatial region [3]. Large-area coverage spectral libraries have been exploited in synergy with remotely sensed hyperspectral imagery to map soil organic carbon and clay content [4–6].

However, the information encoded in the spectra is confounded as absorption bands are weak and often overlapping. The complexity of reflectance spectra in the region of Vis-NIR-SWIR (400-2500 nm) makes it difficult to predict properties by physical theories or models [7]. Therefore, multivariate statistical methods like PLS regression are more suitable to link soil spectra with properties measured in the laboratory. Efforts have been put to develop calibration models that are accurate to infer samples having similar soil composition and spectral characteristics as training data. However, there are still few studies related to extracting effective features from reflectance spectra that are crucial to correlating with soil properties.

1.2 Soil spectra from different platforms

Soil Vis-NIR-SWIR spectra can be acquired at points or by imaging mainly from three different platforms [8], as shown in Figure 1.1. Point spectrometers have demonstrated their capability to accurately determine soil properties in the laboratory, where soil samples are well prepared (seized and dried) and measured under a controlled environment. Thus, laboratory spectroscopy yields the most stable model calibrations. Soil proximal sensing provides a way for rapid in-situ monitoring of soils. It can use either portable point spectrometers or imaging spectrometers. When dedicated to precision agriculture, the sensor is often mounted on a tractor [9]. Although the estimation accuracy is lower due to uncontrollable environmental factors in the field, in situ proximal sensing improves the efficiency of soil data collection by avoiding tedious sampling and preparation procedures [10].

In remote sensing domain, imaging spectrometers can be mounted on aircrafts or satellites, which provide a new perspective for adding spatial details to spectral information [11–13]. Hyperspectral remote sensing has a promising future in the field of soil science and has been adopted to quantify soil properties and study soil degradation [16]. With upcoming new generation space-borne hyperspectral sensors, like the Environmental Mapping and Analysis Program (EnMAP) from Germany, the Hyperspectral Infrared Imager (HyspIRI) from the USA, PRecursore IperSpettrale della Missione Applicativa (PRISMA) from Italy and Space-borne Hyperspectral Applicative Land and Ocean Mission (SHALOM) from the cooperation between Italy and Israel, imaging spectroscopy provides the opportunity to map soil properties at regional and global scales at comparatively low costs. However, the application of hyperspectral remote sensing to the field of soil analysis is restricted by external environmental factors, including the low signal-to-noise ratio, vegetation coverage, atmosphere and BRDF effects.

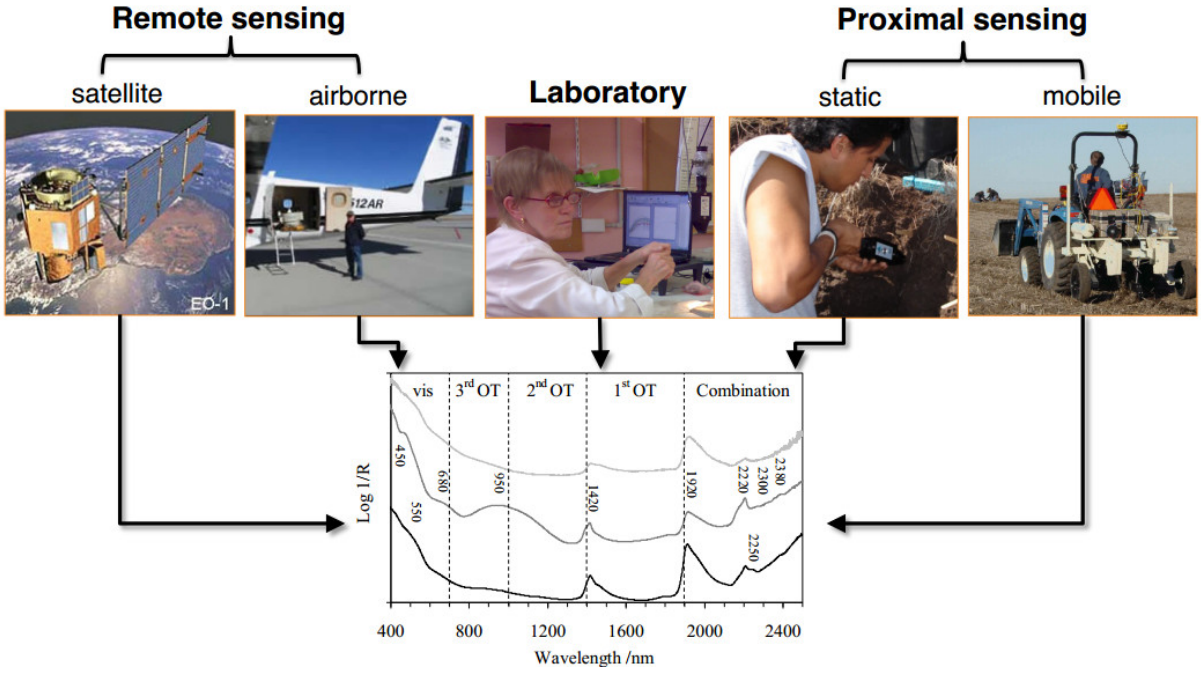

Figure 1.1 Soil spectra measured from different platforms [8]

The soil spectral library can be used as a reference for predicting soil properties by reflectance spectroscopy. Calibrations are not reliable for soil samples not represented in the soil spectral library. Hence there is a need for building libraries representative of the soil diversity [17,18] and an increasing number of large-area coverage soil spectral libraries established at national, continental and even global scales. The ICRAF-ISRIC world soil

spectral library is composed of 4438 samples from 785 soil profiles distributed in 58 countries from Africa, Asia, Europe, North America, and South America selected from the Soil Information System (ISIS) of the International Soil Reference and Information Centre (ISRIC) archives. Samples were scanned in the spectral range of 350-2500 nm. A voluntary collaborative project was started in 2008 to develop a global library of soil spectra, and 23,631 soil spectra have been contributed to the global database by around 45 soil scientists and researchers from 35 institutions [8]. A European spectral library is established within the LUCAS program, which is an extensive and regular topsoil survey that carried out across the EU to derive policy-relevant statistics on the effect of land management on soil characteristics. There will be a new LUCAS sampling campaign undertaken in 2018 [19]. In addition, a number of national and regional soil spectral libraries have been constructed, such as the ones for Australia [20], Czech Republic [21], Brazil [22] and China [10].

1.3 Soil property quantification using spectral data

The complexity of soil prevents a straightforward prediction of reflectance properties by physical theories or models [7]. Therefore, empirical statistical methods are often adopted to relate various properties with soil spectra. Methods including partial least squares (PLS) regression, support vector machine (SVM), extreme learning machine (ELM) and random forest (RF) have been used to derive chemical/physical information from the soil spectra [23–25]. For large-area coverage soil spectral data, soil properties associated to spectrally active constituents cannot be expected to be globally stable [26]. Therefore, it is suggested that local models or memory-based learner (MBL) approaches are suitable for large-scale spectral data instead of global models. The key aim in MBL is not to directly achieve a general or global target function. Instead, when an explanation for a new problem is required, experience in the form of a set of similar related samples is regained from memory. Then, those samples are merged to build the solution to the new problem [27]. A spectrum-based learner (SBL) is further developed based on MBL [24], which selects nearest neighbours from a soil spectral library using distance metrics calculated in the principal component space and optimising the number of components used to identify the nearest neighbours in the selection. Deep learning (DL), as a new area of machine learning research, has also attracted attention from soil

community. Several different deep learning approaches including Restricted Boltzmann Machines (RBM) and one-dimensional convolutional neural networks (1D-CNNs) have been explored for soil property prediction, and 1D-CNN demonstrated to be an effective model for deriving soil properties from high-dimensional spectra [28].

1.4 Feature extraction for soil spectra

Extracting informative and discriminating features is a key component of machine learning and a crucial step for effective soil spectroscopic algorithms. Spectral indices can be viewed as common simple spectral features. It has been widely used for vegetation studies, and more than 100 vegetation indices have been developed such as Ratio Vegetation Index (RVI), Normalized Difference Vegetation Index (NDVI), Enhanced Vegetation Index (EVI) and Soil-Adjusted Vegetation Index (SAVI) [29]. Spectral indices are also proposed to study soil properties. SOC indices based on the summed reflectance and slope in several spectral regions in the Vis-NIR-SWIR were evaluated in [30]. The index developed using the slope of 400-600 nm showed the best performance. For soil clay content, a SWIR Fine particles Index (SWIR FI) was developed based on visual colour indices and absorption peak and shoulders of the absorption feature near 2200-2300 nm [31]. Furthermore, there are several soil indices algorithms implemented in the software of ENSOMAP, which is an open source tool for quantitative soil properties mapping.

Soil Vis-NIR-SWIR spectra are high-dimensional data containing several hundred or thousand bands. Feature extraction is to map the original data to a lower dimensional space without losing significant information to avoid the curse of dimensionality or Hughes phenomenon [32–34]. PCA is commonly used to project raw reflectance spectra to fewer components that describe a large proportion of the variance [30,31]. It is a linear method and reduced dimension representation is generated by linear projections. Several nonlinear PCA, such as kernel PCA and probabilistic PCA, have been proposed to extend the capability of PCA. Manifold learning attempts to model the manifold on which the data lies [34,35]. Local linear embedding (LLE) has been exploited for soil spectral distance and similarity [39]. It can identify the underlying structure of a manifold, while PCA maps faraway data points to nearby points in the plane.

Autoencoder (AE) is an unsupervised learning algorithm and its performance on learning latent representations of soil spectra has been few studied. It is proposed based on neural networks and a basic architecture is shown in Figure 1.2. The AE's codings often reveal useful features from unsupervised data and are useful as dimensionality reduction or feature extraction [40]. AE trains a neural network by constraining the outputs to be equal to the inputs. Thus, the training data do not need to be labelled. By reducing the size of the adjacent layer, the AE is forced to learn a compact representation of the data, in which means that the AE maps the input through an encoder function to generate a latent representation. Ideally, features learned by AE can well represent the input data [41]. There are several approaches proposed to learn features based on AE, like Denoising Autoencoder (DAE), Sparse Autoencoder (SAE), and Variational Autoencoder (VAE).

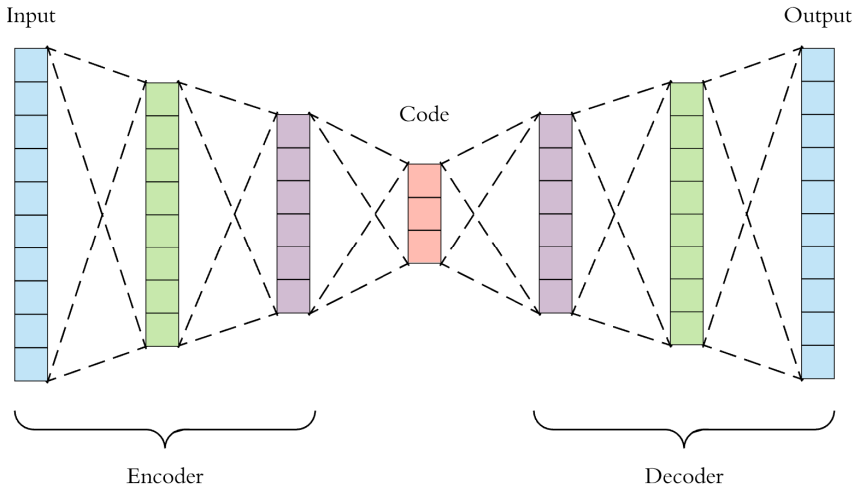

Figure 1.2 The architecture of AE [42]

1.5 Objectives

The main objective of this thesis is to explore feature representation of large-scale soil Vis-NIR-SWIR spectra and its contribution to quantitative soil spectroscopic models. Furthermore, it is intended to use deep learning for quantitative mapping of soil properties by taking advantage of models developed by existing large-scale soil spectral libraries.

The specific objectives are:

1) to assess PLS as a feature extraction tool for soil spectra and the performance by integrating with GBDT on the estimation of soil properties.

2) to develop a new approach to extraction informative features from soil spectra based on fractal geometry using variation estimators with different power indices.

3) to explore the potential of deep learning using 1D-CNN for large-scale soil spectral data modelling.

4) to contribute to soil property mapping using hyperspectral imagery and a large-scale spectral library via transferable features.

1.6 Thesis structure

The main parts of the thesis were prepared as stand-alone manuscripts and published or ready to be submitted to international peer-reviewed journals. The stand-alone manuscripts were written originally by the author of this thesis and subsequently revised by the co-authors. Data collections were carried out by third parties and were identified within the thesis at the appropriate locations. As each of the manuscripts follows the standard structure for a scientific publication, some limited materials are repeated throughout the thesis. The contents of the three published articles have remained unchanged in this thesis and are listed as follows:

Liu, L.; Ji, M.; Buchroithner, M. Combining partial least squares and the gradient-boosting method for soil property retrieval using visible near-infrared shortwave infrared spectra. Remote Sensing. 2017, 9, 1299.

Liu, L.; Ji, M.; Dong, Y.; Zhang, R.; Buchroithner, M. Quantitative retrieval of organic soil properties from visible near-infrared shortwave infrared (Vis-NIR-SWIR) spectroscopy using fractal-based feature extraction. Remote Sensing. 2016, 8, 1035.

Liu, L.; Ji, M.; Buchroithner, M. A case study of Forced Invariance Approach for soil salinity estimation in vegetation-covered terrain using airborne hyperspectral imagery. ISPRS International Journal of Geo-Information. 2018, 7(2), 48.

Thus, the present thesis is divided into the following chapters.

Chapter 1 provides the motivation, objectives and structure of the thesis and related work in soil spectroscopy, including the use of soil Vis-NIR-SWIR spectra for quantifying soil properties and a brief survey of feature extraction or feature representation methods for soil spectra.

Chapter 2 explores PLS components retained from high-dimensional spectral for soil property quantification with the GBDT method. The relative important variables for soil property estimation are also evaluated.

Chapter 3 proposes a novel methodology for soil spectral feature extraction based on fractal geometry. Three variation estimators (rodogram, madogram and variogram) are compared and the effect of step–window sizes on generated fractal features is studied using a grid-search approach. Generated features are compared to PCA-transformed components, and finally these two kinds of features are combined to quantify soil properties using a gradient-boosting regression method.

Chapter 4 presents the potential of transfer learning for soil spectroscopy and its performance on soil clay mapping using hyperspectral data. A 1D-CNN model is developed using LUCAS mineral soils. Its transferability is compared with a clay spectral index using LUCAS organic soils. Then, the 1D-CNN model is fine-tuned and applied to the hyperspectral imagery obtained in the study area.

Chapter 5 demonstrates the Forced Invariance approach for vegetation suppression using hyperspectral data. The performance on improving soil salinity estimation is evaluated.

Finally, Chapter 6 summarises the contributions of the thesis and also discusses the future work as how to further improve the proposed approaches and beyond.

Chapter 2

Combining Partial Least Squares and the Gradient-Boosting Method for Soil Property Retrieval Using Visible Near-Infrared Shortwave Infrared Spectra

Lanfa Liu, Min Ji and Manfred Buchroithner

Contributions:

Lanfa Liu conceived and performed the research and wrote the manuscript.

Min Ji contributed to the design of the research and data analysis.

Manfred Buchroithner reviewed the manuscript and supervised the study at all stages.

Citation:

Liu, L.; Ji, M.; Buchroithner, M. Combining Partial Least Squares and the Gradient-Boosting Method for Soil Property Retrieval Using Visible Near-Infrared Shortwave Infrared Spectra. *Remote Sens.* 2017, *9*, 1299.

2.1 Abstract

Soil spectroscopy has experienced a tremendous increase in soil property characterisation and can be used not only in the laboratory but also from the space (imaging spectroscopy). Partial least squares (PLS) regression is one of the most common approaches for the calibration of soil properties using soil spectra. Besides functioning as a calibration method, PLS can also be used as a dimension reduction tool, which has scarcely been studied in soil spectroscopy. In this study, PLS components retained from high-dimensional spectral data were further explored with the gradient-boosted decision tree (GBDT) method. Three soil sample categories were extracted from the Land Use/Land Cover Area Frame Survey (LUCAS) soil library according to the type of land cover (woodland, grassland, and cropland). First, PLS regression and GBDT were separately applied to build the spectroscopic models for soil organic carbon (SOC), total nitrogen content (N) and clay for each soil category. Then, PLS-derived components were used as input variables for the GBDT model. The results demonstrate that the combined PLS-GBDT approach has better performance than PLS or GBDT alone. The relative important variables for soil property estimation revealed by the proposed method demonstrated that the PLS method is a useful dimension reduction tool for soil spectra to retain target-related information.

2.2 Introduction

Monitoring the status of soil is very important for tackling many challenges including food security, climate change, land degradation, and biodiversity [3]. Traditional laboratory technologies to analyse soil are often time-consuming and expensive and these soil analyses are usually limited to a few samples and lack information on the spatial variability of soil [1]. Soil spectroscopy, as a fast, cost-effective, and environmental-friendly analytical technique, has successfully been utilised to retrieve soil properties and has experienced a tremendous increase in the past years. It has been shown that soil spectra across the Visible Near-Infrared Shortwave Infrared (VIS–NIR–SWIR; 400–2500 nm) spectral region are characterised by significant spectral signals [8,16,43,44], which makes it possible for quantitative analysis of soil properties. Furthermore, the widespread use of visible and infrared spectroscopy can

resolve the trade-off between the growing need for large-scale soil information and its high cost [17]. Using spectral measurements and corresponding soil properties measured by soil analyses, soil spectroscopy can be adopted to quantitatively estimate many soil properties, such as organic matter, heavy metals, clay content, exchangeable potassium, and electrical conductivity [45–47].

The multivariate analysis technique is vital for quantitative analysis of soils. Partial least squares (PLS) regression is frequently used for spectroscopic data and demonstrates the good capability for the estimation of soil properties. The PLS regression method can relate the response variable with relevant information from the spectra while keeping fewer PLS components or factors. It has been successfully demonstrated that the use of soil spectroscopy and PLS regression can quantify soil properties, and an automatic modelling engine PARACUDA®, including PLS, was developed to predict various soil properties using reflectance data [48,49]. PARACUDA® was proposed based on the all-possibilities-approach (APA) concept and a covariate optimisation routine was adopted to select the best pre-processing steps (1st and 2nd derivatives, continuum removal (CR), standard normal variate (SNV), etc.) [50]. Besides PARACUDA®, various calibration methods were also developed based on PLS. The autoPLSR method was proposed to save the need for manual fine-tuning and provided a non-expert, automatic, feature, and latent variable selection, and it was successfully applied for soil clay and iron quantitative mapping using airborne hyperspectral data [15]. The focus of PLS regression is to find the relevant linear subspace of the latent variables, and it has not implemented of variable selection, which could be done based on the selectivity ratio or variable importance in the projection (VIP) before developing PLS models [51–53]. Another option is to use interval PLS (iPLS), which selects only the important variable intervals for PLS regression [54]. Besides, a genetic algorithm was combined with PLS regression (GA-PLSR) to select the most informative spectral variables and thus to improve the prediction accuracies compared with support vector machine regression (SVMR) [46,55]. A memory-based learning (MBL) method called locally weighted partial least squares regression (LWR) was also developed and compared with multiple linear regression (MLR), multiple regression after principal components compression (MLRPC), and PLS. The highest prediction accuracies for most of the soil attributes evaluated were produced by LWR [56]. PLS regression often performs better on a local scale. Therefore, several different local PLS

modelling approaches were proposed and evaluated for predicting soil attributes using a large soil spectral library across the French territory [57]. MBL is a data-driven approach. It is very flexible and can be easily combined with other approaches. MBL describes the target function as a collection of less complex local stable approximations [24,27]. However, it is pointed out that memory-based methods have drawbacks such as high computational costs, and the similarity measure used for recovering samples from the nearest neighbours fails to fit a global function. A spectrum-based learner (SBL) was proposed based on MBL, which can be described as a local linear Gaussian processing modelling approach combining local distance matrices and spectral features as a source of input variables. SBL is able to produce reliable models using regional and global soil spectral libraries [24].

PLS can also be utilised as a dimension reduction (DR) tool [58–61], which has scarcely been explored in soil spectroscopy. The underlying assumption of PLS is that the observed data is generated by a process that is driven by a small number of latent (not directly observed or measured) variables [62]. The reason why PLS regression can perform better than other well-known regression techniques, such as multiple linear regression and ridge regression, is the stability of components derived from the PLS method [63]. The new components can be viewed as retained variables and act as inputs for many other regression approaches. Gradient-boosted decision trees (GBDT), also known as gradient-boosting machine (GBM) or multiple additive regression trees (MART), is one of the most widely used machine learning algorithms and can be viewed as a gradient-boosting algorithm using the decision tree as the weak learner [64,65]. The GBDT method is an additive classification or regression model consisting of an ensemble of trees. It is highly adaptable and many different loss functions can be used during boosting. However, building an accurate GBDT model is time-consuming and often requires extensive parameter tuning. Hence, A GPU-based approach was proposed to accelerate the speed [66].

The relationship between soil properties and soil spectra is very complicated and has an inherently non-linear nature. The objective of the study is to explore the potential of PLS as a dimension reduction tool for soil spectra and the performance of GBDT on the estimation of soil properties. A European-scale soil spectral library has been developed in the framework of Land Use/Land Cover Area Frame Survey (LUCAS) and contains ~20,000 geo-referenced topsoil samples, which is an ideal dataset to evaluate the performance of the proposed PLS-

GBDT method. Three categories of soil samples were extracted from the LUCAS soil spectral library according to the type of land cover (woodland, cropland, and grassland). For each category, SOC, clay, and N were modelled with the proposed method. The evaluation of variable importance was performed and compared with results obtained from PLS and GBDT models.

2.3 Materials and Methods

2.3.1 The LUCAS soil spectral library

As part of the LUCAS project, approximately 20,000 geo-referenced topsoil samples were collected and analysed in the 25 European Union Member States [67,68]. This is the first attempt to build a consistent soil database, which provides an excellent basis to assess topsoil characteristics across the European Union. A standardised sampling procedure was used to collect around 0.5 kg of topsoil (0–20 cm). The collected soils were sampled from different land covers and can be classified as mineral and organic soils. In this paper, the proposed method was applied to mineral soil samples from woodland, cropland, and grassland, the distribution of which can be seen in Figure 2.1.

The soil spectra were measured using a FOSS XDS Rapid Content Analyser (FOSS NIRSystems Inc., Hilleroed, Denmark), operating in the 400–2500 nm wavelength range with 0.5 nm spectral resolution. Pre-processing included removal of the data at wavelengths of 400–500 nm that showed instrumental artefacts, transformation of absorbance (A) spectra into reflectance ($1/10^{A}$) spectra, continuum removal, Savitzky-Golay Filter with a window size of 51 and 2nd order polynomial, and resampling to contain 200 bands. 13 soil properties were analysed in a central laboratory. Three key soil properties, SOC, N, and clay, were selected as our studied properties. A brief statistical summary of soil properties is listed in Table 2.1.

Table 2.1 Summary statistics of soil properties (SOC, N, and clay) for the three soil categories.

Category	Property	N	Mean	SD	Min	Q25	Q50	Q75	Max
Woodland	SOC (g/kg)	4182	37.3	24.1	0.0	18.8	31.4	50.8	125.8
	N (g/kg)	4182	2.0	1.3	0.0	1.0	1.7	2.6	9.1
	Clay (%)	4182	11.3	10.4	0.0	4.0	7.0	16.0	65.0

Category	Property	N	Mean	SD	Min	Q25	Q50	Q75	Max
Cropland	SOC (g/kg)	8341	17.1	10.9	0.0	10.4	14.4	20.5	160.3
	N (g/kg)	8341	1.6	0.79	0.0	1.1	1.5	1.9	9.5
	Clay (%)	8341	22.1	12.7	1.0	13.0	21.0	30.0	79.0
Grassland	SOC (g/kg)	3957	30.2	19.0	0.0	15.7	25.9	39.2	165.7
	N (g/kg)	3957	2.7	1.5	0.0	1.5	2.3	3.4	13.6
	Clay (%)	3957	19.9	12.4	0.0	11.0	18.0	27.0	79.0

SD: Standard Deviation; Q25: lower quartile; Q50: median; Q75: upper quartile.

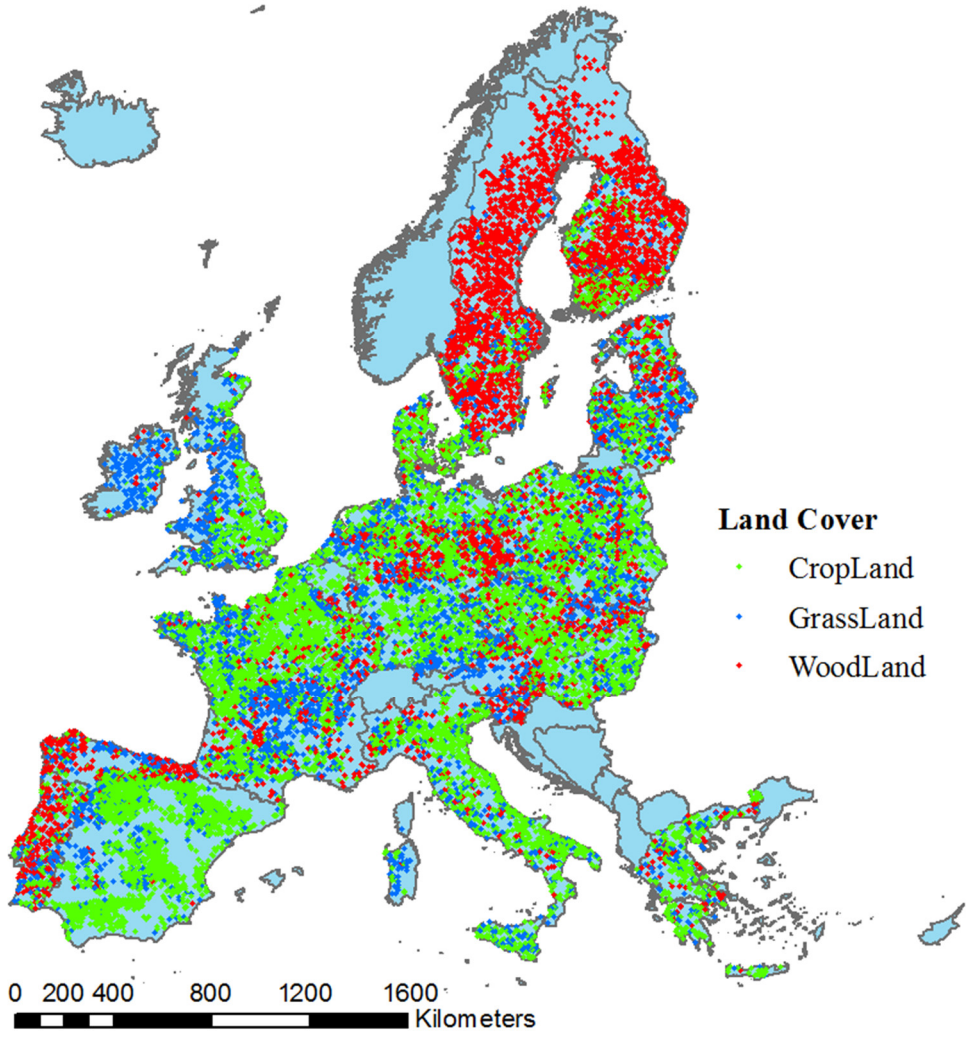

Figure 2.1 Location of selected soil samples from the LUCAS soil spectral library. The colour indicates the corresponding land cover type.

2.3.2 Partial least squares algorithm

PLS regression has proven to be a very successful method for multivariate data analysis. It is a standard tool in chemometrics and has received a great amount of attention in the field of soil spectroscopy. It is similar to principal component regression (PCR), as both can overcome the problems of high dimensionality and multi-collinearity. In its classical form, the PLS method is based on the nonlinear iterative partial least squares (NIPALS) algorithm. To calibrate a PLS regression model for each soil property, the optimal number of latent variables was identified by performing a 10-fold cross-validation, and the root-mean-square error (RMSE) in the cross-validation was used as a decision criterion. Besides directly applying PLS regression to soil spectra, the transformed PLS components were also used as inputs for the following gradient-boosting model.

2.3.3 Gradient-boosted decision trees (GBDT)

Gradient-boosting is a machine learning technique for regression and classification problems, which was developed by Jerome Friedman [69,70]. One of the widely used gradient-boosting methods is GBDT, which is highly adaptable and able to model feature interactions and inherently perform feature selection [71]. These features have made GBDT one of the most widely used machine learning algorithms. Gradient-boosting develops an ensemble of tree-based models by training each of the trees in a sequential manner. Each iteration fits a decision to the residuals left by the previous one, and then the prediction is accomplished by combining the trees. It can produce robust and interpretable procedures for both regression and classification. Mathematically, the model can be viewed as:

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i), \quad f_k \in F \tag{2.1}$$

where k is the number of trees, f is a function in the functional space F, and F is the set of all possible regression trees.

There are several open-source projects that have implemented GBDT, like scikit-learn, XGBoost, and LightGBM [30,38,39]. LightGBM [74] is used in this study, and it is developed by Microsoft. It takes advantage of histogram-based algorithms to accelerate training process and reduce memory consumption by aggregating continuous features into discrete bins [75]. Most decision tree learning algorithms grow trees by the level-wise or depth-wise approach,

as shown in Figure 2.2; LightGBM grows trees by the leaf-wise or best-first approach. It will choose the leaf with max delta loss to grow. When growing the same number of leaves, the leaf-wise algorithm can reduce more loss than a level-wise algorithm. LightGBM also supports parallel and GPU learning, and it is capable of handling large-scale data.

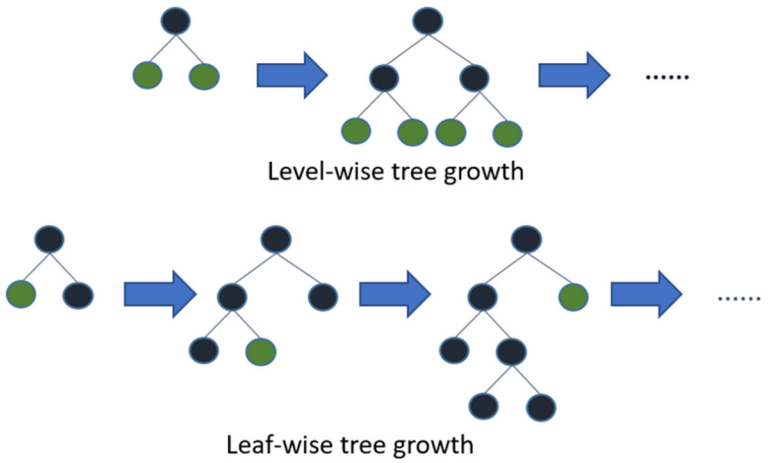

Figure 2.2 Illustration of level-wise and leaf-wise tree growth approaches for gradient-boosted decision trees [74].

Soil spectra quantitatively correlate with soil properties. By fitting a regression model, it is supposed to achieve a good predictive accuracy for the estimation of soil property. There are many parameters that need to be tuned in GBDT, like learning rate or shrinkage, max depth, number of trees, etc. Reducing the learning rate parameter helps prevent overfitting and has a smoothing effect but increases the learning time [65]. The learning rate was set to 0.05. Parameters of max depth and number of trees can also determine whether the model is over-fitted or not, and these two parameters were explored using a grid search strategy.

2.3.4 Calculation of relative variable importance

PLS regression and the gradient-boosting method both can estimate the relative contribution of each input variable or feature. The resultant variable importance measure is useful for understanding the relevance of contributing wavelengths. Ranking based on relative contribution values can help to identify the reflectance bands that are most important for developing soil spectroscopic models. In general, the top few bands contribute most to the model development. For PLS algorithm, the calculation of important input variables is based on weighted sums of the absolute PLS-regression coefficients. A large loading also indicates

the importance of a variable. Here, we use the VIP score derived from coefficients to assess the importance of input variables. It calculates the contribution of independent variables to the contribution of the dependent variable. For the gradient-boosting method, the importance of input variable can be calculated based on metric of "split" or "gain". "Split" is the number of times a variable is used in a model and "gain" is the total gain of splits that use the variable. We use split as the descriptor of relative variable importance in this study. The more a variable is used to make key decisions with decision trees, the higher its relative importance.

2.3.5 Assessment

For each soil property, the soil spectral quantitative model was developed on a random sample of two-thirds of the selected soil samples using PLS regression or the gradient-boosting regression method. The calibrations were tested by predicting the soil properties on validation dataset composed of the remaining one-third samples for each soil category. The model accuracies were evaluated on estimated and measured SOC, N, and clay values using the coefficient of determination (R^2), *RMSE*, and the ratio of performance to deviation (*RPD*) [76].

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2} \tag{2.2}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2} \tag{2.3}$$

$$RPD = \frac{SD}{RMSE} \tag{2.4}$$

where n is the number of validation samples, y is the measured value, $\overline{y}$ is the mean of the measured value, and $\hat{y}$ is the estimated value.

2.4 Results

2.4.1 Overview of the spectral measurement

The mean soil reflectance spectra and standard deviations for soil samples from woodland, cropland, and grassland were plotted in Figure 2.3. The mean spectra of three soil categories have a similar curve shape whose reflectance values increase with increasing

wavelength in the range of 500–1300 nm. Absorption features can be identified near 1400 and 1900 nm, which are assigned to soil hygroscopic water in clay minerals [77]. The mean soil CR spectra and standard deviations for samples from woodland, cropland, and grassland are also shown. CR spectra can be used to isolate and identify characteristic absorptions of minerals, organic compounds, and water in soils [8]. The main spectral difference is that the mean reflectance spectrum for cropland soils demonstrates a higher albedo than spectra for woodland and cropland soils, as cropland soils have a lower mean value of SOC content (17.1 g/kg) than woodland soils (37.3 g/kg) and grassland soils (30.2 g/kg). From the CR spectra, it can be seen that the absorption features are stronger for cropland soils than the other two soil categories, and woodland soils have the weakest absorption features, which can also be explained by the variation of SOC contents. Soil samples with high organic matter content tend to show weak absorption features [24]. Besides, cropland soils have the highest mean value of clay content.

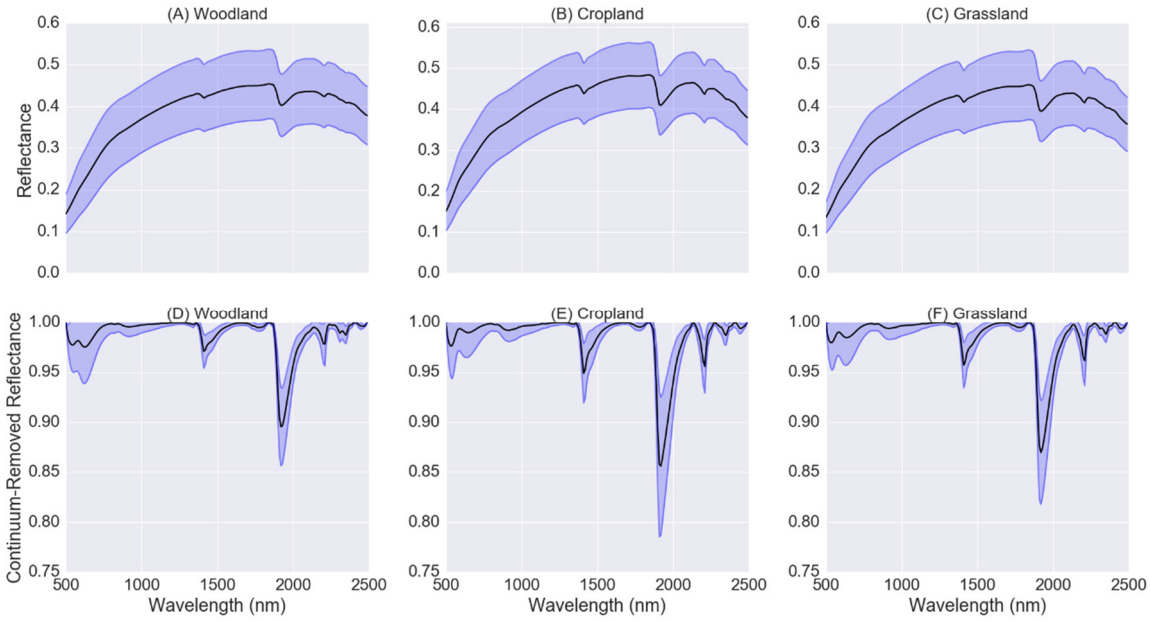

Figure 2.3 (**A**–**C**) are mean soil reflectance spectra (black lines) and standard deviations (blue lines, lower and upper boundaries) for soil samples from woodland, cropland, and grassland; (**D**–**F**) are mean soil continuum-removal spectra (black lines) and standard deviations (blue lines, lower and upper boundaries) for soil samples from woodland, cropland, and grassland. Values are given in reflectance (**A**–**C**) and normalized continuum-removal values (**D**–**F**).

2.4.2 Results of PLS regression for the estimation of soil properties

To make a comparison with the following results obtained from PLS-GBDT, soil spectroscopic models for SOC, N, and clay were developed using PLS regression with the same dataset (Figure 2.4). For each model, the PLS component number was optimised and kept the same as retained by PLS-GBDT (Table 2.2). The accuracies were assessed by R^2, *RMSE*, and *RPD*. Spectroscopic models developed for SOC estimation achieved R^2 values ranging from 0.537 to 0.569 and *RPD* values from 1.51 to 1.57. For N, the highest accuracy (R^2 = 0.652, *RMSE* = 0.78 g/kg, *RPD* = 1.66) was obtained from woodland soils. Models developed for clay estimation achieved comparable good results, and R^2 values vary from 0.656 to 0.732. From *RPD* values, it can be seen that PLS regression can develop fair models for soil spectroscopic analysis that may be used for assessment and correlation.

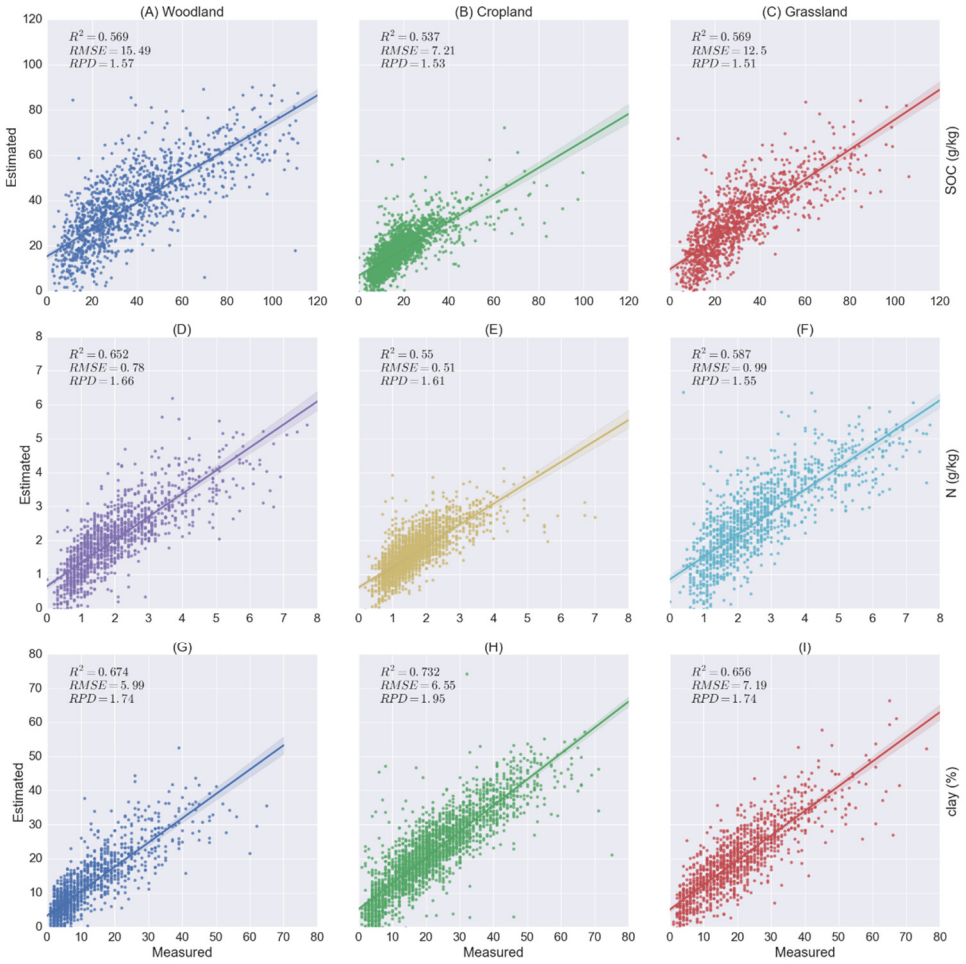

Figure 2.4 Results of soil property estimation accuracies using the partial least squares (PLS) regression method. (**A–C**) are organic carbon accuracies for samples from woodland, cropland, and grassland soils, and (**D–F**) are N accuracies, and (**G–I**) are clay accuracies.

Table 2.2 Optimised parameters for the spectroscopic model using the PLS-gradient-booted decision tree (GBDT) method.

Category	Property	PLS Components	Number of Trees	Maximum Depth
Woodland	SOC (g/kg)	42	300	3
	N (g/kg)	78	1100	4
	Clay (%)	50	500	4
Cropland	SOC (g/kg)	64	1950	4
	N (g/kg)	86	2000	4
	Clay (%)	82	2000	3
Grassland	SOC (g/kg)	60	700	3
	N (g/kg)	72	900	3
	Clay (%)	60	1450	3

Variable selection can be done with PLS. We use VIP scores to rank the relative variable importance. The top 60% variables were kept and further modelled with PLS regression. The results for all three soil categories were shown in Figure 2.5. After variable selection, the accuracy for clay estimation from woodland soils improved with retained variables (R^2 = 0.715, *RMSE* = 5.6 g/kg, *RPD* = 1.86) compared with using full spectrum (R^2 = 0.674, *RMSE* = 5.99 g/kg, *RPD* = 1.74). Variable selection can also increase the SOC estimation accuracy for woodland soils. However, the estimation accuracies for clay from cropland soils and N from woodland soils decreased after variable selection. The R^2 values declined from 0.732 to 0.714 for clay (cropland soils) and 0.652 to 0.636 for N (woodland soils). Soil spectra are complex, especially for large-scale soil spectral data. Soil properties associated with spectrally active constituents cannot be expected to be globally stable [24]. Thus, directly dropping some bands via variable selection may result in a loss of information that is important for some soil samples.

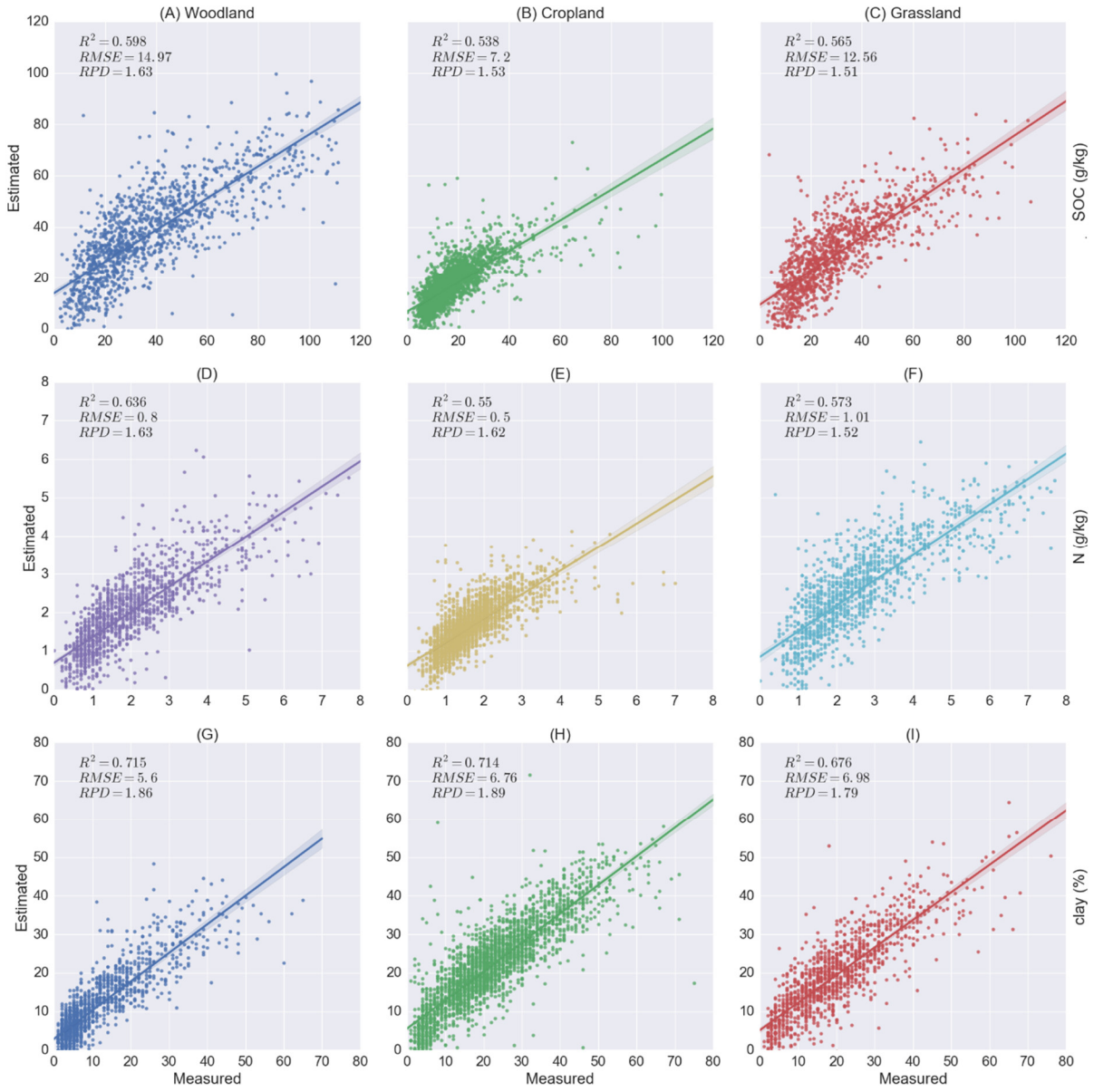

Figure 2.5 Results of soil property estimation accuracies using PLS regression with variable selection using VIP scores. (**A–C**) are organic carbon accuracies for samples from woodland, cropland, and grassland soils, and (**D–F**) are N accuracies, and (**G–I**) are clay accuracies.

2.4.3 Results of PLS-GBDT for the estimation of soil properties

In this study, we propose to transfer soil reflectance spectra data into PLS components to reduce the dimensionality and also decrease the computational complexity. Then, for each category (woodland, cropland, and grassland), soil properties of SOC, N, and clay were modelled using the GBDT method while the input variables were PLS components instead of reflectance spectra. A grid search method was adopted to tune the optimised PLS components

for the first step and also the number of boosted trees and the maximum tree depth for GBDT (Table 2.2).

For SOC, the model built using cropland soil samples achieved the best result (R^2 = 0.679, *RMSE* = 6.0 g/kg, *RPD* = 1.84) compared with soil samples from woodland (R^2 = 0.658, *RMSE* = 13.81 g/kg, *RPD* = 1.76) and grassland (R^2 = 0.671, *RMSE* = 10.92 g/kg, *RPD* = 1.76), which is the same case for the other two soil properties. The spectroscopic model developed from cropland soils has an *RPD* value of 1.94 for N and 2.34 for clay, and both are higher than models developed for woodland soils and grassland soils. This might be due to the complexity of the soil sampling matrix and soil sampling density. From Figure 2.1, it can be seen that cropland soils have the largest proportion of samples because of their ease of access and thus distribute more homogenously compared with woodland soils and grassland soils. The accuracy of clay obtained from the developed PLS-GBDT model has the highest value compared with the other two properties, R^2 values ranging from 0.736 to 0.812 and *RPD* from 1.94 to 2.34.

Compared with Figures 2.4 and 2.6, it can be seen that the results achieved by PLS regression with or without variable selection are worse than by PLS-GBDT. For woodland soils, the R^2 value for SOC reduced from 0.679 to 0.537 and the *RPD* value from 1.84 to 1.53, the R^2 value for N dropped from 0.687 to 0.55, the *RPD* value from 1.94 to 1.61, and the estimation of clay also has the same trend. Therefore, the model developed by non-linear regression method such as PLS-GBDT is suitable for quantitative retrieval of soil properties as reported by [8,78].

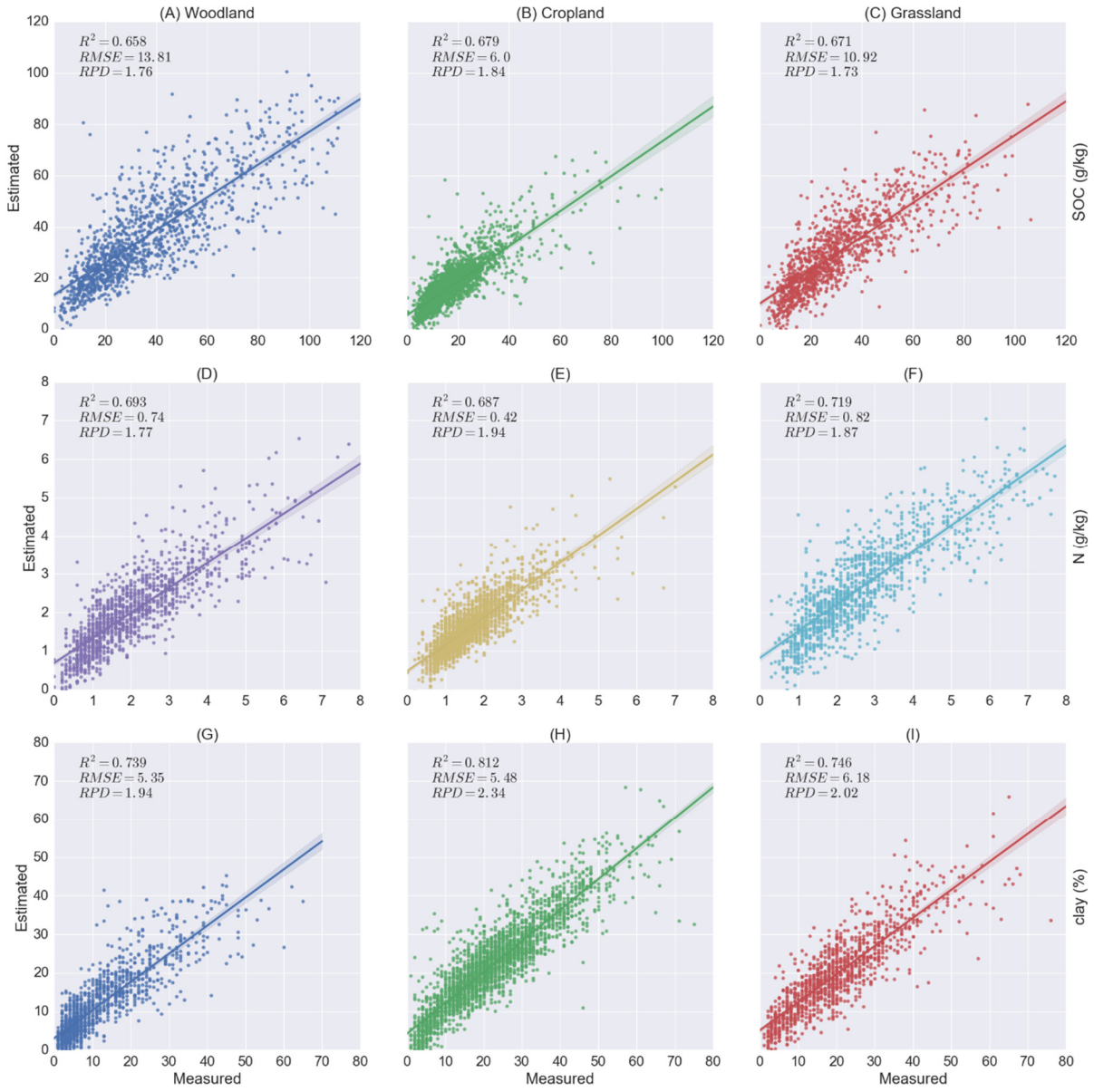

Figure 2.6 Results of soil property estimation accuracies using the PLS gradient-boosting regression method. (**A–C**) are organic carbon accuracies for samples from woodland, cropland, and grassland soils, and (**D–F**) are N accuracies, and (**G–I**) are clay accuracies.

To further evaluate the performance of the PLS-GBDT method, we also directly applied GBDT to soil reflectance spectra. We take samples from woodland soils as an example and use the mean square error (MSE) as the evaluation metric. From Figure 2.7, it can be seen that GBDT model did not perform well, and it is not easy for it to be convergent with the increase of epochs in the training step, as the model tends to be complex when the data dimensionality is too high. PLS-GBDT models achieved much lower MSE values compared with GBDT models, both in the training and validation steps.

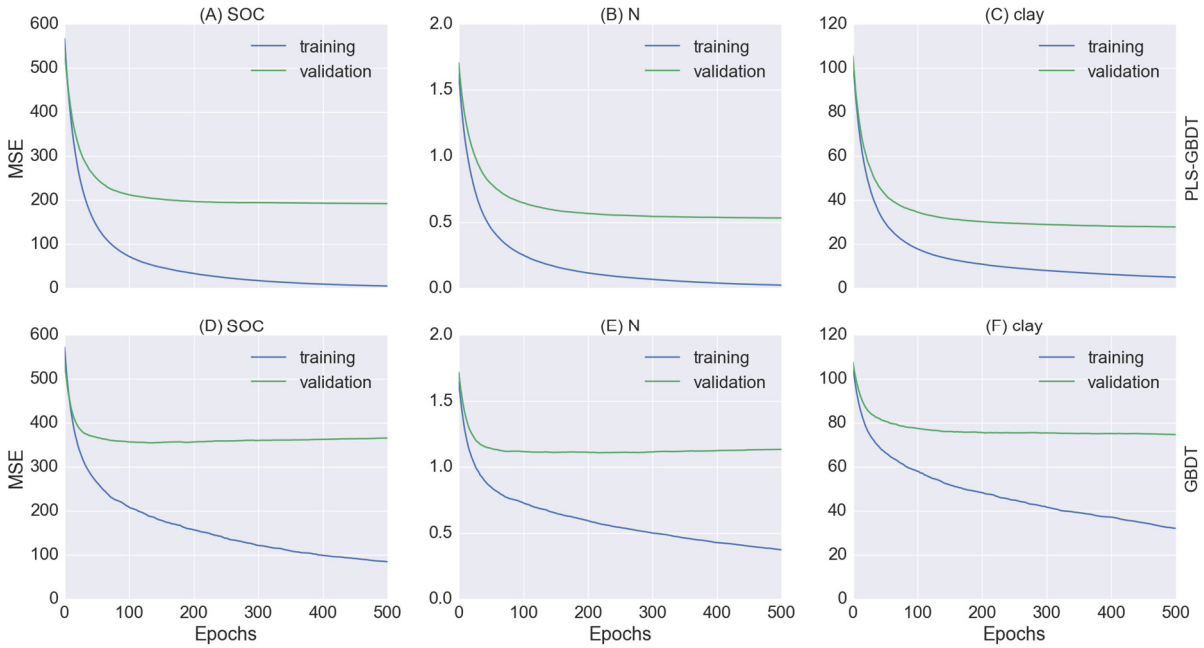

Figure 2.7 Training and validation curves of soil spectroscopic models developed by PLS-GDBT (**A**–**C**) and GBDT (**D**–**F**) in 500 epochs for woodland soil samples.

2.4.4 Relative important variables derived from PLS regression and the gradient-boosting method

A benefit of PLS regression and GBDT is that they can provide the estimation of variable importance from the trained calibration model. As it is time-consuming to tune hyperparameters for GBDT models with very high dimensional data, soil spectra were resampled to 200 bands. The top 13 relative important variables derived from PLS regression models can be seen from Figure 2.8. For SOC, the most important bands for these three soil categories are at 1920, 2170, and 2050 nm. The top-ranked bands for N are similar to SOC (2160, 1940, and 2000 nm). For clay, the derived important variables are at 2070, 1950, and 2230 nm. In previous study [79], the bands near 800, 1000, 1400, and 1900–2450 nm were confirmed to be important for SOC estimation, and the bands around 1100, 1600, 1700 to 1800, 2000, and 2200 to 2400 nm were also identified as key bands for SOC and N estimation [26]. The results are basically in agreement with previous research.

For all of these three soil properties, the top-ranked variables derived from GBDT model were basically at the beginning and the end of the spectrum (Figure 2.9). It can be seen that the GBDT method failed to select meaningful bands for quantitative estimation of SOC, N, and clay when directly using the full spectrum as input variables, which also explained why the accuracy of the GBDT model is worse than the results obtained from PLS and PLS-GBDT

models. Conversely, relative important variables derived from PLS regression are more reasonable.

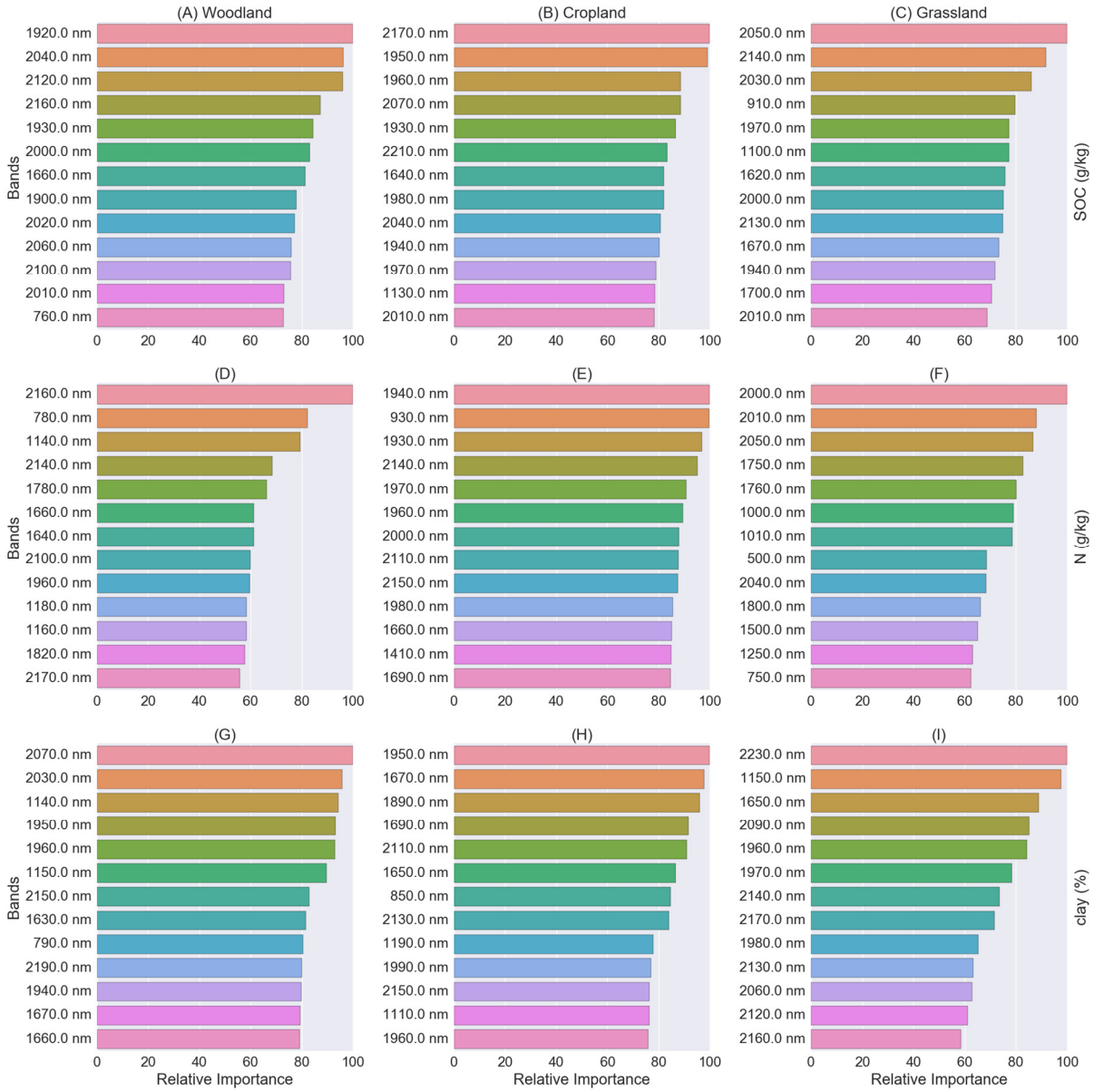

Figure 2.8 Top 13 relative important variables derived from PLS regression models. (**A–C**) are relative important variables derived from SOC models for woodland, cropland, and grassland soils, and (**D–F**) are relative important variables derived from N models, and (**G–I**) are relative important variables derived from clay models.

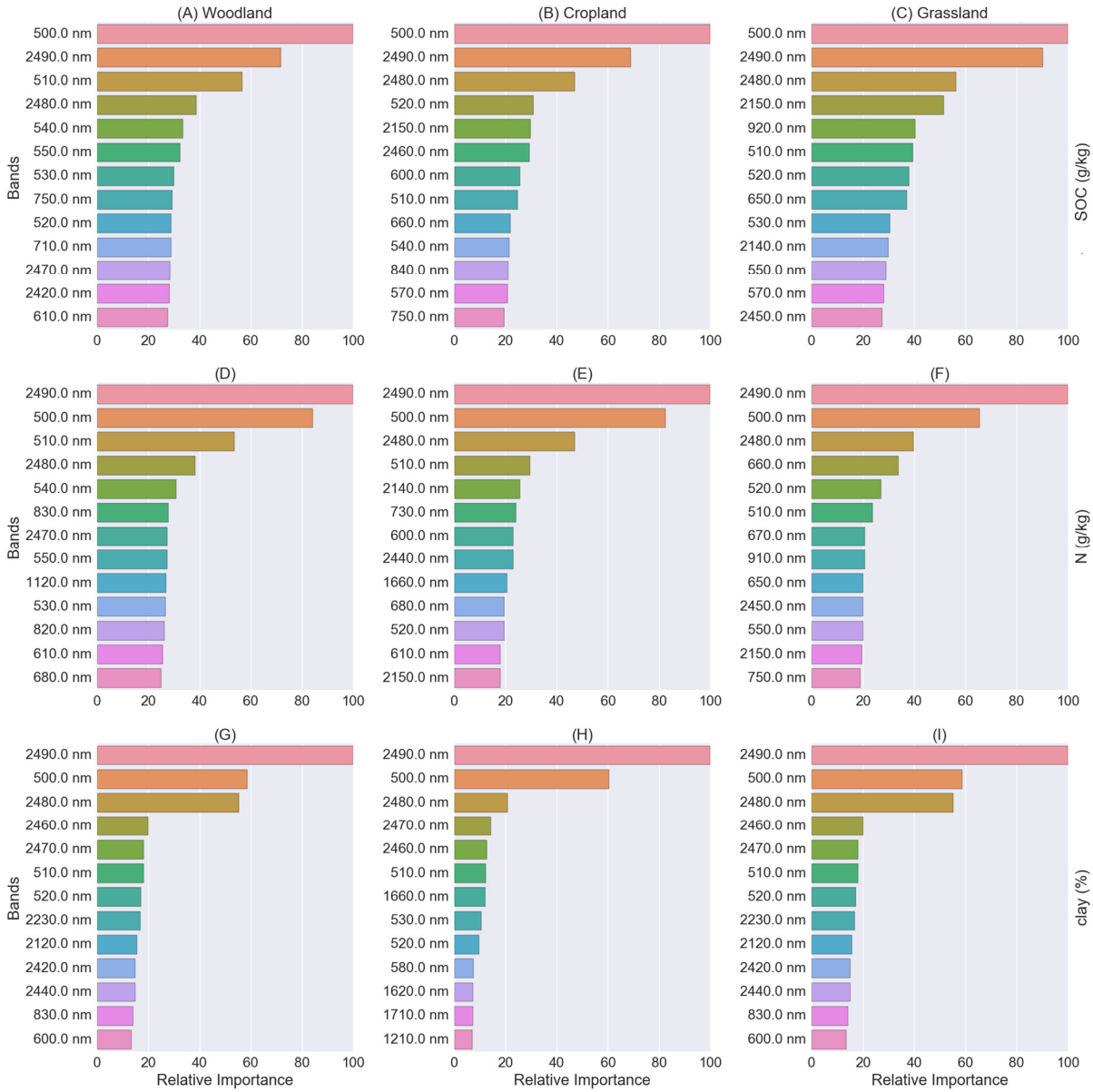

Figure 2.9 Top 13 relative important variables derived from GBDT models. (**A–C**) are relative important variables derived from SOC models for woodland, cropland, and grassland soils, and (**D–F**) are relative important variables derived from N models, and (**G–I**) are relative important variables derived from clay models.

Although the GBDT method failed to derive the relative important variables when using full spectrum, it does not mean that this method is not suitable for soil spectroscopic analysis. The obtained relative important variables were demonstrated (Figure 2.10) when the model was combined with retained PLS components. For PLS-GBDT, the first PLS component is supposed to be the most important variable for the estimation of corresponding soil properties, as PLS retains target-related information. The results demonstrate that the most important

variables are the first PLS component for SOC and N, while the second PLS component is ranked first for clay. In general, the top-ranked PLS components are also important to the gradient-boosting model, as revealed by Figure 2.10. This also means that PLS performs well on the extraction of target-related information.

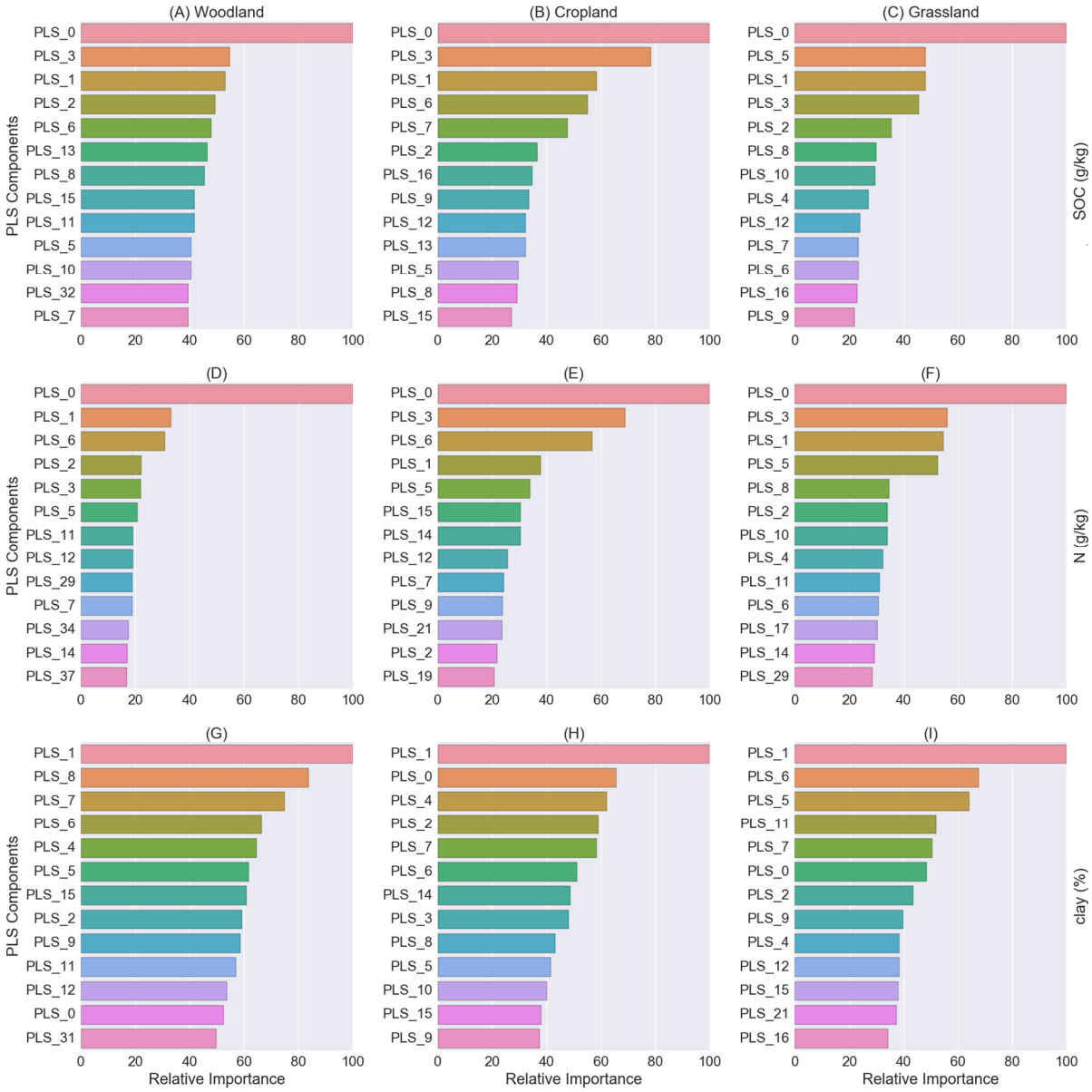

Figure 2.10 The top 13 relative important variables (PLS components) derived from PLS-GBDT. (**A–C**) are relative important variables derived from SOC models for woodland, cropland, and grassland soils, and (**D–F**) are relative important variables derived from N models, and (**G–I**) are relative important variables derived from clay models.

2.5 Discussion

2.5.1 Dimension reduction for high-dimensional soil spectra

High-dimensional data like soil spectra often contain redundant information and will increase computation complexity, which is known as the curse of dimensionality or Hughes phenomenon [32–34]. Variable selection can reduce the complexity and improve the robustness of the model. By selecting the most informative spectral bands instead of using the full spectrum, the calibration model is supposed to be more accurate [80]. Variable selection can be based on physical background by identifying key wavelengths for the target property. It is also possible to evaluate it using the statistics of the resulting calibration model, like the VIP score derived from the PLS regression model in this study.

High-dimensional spectral data can be projected to a lower dimensional space without actually losing significant information using methods like principal component analysis (PCA). PCA reduces the dimensionality of the data to fewer components that describe a large proportion of the variance. The first principal component accounts for the largest variance, while subsequent components account for decreasingly smaller proportions [35]. Local linear embedding (LLE) is a nonlinear dimensionality reduction method, and it can identify the underlying structure of a manifold [38]. PCA and LLE have been exploited in a comparative way for soil spectral distance and similarity in projected space [39]. Autoencoder (AE) is an unsupervised learning algorithm and its performance on reducing the dimensionality of soil spectra has not been well studied. Several approaches were developed based on it, such as stacked autoencoder and sparse autoencoder [81,82]. AE trains a neural network by constraining the output values to be equal to the input values. The reconstruction error between the input and the output is used to adjust the weights of each layer. Ideally, features learned by AE can well represent the input data [41]. The difference between PLS and the above mentioned DR methods is that PLS is able to retain target-related information and can be viewed as a supervised DR method. It has the potential to explore the intrinsic structure of spectra, and it can not only reduce the data redundancy but also improve the estimation accuracy. Besides, it is worth making a comparison between PLS and other mentioned DR methods (PCA, LLE, AE, etc.) for soil spectral analysis in the following studies.

2.5.2 GBDT for quantitative soil spectroscopic modelling

Modelling soil properties using large and diverse soil spectral libraries is still a challenging task. PLS regression, as a common approach in soil spectroscopy, has limitations in handling large-scale soil spectral data. With variable selection using VIP scores, the performance with regard to improving the estimation accuracies is still not satisfying in this study. GBDT has been used to win machine learning competitions on Kaggle and has gained a lot of attention. In this study, we proposed to take advantage of GBDT for the estimation of soil properties by using PLS components as the input variables instead of raw reflectance spectra. The result demonstrated that the combined PLS-GBDT approach performs better than PLS or GBDT alone. It also confirmed the experiments in [83], in which the boosted decision trees method performed exceptionally well when dimensionality was low. The model is prone to being complex when the dimension is too high, and it tends to need more trees and a high degree of tree depth, which could be a serious problem in high dimensions [84]. Therefore, it is suggested to reduce the number of input features via dimension reduction or feature selection when facing high-dimensional data. There are several studies related to soil spectroscopic modelling using large-scale soil spectral libraries. Local or MBL approaches are reported to have better performance on large-scale soil data. PLS, SVM, LWR, and SBL were comparatively studied on a regional soil spectral library in Brazil and a global soil spectral library [24]. SBL algorithm achieved the best performance for SOC estimation in the regional (R^2 = 0.59) and the global data (R^2 = 0.68). MBL approaches are very flexible and can be easily integrated with PLS-GBDT. Besides, additional soil information like texture (sand, clay, and silt) can contribute to soil spectroscopic model. By only using spectral bands as the input variables in [85], SVM obtained a similar result for SOC estimation of cropland soils as achieved by PLS-GBDT. However, the R^2 value improved from 0.67 to 0.71 with variable selection and clay content as an auxiliary variable. A higher accuracy of the SOC estimation model was also obtained by [86] when considering sand content. Therefore, additional soil information is very important to calibration models for large-scale soil spectral data.

2.6 Conclusions

Soil spectra measured in the laboratory typical have several hundred or even thousand bands, which would be a problem for the gradient-boosting model when directly using such high-dimensional data as inputs. This study presents a PLS-GBDT method to retrieve soil properties from reflectance spectra. The LUCAS soil spectral library was used to evaluate its performance. For three soil categories (woodland, grassland, and cropland), R^2 achieved values of 0.658–0.679 for SOC, 0.687–0.719 for N, and 0.739–0.812 for clay. Both PLS and GBDT can estimate the relative contributions of input variables. However, GBDT failed in this task when directly using high-dimensional soil spectra as input data. The GBDT method is a well-known machine learning algorithm that uses the decision tree as the weak learner, and it has successfully been applied in numerous areas. By using PLS components as input variables, which are retained with target variable-related information, GBDT is able to perform well on soil quantitative analysis. Although the PLS-GBDT method is directly used to develop a global model to fit the whole soil spectral library in this study, it is possible to combine it with MBL if it functions as a basic or local model.

Acknowledgments

We acknowledge support by the German Research Foundation and the Open Access Publication Funds of the TU Dresden. The first author wants to acknowledge the China Scholarship Council (CSC) for providing financial support to study at TU Dresden. The LUCAS topsoil dataset in this work was made available by the European Commission through the European Soil Data Centre and managed by the Joint Research Centre (JRC) http://esdac.jrc.europa.edu/.

Chapter 3

Quantitative Retrieval of Organic Soil Properties from Visible Near-Infrared Shortwave Infrared Spectroscopy Using Fractal-Based Feature Extraction

Lanfa Liu, Min Ji, Yunyun Dong, Rongchung Zhang and Manfred Buchroithner

Contributions:

Lanfa Liu conceived, designed and performed the research.

Min Ji, Yunyun Dong, Rongchung Zhang contributed to the analysis of the data.

Lanfa Liu wrote the draft, and Manfred Buchroithner reviewed and edited it.

Citation:

Liu, L.; Ji, M.; Dong, Y.; Zhang, R.; Buchroithner, M. Quantitative Retrieval of Organic Soil Properties from Visible Near-Infrared Shortwave Infrared (Vis-NIR-SWIR) Spectroscopy Using Fractal-Based Feature Extraction. *Remote Sens.* 2016, *8*, 1035.

3.1 Abstract

Visible and near-infrared diffuse reflectance spectroscopy has been demonstrated to be a fast and cheap tool for estimating a large number of chemical and physical soil properties, and effective features extracted from spectra are crucial to correlating with these properties. We adopt a novel methodology for feature extraction of soil spectroscopy based on fractal geometry. The spectrum can be divided into multiple segments with different step–window pairs. For each segmented spectral curve, the fractal dimension value was calculated using variation estimators with power indices 0.5, 1.0 and 2.0. Thus, the fractal feature can be generated by multiplying the fractal dimension value with spectral energy. To assess and compare the performance of new generated features, we took advantage of organic soil samples from the large-scale European Land Use/Land Cover Area Frame Survey (LUCAS). Gradient-boosting regression models built using XGBoost library with soil spectral library were developed to estimate N, pH and soil organic carbon (SOC) contents. Features generated by a variogram estimator performed better than the other two estimators and the principal component analysis (PCA). The estimation results for SOC were coefficient of determination (R^2) = 0.85, root mean square error (*RMSE*) = 56.7 g/kg, the ratio of percent deviation (*RPD*) = 2.59; for pH: R^2 = 0.82, *RMSE* = 0.49 g/kg, *RPD* = 2.31; and for N: R^2 = 0.77, *RMSE* = 3.01 g/kg, *RPD* = 2.09. Even better results could be achieved when fractal features were combined with PCA components. Fractal features generated by the proposed method can improve estimation accuracies of soil properties and simultaneously maintain the original spectral curve shape.

3.2 Introduction

Quantitative assessment of soil properties using visible near-infrared shortwave infrared (Vis-NIR-SWIR) spectroscopy has been demonstrated as a fast and non-destructive method [3,8,26,77,85,87]. Over the past 30 years, numerous soil physical and chemical properties, such as soil texture, soil organic carbon (SOC), cationic exchange capacity (CEC), total nitrogen (N) and exchangeable potassium (K), have been investigated using the spectroscopic approach based on various multivariate statistics and machine learning approaches [1,24,88–90], and outcomes were applied in soil contamination, soil degradation, environmental monitoring

and precision agriculture [77,91–93]. As one of the attractive advantages, soil spectra can be recorded at points or by imaging from different platforms [8,78]. The technique is mainly used in the laboratory, where soil samples are prepared and measured under controlled conditions, and it can be considered as an alternative to traditional analytical techniques. Portable Vis-NIR-SWIR spectrometers allow measurements operated directly in situ. Although the estimation accuracy is lower when compared to results achieved in the laboratory due to uncontrollable environmental factors in the field, in situ proximal sensing improves the efficiency of soil data collection by avoiding tedious sampling and preparation procedures [10]. Sensors can also operate from high above, termed as air- or space-borne imaging spectroscopy [11–13]. However, there are still some limitations with respect to the application of imaging spectroscopy to the field of soil analysis, especially when vegetation is present. They have already shown the potential to map and quantify soil properties [20,21]. With upcoming space-borne sensors, like the Environmental Mapping and Analysis Program (EnMAP) from Germany and the Hyperspectral Infrared Imager (HyspIRI) from the USA, imaging spectroscopy provides the opportunity to map soil properties at regional and global scales at comparatively low costs.

Reflectance spectra of soil can be viewed as cumulative properties that reflect the inherent spectral behaviour of soil components and can be used to quantify these components simultaneously [3]. However, due to the complexity of scattering effects caused by soil structure and specific constituents, the absorption wavelengths are largely overlapping and result in complex absorption patterns [26]. Besides, soil spectra often tend to have a very high dimensionality. For example, each spectrum in the Land Use/Land Cover Area Frame Survey (LUCAS) [67] soil spectral library has 4,200 Vis-NIR-SWIR absorbance measurements, while the Africa Soil Information Service (AfSIS) [94] soil spectra has more than 3000 mid-infrared absorbance measurements. The LUCAS Project aims to sample and analyse the main properties of topsoil across Europe, and the AfSIS Project aims to narrow the sub-Saharan soil information gap and to provide a consistent baseline for monitoring soil ecosystem services. Laboratory spectroscopy was used in both projects. High-dimensional data often contain redundant information and increase computation complexity. It has been proven that most of the data are concentrated in the corners of high dimensional space and the model's accuracy tends to firstly improve and then decline with an increase of features, which is also known as

the curse of dimensionality or Hughes phenomenon [32,33,95]. Therefore, simply relying on different multivariate statistics in raw feature space is not enough, and methods to reduce the dimensionality and extract information from the spectra that can be better correlated with soil properties of interest should be investigated.

Feature extraction has been proved to be successful in imaging-spectroscopy classification [32,41,96–98]. The high-dimensional spectral data can be projected to a lower dimensional space with feature extraction methods, without actually losing significant information. Reduced features may increase the separation between spectrally similar classes and the classification model can perform well with a reduced number of features. In soil spectroscopy, a common approach is principal component analysis (PCA). In [36], PCA was used to reduce the Vis-NIR-SWIR data with more than 2,000 wavelengths to a few components, the first component of which accounting for the largest variance. Also, soil information contents of the spectra consisted of PCA components, and a predictive spatial model was developed across Australia. Effective information can also be extracted with wavelet analysis [99]. It can substantially reduce the factors outside the parameters of the spectrum directly or indirectly. PCA and local linear embedding (LLE) have, in a comparative way, been exploited for soil spectral distance and similarity in projected space [39]. LLE is a nonlinear dimensionality reduction method [34,35]. It can identify the underlying structure of a manifold, while PCA maps faraway data points to nearby points in the plane. The results indicate that the distances computed in the raw space have comparatively lower performance than the ones computed in low reduced spaces. Methods using PCA and LLE with Mahalanobis distance outperformed other approaches. It can be seen that an effective feature extraction method has the potential to explore the intrinsic structure of spectra, and does not only reduce the data redundancy but also improves the estimation accuracy [100].

Knowing how to effectively extract features from the spectra is crucial for a successful soil-spectral quantitative model. Studies focused on feature extraction from soil Vis-NIR-SWIR spectra are still limited. In this paper, we adopt a novel approach of fractal features based on fractal geometry using variation estimators with the different power indices 0.5, 1.0 and 2.0, which can be termed as rodogram, madogram and variogram respectively. The concept of fractal dimension was introduced by [37,38] to reduce the dimensionality of imaging spectroscopy data. Kriti Mukherjee [32,102] proposed a method to generate multiple

fractal-based features from imaging spectroscopy data and then further compared the performance of fractal-based dimensionality reduction using Sevcik's, power spectrum and variogram methods with conventional methods like PCA, minimum noise fraction (MNF), independent component analysis (ICA) and decision boundary feature extraction (DBFE) methods. They concluded that the classification accuracy is similar but the computational complexity is reduced. The aims of the present study are to explore fractal-based feature extraction from soil spectra and to examine its performance on the estimation of SOC, N and pH contents with soil Vis-NIR-SWIR diffuse reflectance spectra. Features generated by the fractal method were compared to PCA-transformed components, and then these two kinds of features were combined to quantify soil properties using a gradient-boosting regression method. The proposed method is further compared to partial least squares (PLS) regression, which is a frequently adopted method for the quantification of soil properties.

3.3 Materials and Methods

3.3.1 The LUCAS topsoil database

As part of Land Use/Land Cover Area Frame Survey, approximately 20,000 geo-referenced topsoil samples were collected and analysed for the 25 European Union member states [67,68]. Stratified random sampling was applied to collect around 0.5 kg of topsoil (0–20 cm) [103]. The LUCAS topsoil dataset is obtained from the Joint Research Centre (JRC) and can be used for non-commercial purposes [67]. The collected samples can be classified as mineral and organic soils. In this paper, the proposed feature extraction method was tested using the LUCAS organic soil samples, the distribution of which was explored in ArcGIS 10.4 and can be seen in Figure 3.1.

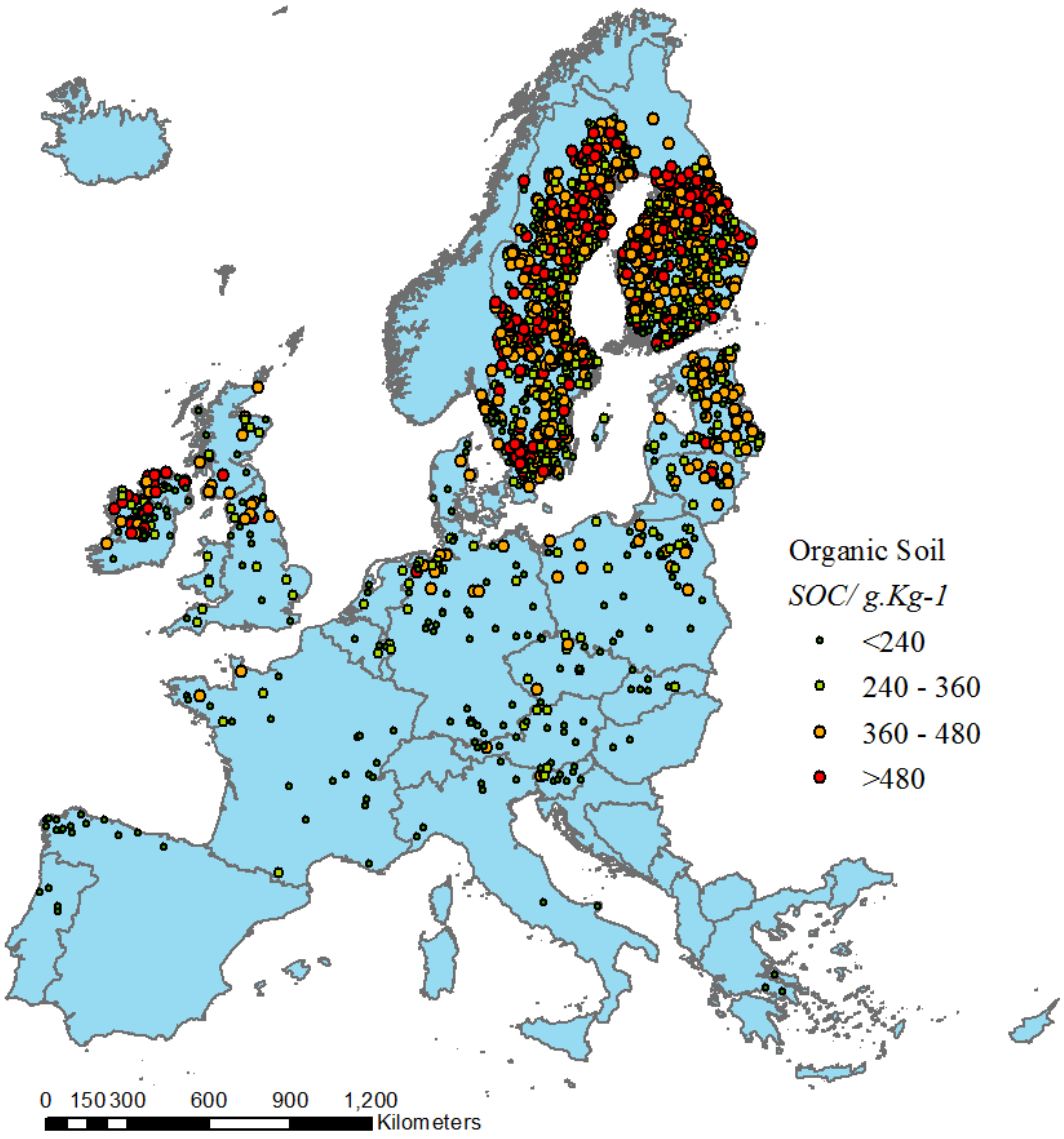

Figure 3.1 Distribution of organic soil samples in the LUCAS topsoil database. Colours indicate amounts of soil organic carbon (SOC) content.

The Vis-NIR-SWIR soil spectra were measured using a FOSS XDS Rapid Content Analyser (FOSS NIRSystems Inc., Denmark) [67], operating in the 400–2500 nm wavelength range, with 0.5 nm spectral resolution. Organic soil spectra were pre-processed by removing the data at wavelengths of 400–500 nm that showed instrumental artefacts, transformation of absorbance (A) spectra into reflectance ($1/10^{A}$) spectra, continuum removal, Savitzky-Golay filter with a window size of 50, second-order polynomial and first derivative. Thirteen soil properties have been analysed in a central laboratory [67], including the percentage of coarse fragments, particle size distribution (% clay, silt and sand content), pH (in $CaCl_2$ and H_2O), soil organic carbon (g/kg), carbonate content (g/kg), phosphorous content (mg/kg), total

nitrogen content (g/kg), extractable potassium content (mg/kg), and cation exchange capacity (cmol(+)/kg). Three key soil fertility properties, soil organic carbon (SOC), total nitrogen content (N) and pH in $CaCl_2$ (pH), were selected as our studied properties.

3.3.2 Fractal feature extraction method

3.3.2.1 Concept of fractal dimension

Fractal dimension is a robust method for describing natural or man-made fractals having the fundamental feature known and referred to as self-similarity [104]. Within the fractal lies another copy of the same fractal, smaller but complete. If we have a strictly self-similar fractal which can be decomposed into N pieces, each of which is a copy of the original fractal scaled by a factor of S, then,

$$S^D = \mathrm{N} \tag{3.1}$$

where D is the Hausdorff Dimension. D is a non-integer number, describing how the irregular structure of objects and/or phenomena is replicated in an iterative way from small to large scales. Anything that appears random and irregular can be a fractal, strictly or statistically, including the soil Vis-NIR-SWIR spectrum, which cannot be defined by any mathematical equation and is therefore considered as an irregular curve. There are numerous methods which have been developed for fractal dimension estimation, including box-count [105], variogram [106], power spectrum [32] and spectral [107] methods.

3.3.2.2 Variation method for fractal dimension

The variogram estimator is widely used in the determination of the fractal dimension and it is known for its ease of use [108]. By sampling a large number of pairs of points along the spectral curve and computing the differences in their reflectance values, the fractal dimension is easily derived from the log-log plot of variogram and lags. X_u and X_{t+u} are two reflectance values located at points u and $t+u$, and these two points are separated by the lag of t. The variogram can be calculated as the mean sum of squares of all differences between pairs of values with a given distance divided by two.

$$\gamma(t) = \frac{1}{2}E(X_u - X_{t+u})^2 \tag{3.2}$$

The variogram estimator is a stochastic process with stationary increments as half times the expectation of the square of an increment at lag t, and a generalisation of the variation estimator can be obtained with different order p of a stochastic process [109]:

$$\gamma_p(t) = \frac{1}{2} E|X_u - X_{t+u}|^p \tag{3.3}$$

where p = 1.0, it represents the madogram, which instead of calculating squares of the differences takes the absolute values. Where p = 1/2, the rodogram is derived by calculating the square root of absolute differences. Fractal dimension is estimated using the slope (θ) of the corresponding log–log regression plot of $\gamma_p(t)$ and t, as shown in Figure 3.2.

$$D = 2 - \frac{\theta}{2} \tag{3.4}$$

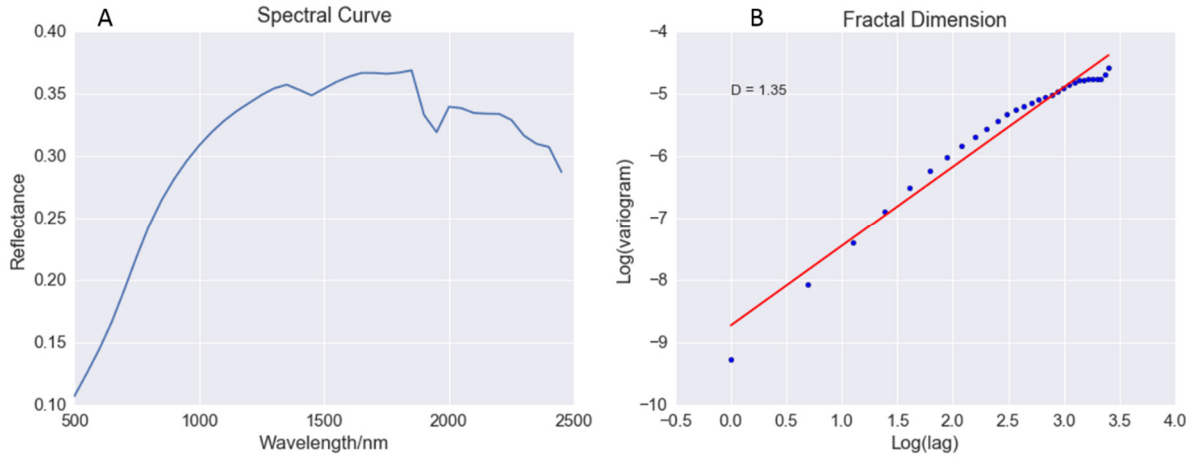

Figure 3.2 Illustration of fractal dimension calculation. (**A**) is the spectral curve and (**B**) is the corresponding log-log plot of variogram and lags and the fitted regression line.

3.3.2.3 Fractal feature generation

Fractal features are generated by multiplying spectral energy with the corresponding fractal dimension. As the fractal dimension can be calculated using the whole curve or only part of the curve, the spectrum can be segmented into several parts and each part corresponds to a new fractal feature. For a soil spectral curve, a common approach is to evenly divide the whole curve into the desired number of segments [110], which means the step and window

size are the same. In this study, we explored the effect of different combinations of step and window sizes on generated fractal features. The final feature number N_f can be calculated as:

$$N_f = \frac{N_r - W}{P} + 1 \tag{3.5}$$

N_r is the number of raw spectral measurements, P is the value of step size and W is the value of moving window size. The window size is often defined as larger than the step size, which means segments of the same spectral curve are overlapping. Step size is defined as 100.0 nm and moving window size as 200.0 nm, as shown in Figure 3.3, which means P = 200 and w = 400 (the spectral resolution is 0.5 nm in our case). New fractal features can be generated when the wavelength window moves along the spectral curve at step 100.0 nm. With the increase of the step size, the final fractal feature number (N_f) correspondingly decreases, which can be used as a means of dimension reduction.

N_f numbers of fractal dimension values can be obtained by moving along the spectral curve at step size p. For each segment, the number of points are marked as n and can be calculated by Equation (3.5). The reflectance value as Z_j (j = 1, 2…, n) and the corresponding fractal dimension value can be calculated according to Equation (3.4) as D_m (m = 1, 2…, Nf), and fractal features by:

$$F_m = \mathrm{D}_m \times \mathrm{E}_m \tag{3.6}$$

where E_m is the spectral energy and can be derived from the following equation:

$$\mathrm{E}_m = \sum_{j=1}^{n} Z_{m,j}^2 \tag{3.7}$$

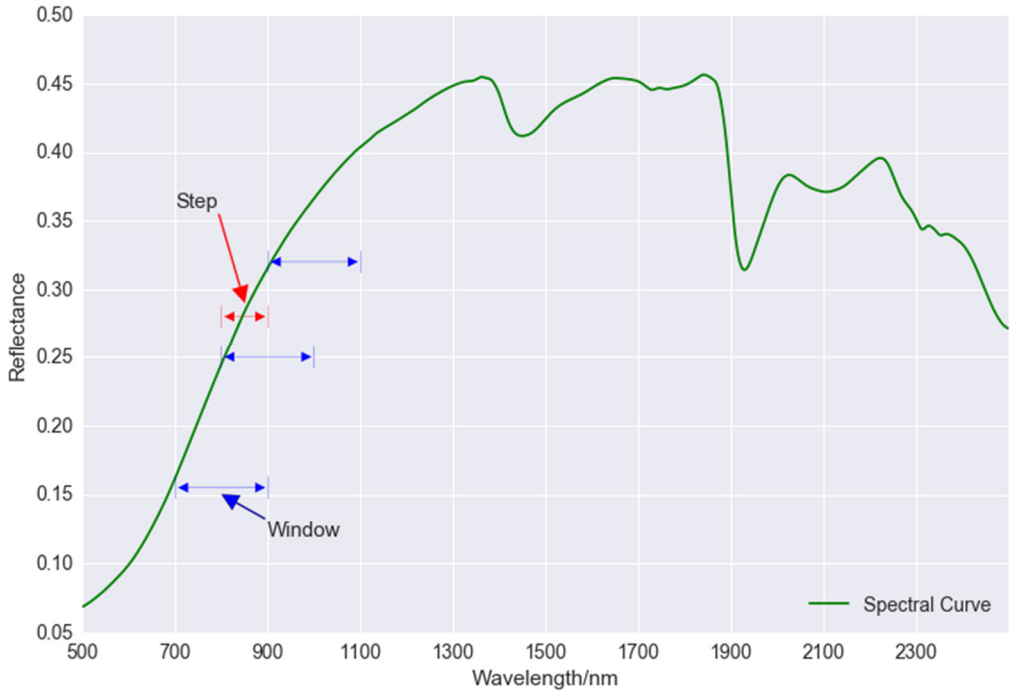

Figure 3.3 Illustration of the meaning of step and window size for multiple fractal feature generation. (step size = 100.0 nm, window size = 200.0 nm).

3.3.3 Gradient-boosting regression model

Soil spectroscopy quantitatively correlates with soil properties, which supposes that fitting a regression model with features extracted from spectra will have good predictive accuracies with respect to continuous soil properties. Gradient-boosting is a highly effective and widely used machine-learning approach [69]. Gradient-boosting develops an ensemble of tree-based models by training each of the trees in the ensemble on different labels and then combining the trees. It can produce robust and interpretable procedures for both regression and classification. For a regression problem where the objective is to maximize the coefficient of determination (R^2) or to minimize the root mean square error (*RMSE*), each successive tree is trained on the errors left over by the collection of earlier trees. XGBoost is a scalable and flexible gradient-boosting library [64,111,112], which is adopted to build the soil spectral quantitative model in our study. XGBoost uses more regularised model formalisation to control over-fitting, which gives it better performance. Mathematically, the model can be viewed as:

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i), \quad f_k \in F \tag{3.8}$$

where K is the number of trees, f is a function in the functional space F, and F is the set of all possible regression trees. Therefore, the objective of optimization can be written as:

$$obj(\theta) = \sum_{i}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k) \tag{3.9}$$

where $l(y_i, \hat{y}_i)$ is the training loss function, and $\Omega(f_k)$ is the regularization term. The goal of XGBoost model is to minimize $obj(\theta)$.

3.3.4 Evaluation

For each soil property, the soil spectral quantitative model was developed on a random sample of two-thirds of the selected soil samples using the gradient-boosting regression method. The calibrations were tested by predicting the soil properties on validation data sets composed of the remaining one-third of the organic soil samples. No samples were omitted from the analysis, nor the calibration or validation data sets. The model accuracies were evaluated on estimated and measured soil SOC, N and pH values using *RMSE*, R^2 and the ratio of percent deviation (*RPD*).

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2}{\sum_{i=1}^{n}(Y_i - \overline{Y})^2} \tag{3.10}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2} \tag{3.11}$$

$$RPD = \frac{SD}{RMSE} \tag{3.12}$$

where n is the number of validation samples, y is the measured values, $\overline{y}$ is the mean of the measured values, and $\hat{y}$ is the estimated values. *RPD* is the ratio of the standard deviation (SD) of the calibration data to the RMSE of the validation data [76]. An *RPD* <1.0 indicates a very poor model and its use is not recommended; an *RPD* between 1.0 and 1.4 indicates a poor model where only high and low values are distinguishable; an *RPD* between 1.4 and 1.8 indicates a fair model which may be used for assessment and correlation; *RPD* values between 1.8 and 2.0 indicate a good model where quantitative predictions are possible; an *RPD*

between 2.0 and 2.5 indicates a very good, quantitative model, and an *RPD* >2.5 indicates an excellent model.

3.4 Results

3.4.1 Fractal features for soil spectroscopy

For a single soil Vis-NIR-SWIR spectrum, the fractal dimension can be calculated by Equation (3.4). Before extracting fractal features from soil spectra, we first examined the relationship between soil properties and the corresponding fractal dimension. Spectral values of soil are relatively low and the curve appears smoother compared with other objects like vegetation. Thus, the resulting fractal dimension values are comparatively low. Since the fractal dimension is derived from the slope of the regression line obtained from the log-log plot of $\gamma_p(t)$ and lag t, one problem is how many lag increments are necessary to produce reliable results. Theoretically only a minimum of two points is necessary to make such a plot [108]. However, the results of such an analysis tend not to be reliable or representative. In this study, the value of lag increments was set as 5, and the Pearson correlations of soil properties and fractal dimensions are shown in Table 3.1. The Pearson is a standardised covariance and ranges from −1 to +1, which indicates a perfect negative (−1) or positive (+1) linear relationship respectively. A value of zero is not related to the independence between the two variables, and it only suggests no linear association. It can be seen that SOC, N and pH have negative relationships with fractal dimension. SOC and N have similar correlations with fractal dimension. Among these three estimators, the variogram-based fractal dimension calculation method achieved the best correlation between fractal dimension values and soil properties SOC (correlation coefficient (r) = −0.54), N (r = −0.50) and pH (r = −0.12).

Table 3.1 Pearson correlation coefficients between soil properties and fractal dimensions calculated by rodogram, madogram and variogram estimators.

	Rodogram	**Madogram**	**Variogram**
SOC	−0.40	−0.47	−0.54
N	−0.38	−0.43	−0.50
pH	−0.12	−0.13	−0.12

An intact spectrum can be divided into multiple segments, overlapping or non-overlapping. Each segment is corresponding to a fractal feature. When step size and window size are respectively set to 2.5 nm and 50.0 nm, a total number of 791 fractal features can be derived by rodogram, madogram or variogram methods, resulting in the original spectral dimension reduced from 4000 to 791. In order to make a proper comparison between the generated fractal feature-based curve and the raw spectral curve, the centre wavelength value of the spectral segment is assigned to the fractal feature as the corresponding "wavelength number".

A great advantage of fractal-based feature extraction is that the curve shape of fractal features is similar to the shape of the raw spectrum, which makes it possible to apply methods like continuum removal (CR) not only to the raw spectrum but also to the fractal-based "spectrum". The organic soil samples can be divided into four groups according to the content of SOC. Average spectral reflectance and continuum removal reflectance of LUCAS organic soil samples were computed by SOC classes (Figure 3.4A). For fractal features, average fractal energy and continuum removal responses of organic soil samples were also computed and shown in Figure 3.4B–D. The highest SOC class that was above 480 g/kg showed the highest mean reflectance in the wavelength range from 1000.0 nm to 2000.0 nm, which is consistent with observations in the literature [26]. The continuum removal reflectance showed a strong correlation with SOC content at a wavelength of near 600.0 nm. The difference between the raw spectral curve and fractal feature curve was not obvious from the view of shape. Fractal features showed shallow absorption peak in proportion for SOC classes at a wavelength of 600.0 nm. The fractal energy values were larger than reflectance values, as the former were multiplied by spectral energy and fractal dimension, which was supposed to be larger than 1.0.

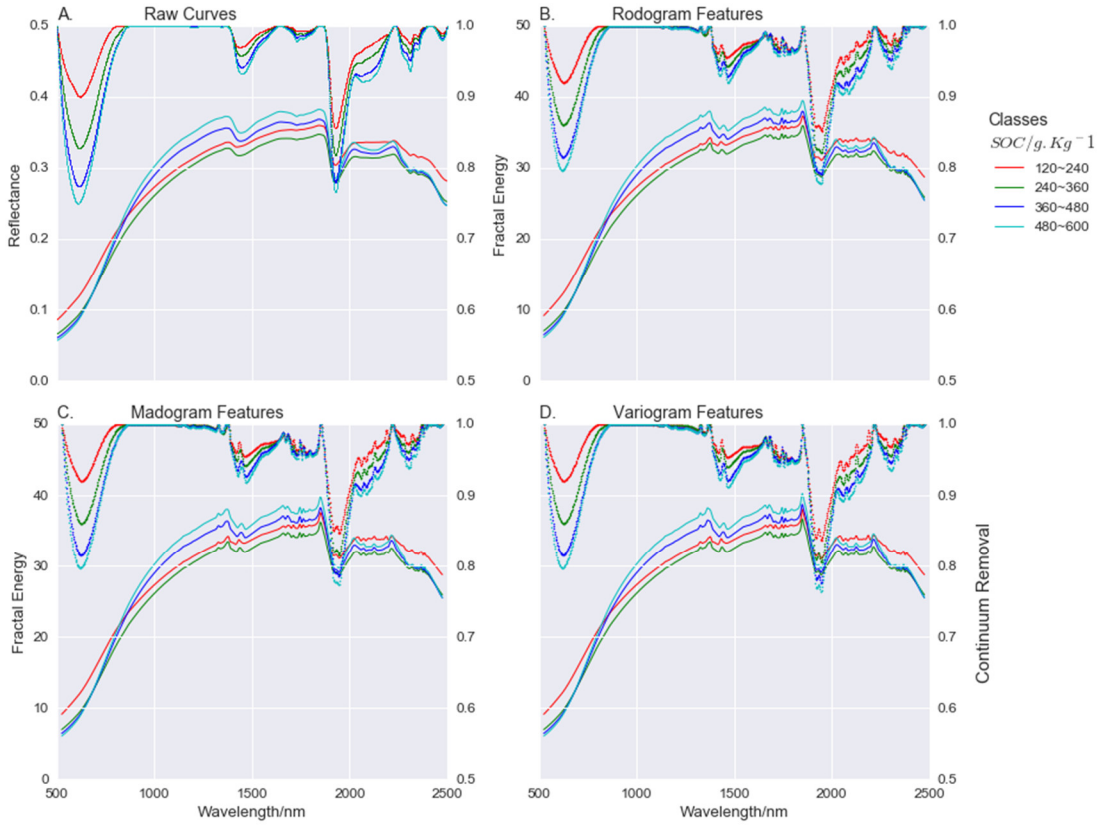

Figure 3.4 (**A**) Average spectral reflectance and continuum removal reflectance of LUCAS organic soil samples computed by SOC classes; (**B–D**) Average fractal energy and continuum removal responses of organic soil samples computed by SOC classes using rodogram, madogram and variogram estimators respectively. The central wavelength number of the corresponding spectral segment is assigned to the fractal feature.

To demonstrate the effects of step and window size on extracted fractal features, the combinations of the two parameters were tested. When the step size was fixed at 2.5 nm, a series of fractal feature curves were derived by defining window sizes as 15.0 nm, 35.0 nm, 55.0 nm, 75.0 nm and 95.0 nm. With the increase of window size, fractal energies correspondingly increased and the shapes of fractal features were also gradually exaggerated, as shown in Figure 3.5A. The number of fractal features derived at different window sizes was equal but less than raw spectral features. When the window size was fixed at 50.0 nm and step size increased from 10.0 to 50.0 nm at an interval of 2.5 nm, the number of fractal features was non-linearly decreased from 196 to 40 as shown in Figure 3.5B.

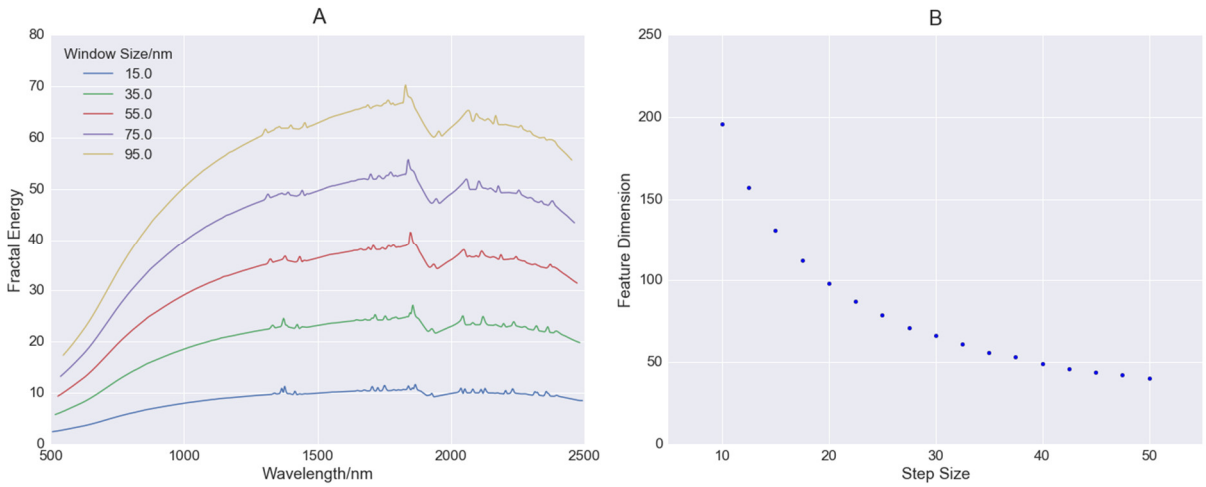

Figure 3.5 The effect of step and window size on generated fractal features. (**A**) are fractal feature curves when window sizes were at 15.0–95.0 nm (step size fixed at 2.5 nm); (**B**) is the number of fractal features when step sizes were increased from 10.0 to 50.0 nm (window size fixed at 50.0 nm).

3.4.2 Effects of different step and window size on extracted fractal features

For further analysis about effects of step and window size on the relationship between fractal features and soil properties, a matrix of step–window pairs was generated by defining step size ranging from 2.5 nm to 50.0 nm at an interval of 2.5 nm and window size ranging from 10.0 nm to 100.0 nm at an interval of 5.0 nm. For each pair of these two parameters, fractal features were derived according to Equation (3.6). A gradient-boosting regression model using the XGBoost tool was built on a random sample of two-thirds of organic soil samples, and then applied to the estimation of each sample from the validation dataset. Pre-processing methods for soil spectra could also be applied to new fractal features because of the shape similarity between fractal features and the raw spectral curve. For example, fractal features were smoothed by use of Savitzky-Golay filter. R^2 derived by step–window pairs for SOC using rodogram, madogram and variogram methods are shown in Figure 3.6A2–A4 respectively, as is the case for N and pH in Figure 3.6B-C. For a comparable study, the regression model was also applied to raw spectral values and PCA-transformed data.

Taking advantage of fractal features, models developed for SOC estimation achieved comparably good results, R^2 varies from 0.64 to 0.83 (rodogram), 0.70 to 0.84 (madogram) and 0.72 to 0.84 (variogram). For pH, R^2 varies from 0.61 to 0.80 (rodogram), 0.63 to 0.80 (madogram) and 0.63 to 0.82 (variogram. However, the accuracies are comparatively lower

for N. R^2 varies from 0.52 to 0.74 (rodogram), 0.53 to 0.75 (madogram) and 0.55 to 0.76 (variogram). Models with raw spectra were developed by evenly selecting the desired number of spectral measurements. The Hughes phenomenon can be seen well in models built with raw spectra. R^2 increased first and then declined with the increase of feature numbers. It can be seen that models with raw spectra had the poorest performance. For SOC and N, fractal features outperformed PCA-transformed features and raw spectra. Fractal features for pH achieved similar accuracies compared to PCA-transformed features.

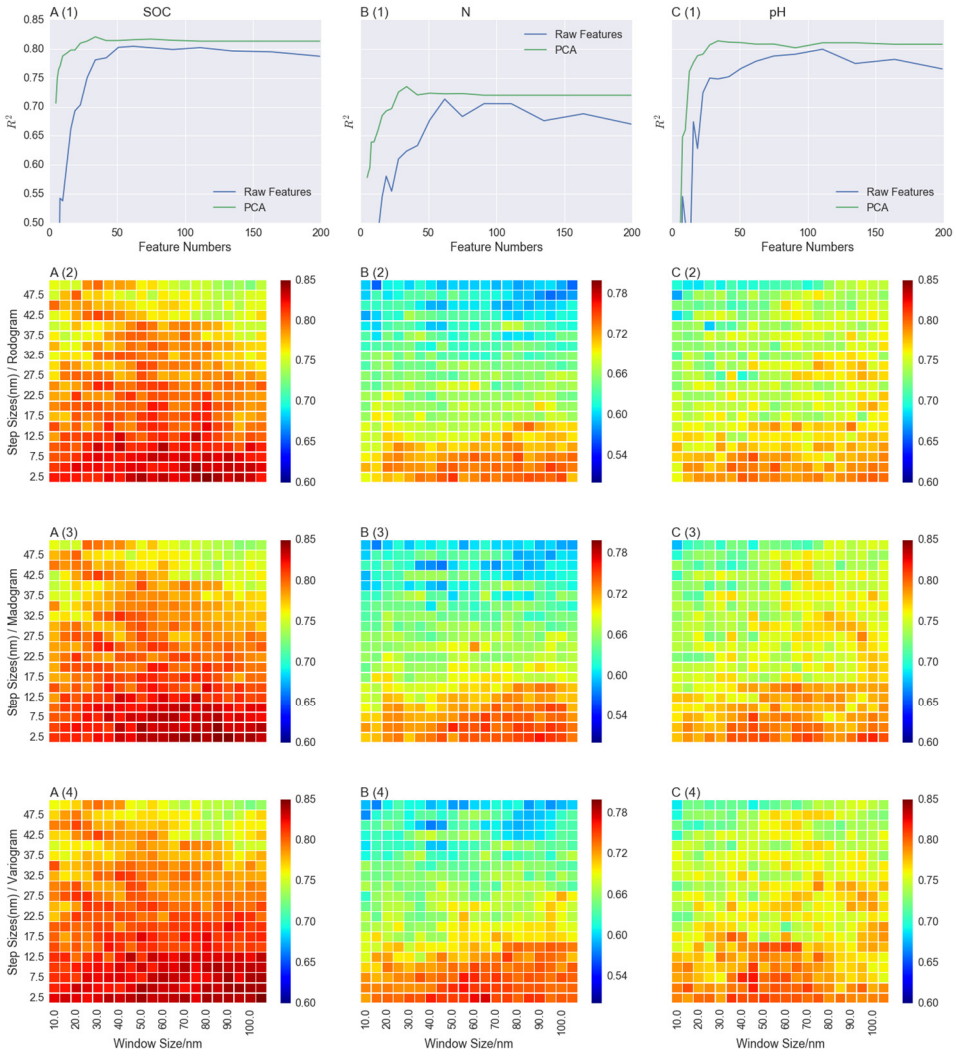

Figure 3.6 Gradient-boosting regression modelling accuracies for SOC, N and pH. (**A1**), (**B1**) and (**C1**) were with principal component analysis (PCA)-transformed features and raw spectra; (**A2**), (**B2**) and (**C2**) were with fractal features derived by the rodogram method with various step-window pairs. (**A3**), (**B3**) and (**C3**) were with fractal features derived by the madogram method with various step-window pairs. (**A4**), (**B4**) and (**C4**) were with fractal features derived by the variogram method with various step-window pairs.

3.4.3 Modelling soil properties with fractal features

Window sizes and step sizes adopted to optimise the gradient-boosting regression model can be seen in Section 3.2. Fractal feature numbers approximately ranged from 40 to 800. The optimal pairs of step–window sizes for SOC, N and pH can be seen in Table 3.2. For each gradient-boosting regression model built with XGBoost library, the maximum tree depth was 4 and a maximum number of trees was 100. R^2 was used as the evaluation metric for validation data.

The best trade-off between step and window size for SOC (R^2 = 0.851, *RMSE* = 56.7 g/kg, *RPD* = 2.59) was 2.5 nm for the former and 105.0 nm for the latter with variogram estimator. The best performance step–window sizes for N (R^2 = 0.776, *RMSE* = 3.01 g/kg, *RPD* = 2.09) were step size at 2.5 nm and window size at 65.0 nm with the variogram estimator. The best performance step–window size for N (R^2 = 0.822, *RMSE* = 0.49, *RPD* = 2.31) were step size at 7.5 nm and window size at 45.0 nm with the variogram estimator. From Table 3.2, it can be seen that fractal-based feature extraction methods tend to keep a much larger number of features compared to PCA. To achieve similar performance of PCA, fractal-based approaches need to retain ~200 features, such as 190 for SOC (R^2 = 0.819, *RMSE* = 62.49 g/kg, *RPD* = 2.34) where step size and window size were respectively 10.0 nm and 105.0 nm, 128 features for N (R^2 = 0.736, *RMSE* = 3.26 g/kg, *RPD* = 1.92) where step size and window size were respectively 15.0 nm and 135.0 nm, and 131 features for pH (R^2 = 0.807, *RMSE* = 0.50, *RPD* = 2.22) where step size and window size were respectively 15.0 nm and 50.0 nm.

In real-world examples, there are many ways to extract features from a dataset. Often it is beneficial to combine several methods to obtain good performance. To assess whether predictive accuracy could be enhanced by integrating multiple features, the first 30 PCA components were combined with fractal features and then ingested into the gradient-boosting regression model. Combined features showed better performance when applied for the estimation of all three soil properties, SOC (R^2 = 0.86, *RMSE* = 55.16 g/kg, *RPD* = 2.7), N (R^2 = 0.78, *RMSE* = 2.96 g/kg, *RPD* = 2.19) and pH (R^2 = 0.85, *RMSE* = 0.44, *RPD* = 2.59), as shown in Figure 3.7.

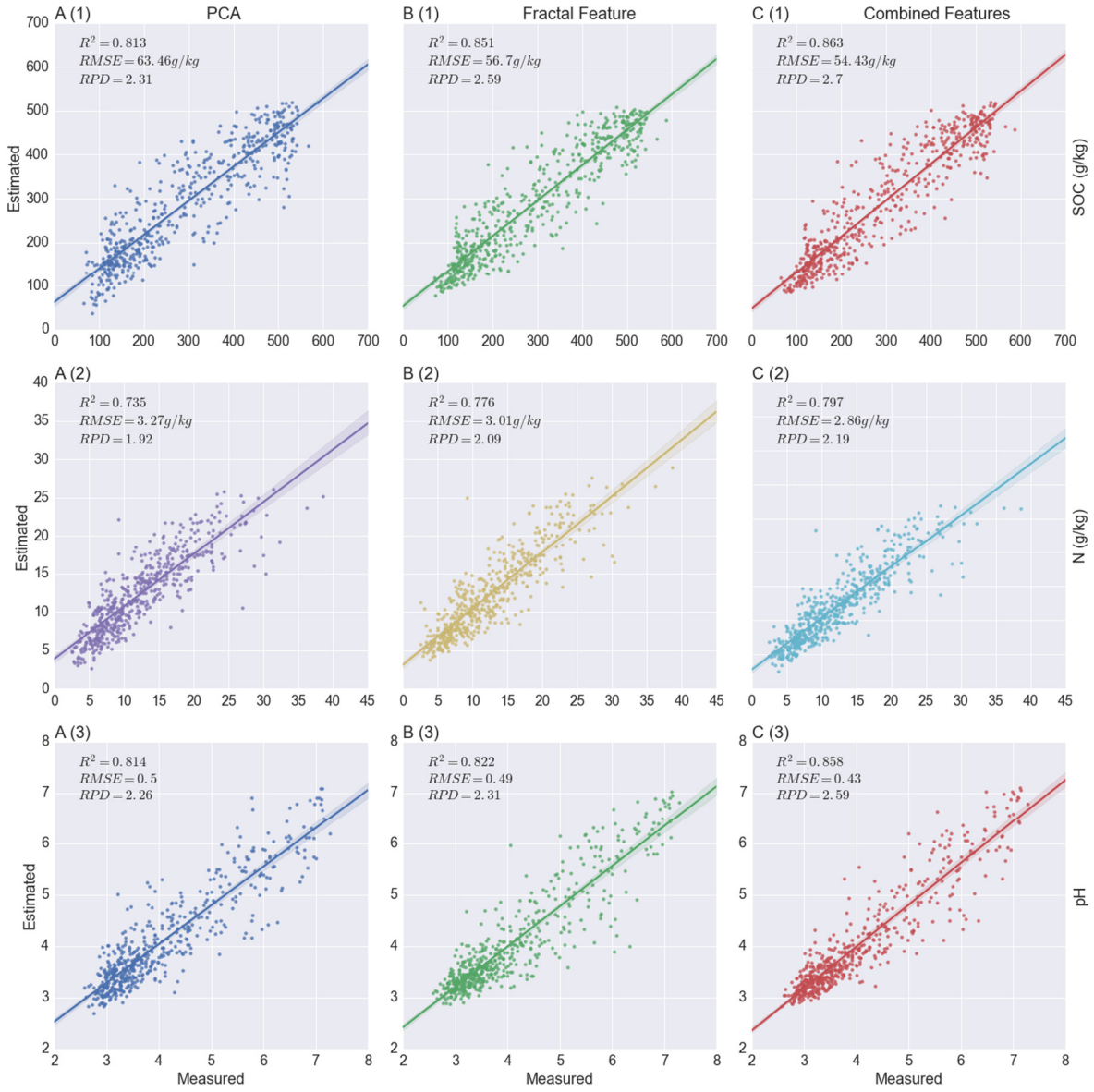

Figure 3.7 Best performance of gradient-boosting regression modelling accuracies for SOC, N and pH. (**A1**), (**A2**) and (**A3**) were with PCA-transformed features. (**B1**), (**B2**) and (**B3**) were with fractal features. (**C1**), (**C2**) and (**C3**) were with features combined by PCA-transformed features and fractal features. R^2: coefficient of determination; RMSE: root mean square error; RPD: the ratio of percent deviation.

Table 3.2 Best Performance step–window pairs for soil properties estimation using fractal-based feature extraction and comparison with PCA.

	Method	**Step Size/nm**	**Window Size/nm**	**Dimension**	R^2
SOC	PCA	-	-	28	0.813
	Rodogram	2.5	80	769	0.847

	Method	Step Size/nm	Window Size/nm	Dimension	R^2
	Madogram	2.5	90	765	0.847
	Variogram	2.5	105	759	0.851
N	PCA	-	-	34	0.735
	Rodogram	2.5	50	781	0.756
	Madogram	2.5	90	765	0.767
	Variogram	2.5	65	775	0.776
pH	PCA	-	-	34	0.814
	Rodogram	5	55	390	0.806
	Madogram	2.5	100	761	0.818
	Variogram	7.5	45	261	0.821

3.4.4 Comparison with PLS regression

PLS regression is frequently used to calibrate soil properties with soil spectra, and it can maximise the covariance between the spectra and a measured soil property [7]. To make a comparison, PLS regression, named as method A for the sake of convenience, was applied to the raw spectra of the LUCAS organic soil to estimate organic carbon (SOC) contents, and the best performance (R^2 = 0.834) was achieved when the number of components was 60 (Figure 3.8).

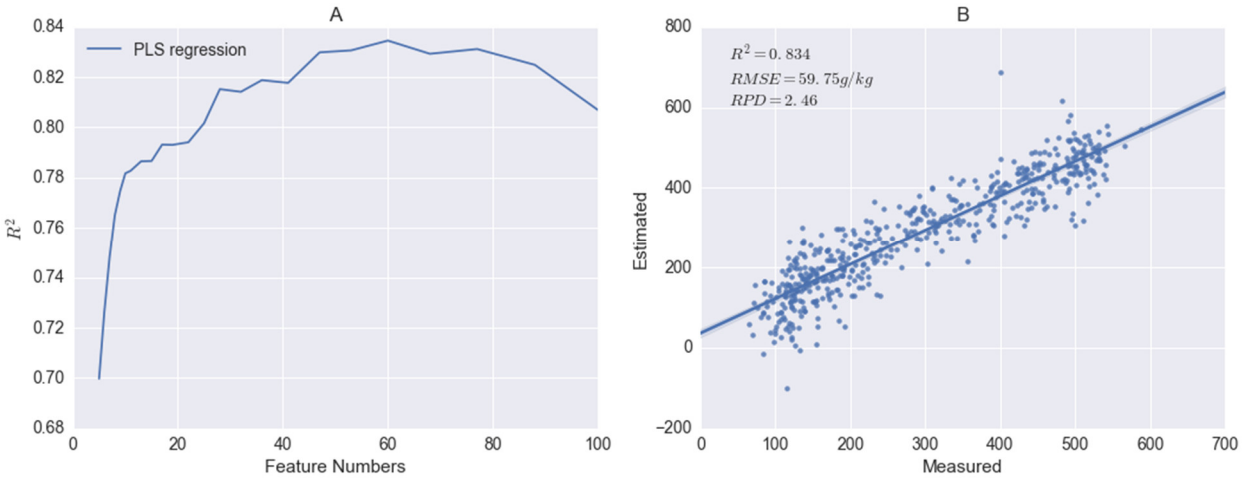

Figure 3.8 The change of R^2 with the increase of the partial least squares (PLS) component number (**A**) and the PLS regression model when the component number was 60 (**B**).

PLS regression integrates the compression and regression steps, and it can be viewed as a combination of PLS components and linear regression [54]. Therefore, it is also possible to transform the raw spectra into PLS components and then ingest them into the gradient-boosting regression model (method B). The same gradient-boosting model parameters were adopted. When the number of retained PLS components was 60, the achieved R^2 for the estimation of SOC contents was 0.846 (Figure 3.9).

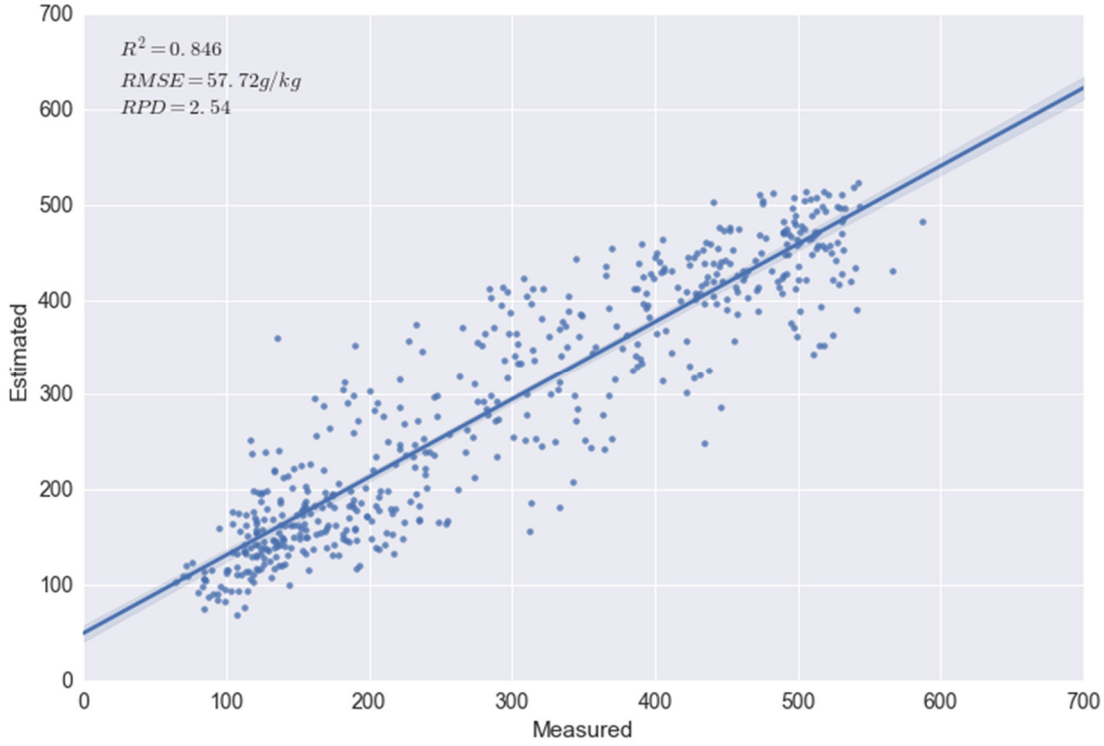

Figure 3.9 The gradient-boosting regression model with PLS components for the estimation of SOC contents.

The quantitative method proposed in the paper can be viewed as a combination of fractal features and gradient-boosting regression (method C), and it achieved the best performance (R^2 = 0.851) for the estimation of SOC contents of these three methods. We also applied methods A and B to the estimation of N and pH contents. For N, the same case applied; method C showed the highest R^2. Although method A (PLS regression) achieved the best performance for the estimation of pH contents, when focusing on extracted features, fractal features had similar performance compared with PLS components, the R^2 for method C being 0.821 and for method B, 0.823. The only difference between these two methods was the

ingested features. The results are summarised in Table 3.3, and it can be seen that fractal features can achieve similar or even better results compared with PLS components.

Table 3.3 Comparison of three methods for the quantitative retrieval of soil properties.

	Features	Modelling	SOC (R^2)	N (R^2)	pH (R^2)
Method A	PLS components	Linear regression	0.834	0.743	0.87
Method B	PLS components	Gradient-boosting regression	0.846	0.759	0.823
Method C	Fractal features	Gradient-boosting regression	0.851	0.776	0.821

3.5 Discussion

3.5.1 The importance of fractal dimension for soil spectra

The correlations between fractal dimension and soil properties were assessed by means of Pearson correlation analysis when the fractal dimension calculation was applied to the whole spectrum. Significant negative correlations for SOC (r = −0.54) and N (r = −0.50) with the fractal dimension were found, which means that values of SOC and N could have effects on the shape of soil spectra and therefore diagnostic wavelengths exist for SOC and N. In [86] an absorption peak centred at 600 nm was observed, which seems to be related to SOC content. At 2100 nm, there was an absorption determined by N content. In [26] the authors also highlighted that wavelengths of around 1100, 1600, 1700–1800, 2000, and 2200–2400 nm have been identified as being particularly important for SOC and N estimation.

The pH showed a very weak correlation with the fractal dimension (r = −0.12), which could be caused by a lower direct spectral response to soil pH [26]. It has to be pointed out that the weak correlation between pH and fractal dimension does not mean that soil spectra cannot be used to quantify soil pH values. The variation of soil pH values does not significantly contribute to the smoothness or roughness of the spectral curve. Soil pH value can still be well estimated in the laboratory or the field [55,56] using raw spectral data, which might be due to the mutual effects of spectrally active soil constituents such as organic matter and clay [43]. It also can be seen that the Pearson correlation between fractal dimension and soil properties has a positive relationship with the performance of fractal features.

3.5.2 Modelling soil properties with fractal features

Three methods for the fractal dimension calculation and further feature extraction were studied in this paper. The results demonstrated that the variogram estimator had slightly better performance than the madogram estimator when applied to fractal feature generation for soil property estimation, and methods using these two estimators achieved better R^2 than the method using the rodogram estimator. In [114] the classification achieved better results with texture layers derived from the madogram. Since the madogram estimator calculates the sum of the absolute value of the semivariance for all observed lags, it yields a softer effect on the presence of outliers compared to the variogram estimator. However, in our study, soil spectra were well pre-processed by the Savitzky–Golay filter and generated fractal features. Fractal features generated by these three estimators have a similar curve shape and achieved very close estimation accuracies for tested soil properties.

Step–window pairs have a significant impact on estimation accuracies of soil properties. When the window size is fixed, accuracies are decreased with the increase of step size. However, when the step size is fixed, accuracies are prone to ascend slightly and then clearly descend. A higher R^2 was found to be located at the bottom of the step–window matrix. However, there is no guarantee as to which step–window pair is the best parameter for soil property estimation. Therefore, a hyper-parameter optimisation method should be adopted for each of the soil properties.

In general, fractal features achieved better results compared to PCA-transformed features and raw spectra. This demonstrates that by taking advantage of fractal information encoded in the soil spectral shape, soil properties can be estimated in a better way. Besides, when raw data are transformed or projected via PCA, measurement units and shape are lost. However, fractal-based feature extraction is prone to retaining a much larger number of features compared to PCA. To achieve similar performance, the fractal-based approach needs ~200 feature numbers while PCA only needs ~30. When compared with PLS components, fractal features also had better performance for the estimation of OC and N contents. However, there is no conflict between common feature extraction practices with the proposed fractal method. When integrating different kinds of features, like PCA-transformed features and fractal features, the performance is expected to be improved for the retrieval of soil properties.

3.6 Conclusions

Data acquisition with Vis-NIR-SWIR spectroscopy is relatively easy, and a wide range of soil properties can be analysed within a comparatively short time with relatively little effort for sample preparation. Soil spectroscopy has recently been identified as a method that has the potential to rapidly estimate soil properties. Many soil-spectral libraries are already built at regional, continental or even global scales. Various multivariate statistical methods have been successfully adopted to explore the relationship between soil spectra and soil physical/chemical properties. However, few studies are focused on feature extraction from measured soil spectra, which is also crucial to correlating spectra with soil properties.

This study presents a novel methodology for feature extraction based on fractal geometry. Each Vis-NIR-SWIR spectrum can be divided into multiple segments by defining the moving window size and the step size. For each segmented spectral curve, the fractal dimension value was calculated using variation estimators. Fractal features, generated by multiplying the fractal dimension value with spectral energy, were further combined with PCA-transformed features, and the gradient-boosting regression model achieved good performance with respect to the retrieval of SOC (R^2 = 0.86, *RMSE* = 55.16 g/kg, *RPD* = 2.7), N (R^2 = 0.78, *RMSE* = 2.96 g/kg, *RPD* = 2.19) and pH (R^2 = 0.85, *RMSE* = 0.44, *RPD* = 2.59). Fractal analysis can be functionalised as an approach to examine the relationship between soil spectra and soil properties, which can characterise statistical self-similarity and further quantify the irregularity of soil spectra [109]. Fractal features, by taking advantage of fractal information encoded in the shape of soil spectral curve, can reflect the impact of various properties on soil spectra except when the properties have a less direct spectral response. In this case, fractal features can still be functioned to quantify the corresponding soil property. Fractal features performed well when ingested into quantitative soil spectroscopic models, and the proposed fractal method can not only reduce the dimensionality in the original space, but also simultaneously maintain the spectral shape, which means that methods for raw spectra can also be applied to extracted fractal features, for example, calibrating soil properties using PLS regression with fractal features.

Acknowledgments

The first author wants to express acknowledgement to the China Scholarship Council (CSC) for providing financial support to study at TU Dresden. The LUCAS topsoil dataset in this work was made available by the European Commission through the European Soil Data Centre and managed by the Joint Research Centre (JRC) http://esdac.jrc.europa.edu/. We acknowledge support by the German Research Foundation and the Open Access Publication Fund of the TU Dresden. We also thank the academic editors and the anonymous reviewers for their valuable comments.

Chapter 4

Transfer Learning for Soil Spectroscopy Based on Convolutional Neural Networks and Its Application in Soil Clay Mapping Using Hyperspectral Imagery

4.1 Abstract

Soil spectra are often measured in the laboratory, and there is an increasing number of large-scale soil spectral libraries establishing across the world. However, calibration models developed from soil libraries are difficult to apply to spectral data acquired from the field or space. Transfer learning has the potential to bridge the gap and make the calibration model transferrable from one sensor to another. The objective of this study is to explore the potential of transfer learning for soil spectroscopy and its performance on soil clay content estimation using hyperspectral data. First, a one-dimensional convolutional neural network (1D-CNN) is used on Land Use/Land Cover Area Frame Survey (LUCAS) mineral soils. To evaluate whether the pre-trained 1D-CNN model was transferrable, LUCAS organic soils were used to fine-tune and validate the model. The fine-tuned model achieved a good accuracy (coefficient of determination (R^2) = 0.756, root-mean-square error (*RMSE*) = 7.07 and ratio of percent deviation (*RPD*) = 2.26) for the estimation of clay content. Spectral index, as suggested as a simple transferrable feature, was also explored on LUCAS data, but did not performed well on the estimation of clay content. Then, the pre-trained 1D-CNN model was further fine-tuned by field samples collect in the study area with spectra extracted from HyMap imagery,

achieved an accuracy of R^2 = 0.601, *RMSE* = 8.62 and *RPD* = 1.54. Finally, the soil clay map was generated with the fine-tuned 1D-CNN model and hyperspectral data.

4.2 Introduction

Soil spectroscopy has the capability to rapidly and non-destructively analyse soil properties by taking advantage of visible near-infrared shortwave infrared (Vis–NIR–SWIR) spectral information [3,26,77,87,115]. There are numeral studies related to the reliable estimation of soil properties using prepared soil samples and measured spectral data [7,8,34,116]. Although the relationship between soil properties and the corresponding spectra is complex and soil spectroscopy is less accurate than wet chemistry, it still achieved great success in laboratory studies, which naturally leads to the exploration of imaging spectroscopy (IS) for characterising soil properties at large scales. It not only has the capability of obtaining spectral information at several hundred spectral bands as laboratory spectroscopy does, but also provides a spatial view, which cannot be achieved by laboratory techniques [117]. IS technology provides the opportunity to map various soil properties at regional and global scales at comparatively low costs.

The spectral features and quantitative estimation of clay content in soil have been explored in previous studies [118–121]. In Reference [45], the clay content was demonstrated to be strongly correlated with the clay minerals in soil and the principal characteristic bands were related to the lattice hydroxyl groups. Clay minerals have characteristic absorptions near 1400 nm and 2200 nm [26]. The absorption feature near 1400 nm is due to overtones of the O-H stretch vibration, while the absorption near 2200 nm is due to Al-OH bend plus O-H stretch combinations. A clay spectral index was further proposed using the absorption feature near 2200–2300 nm in Reference [31]. The performance of spectra measured in the well-controlled laboratory and acquired from IS sensors has been assessed by many case studies for soil property estimation [122–126]. The accuracy using imaging spectroscopy is comparatively lower than the result obtained from laboratory spectroscopy, as the application of imaging spectroscopy in the assessment of topsoil properties is constrained by many factors, such as the low signal-to-noise ratio, atmosphere attenuation, revisiting time, sensor radiometric and spatial resolutions, vegetation coverage and Bidirectional Reflectance Distribution Functional

(BRDF) [16,127]. The distance from the sensor to soil samples is often between 1 and 140 cm in the laboratory [22] while the IS has a much far distance, like the satellite-borne Hyperion hyperspectral flying at 705 km altitude [128]. The calibration and performance of different sensors are also determinants of the quality of measured spectra. Soil samples are often illuminated with 1-6 light sources [22] in the laboratory while air- or satellite-borne hyperspectral imagery is obtained under solar illumination. Besides, laboratory soil samples are dried, crushed and sieved while imaging targets in the field are natural surfaces with heterogeneous surface temperatures, moisture levels and roughness [6]. Moisture effects on the soil spectral reflectance have been studied extensively [115,129]. The overall reflectance generally decreased, with an increasing amount of moisture. Furthermore, the absorption by water in the SWIR region impacted clay-associated absorption features [130]. These factors lead to spectral differences between laboratory and remotely sensed data.

Soil spectral libraries can be used as a reference for retrieving soil attributes by reflectance spectroscopy. Calibrations are not reliable for soils not represented in the soil spectral library, hence there is a need for building libraries representative of the soil diversity [17,18] and an increasing number of large-scale soil spectral libraries established at national, continental and even global levels. As a key innovation, near and mid-infrared spectroscopy are used for soil analysis in the collaborative Africa Soil Information Service (AfSIS) project, which covers an area, including about 17.5 million km^2 of continental sub-Saharan Africa (SSA) and almost 0.6 million km^2 of Madagascar [94]. In the first period (2009–2012) of Land Use/Land Cover Area Frame Survey (LUCAS), which is an extensive topsoil survey that is carried out across the European Union to derive policy-relevant statistics on the effect of land management on soil characteristics, soil spectra of about 20,000 topsoil samples were acquired in the range of 400–2500 nm and extensively studied [24,34,85,86,103,131]. A new LUCAS sampling campaign will be undertaken in 2018 [19]. A voluntary collaborative project was started in 2008 to develop a global library of soil spectra, and 23,631 soil spectra have been contributed to the global database by around 45 soil scientists and researchers from 35 institutions [8]. In addition, there are a number of national and regional soil spectral libraries have been established, such as the ones for Australia [20], Czech Republic [21], Brazil [22] and China [10]. A soil library typically contains soil attributes as done by wet chemistry standard methods and reflectance spectra acquired under a routine protocol and spectrometer. However, there is still lack of protocols

for soil spectral measurements. The internal soil standard (ISS) concept is proposed to make soil spectra from different libraries sharable by minimising the systematic effects [132,133].

Large soil spectral libraries should help to reduce or even save the need to collect and analyse new samples for site-specific calibrations to estimate soil properties, and it could be a strong base for hyperspectral remote sensing of soils from space [3]. The laboratory soil spectra may enable appropriate validation of the reflectance information acquired from IS sensors. However, there are still few studies integrating IS with laboratory studies [4–6,134,135]. In Reference [127], it is pointed out that calibration models developed from laboratory processed samples cannot be utilised for field spectroscopy, due to the influence of external environmental factors (such as soil moisture, soil roughness, atmospheric effect and vegetation coverage). Furthermore, spectroscopic models achieved by common calibration methods are usually not transferrable. An important drawback of Partial least squares (PLS) regression is the complexity of the transfer of spectroscopic models from one sensor to another [5,136]. When samples to be predicted are far away from the spectral library, the regression algorithm is prone to fail in producing reliable model soil predictions [24].

It is suggested that spectral indices may provide an alternative method to PLS regression for quantifying soil contents in situations where calibration models should be transferred between different spectrophotometers [137]. The soil organic carbon (SOC) estimation was carried out using simple and multiple linear regression techniques based on image reflectance values and spectral indices, which confirmed that spectral indices have potential to be transferred among airborne and satellite hyperspectral sensors [138]. Spectral indices can be viewed as simple transferable features developed by combining surface reflectance at two or more wavelengths that indicate relative abundance of features of interest. A number of soil spectral indices have been proposed for the estimation of SOC, soil salinity, soil clay and iron [30,31,137,139]. Transfer learning aims to propagate the knowledge from a source domain to a target domain [140]. Therefore, it has the potential to make calibration models transferable from one sensor to another. Transfer learning with the pre-trained convolutional neural network (CNN) model has been proposed for remote sensing. CNNs can learn representative and discriminative features in a hierarchical manner from the data [81], and have recently been widely used in various remote sensing data analysis tasks, such as classification, segmentation, object detection, image registration, and change detection [141–145]. A

comprehensive review and list of resources using CNNs for remotely sensed data can be found in Reference [146]. The transferability of the natural image features from the pre-trained CNN models has been explored to the limited amount of high-resolution remote sensing scene datasets with the feature coding methods [147]. The advantage of adopting pre-trained CNN models is the effective extensible properties for dealing with the high-resolution remote sensing imagery scenes with limited labelling. A transfer learning method with fully pre-trained CNNs (CNN-FT-Full) was proposed to overcome the separation of asynchrony of different parts of the transferred CNNs during the learning process, and it performed well on land-use classification with high-resolution remote sensing images [148]. In Reference [149], transfer learning was proposed to transfer knowledge learned from a large amount of unlabelled SAR scene data (50,000 image patches extracted from TerraSAR-X scene images) to SAR target recognition tasks. However, there are still few studies using pre-trained CNN models in soil spectroscopy.

The objective of this study is to explore the potential of transfer learning for soil clay mapping using hyperspectral imagery and a pre-trained CNN model developed from a large number of spectra measured in the laboratory. Descriptions of laboratory and airborne spectral data are given in Section 2.1. The proposed workflow and model performance metrics are presented in Section 2.2. The results of the calibration and validation for soil clay content retrieval using laboratory-derived spectral library and the transferability for airborne spectral data are presented and subsequently discussed in Section 3. Conclusions are given in Section 4.

4.3 Materials and Methods

4.3.1 Datasets

4.3.1.1 The LUCAS Soil Spectral Library

The first dataset utilised for developing and evaluating the pre-trained one-dimensional convolutional neural network (1D-CNN) model is LUCAS soil spectral library, which contains approximately 20,000 geo-referenced soil samples that collected and analysed across Europe [67,68]. A standardised sampling procedure was adopted to collect around 0.5 kg of topsoil

(0–20 cm) in the field. The distribution of LUCAS soil samples can be seen in Figure 4.1. Soil samples can be divided into mineral and organic soils according to [86]. Soil spectra were measured using a FOSS XDS Rapid Content Analyser, operating in the 400–2500 nm wavelength range, with 0.5 nm spectral resolution. Pre-processed included transformation of absorbance (A) spectra into reflectance ($1/10^{A}$) spectra and Savitzky-Golay Filter with a window size of 50, second order polynomial. The laboratory spectral data were resampled to be in consistent with bands of the HyMap imagery, so that the model developed using LUCAS data can also accept HyMap data as inputs.

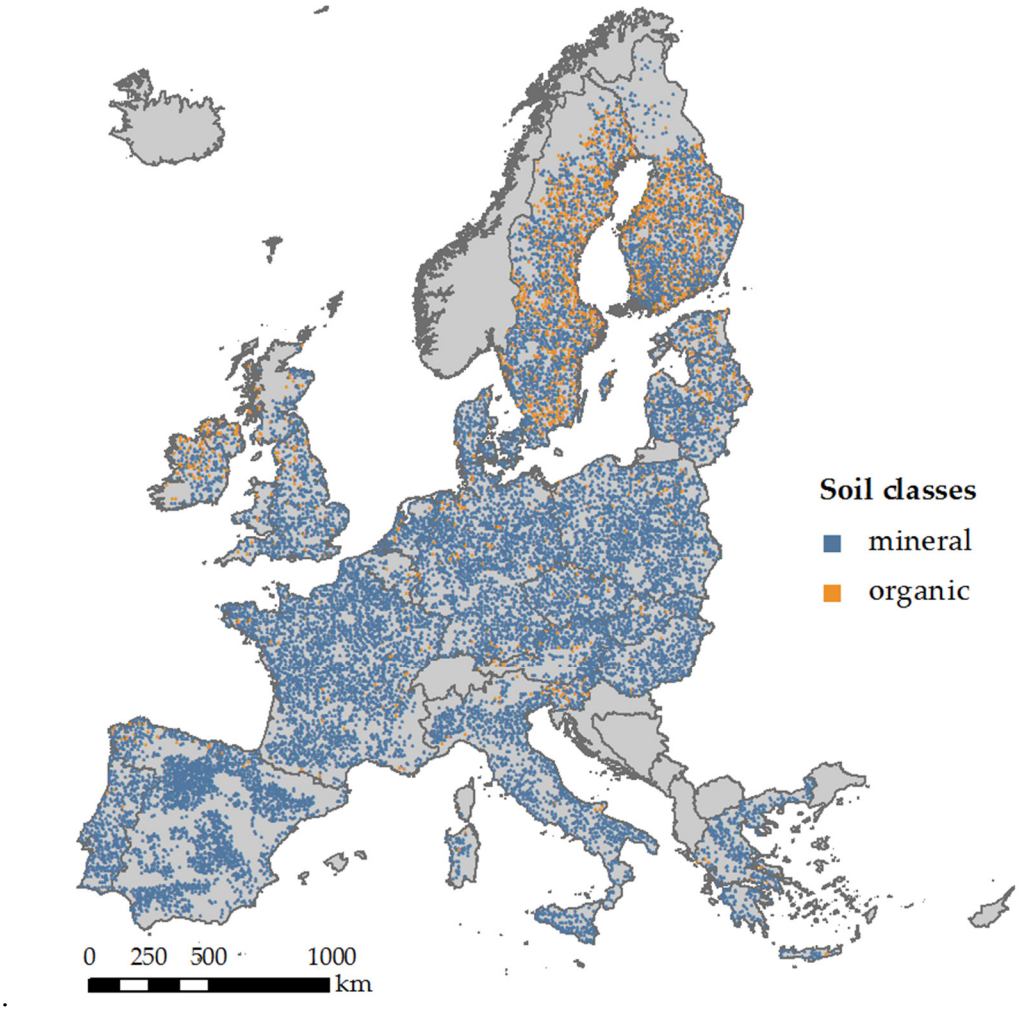

Figure 4.1 Distribution of mineral and organic soils from the LUCAS soil spectral library.

4.3.1.2 Cabo de Gata-Nijar hyperspectral imagery

The second dataset is the hyperspectral imagery acquired in the Natural Park Cabo de Gata-Níjar in the Almeria province of southeastern Spain. Our study focuses on a small area at Cortijo del Fraile, which is an agricultural area in the middle of the park with mostly bare

fields at the time of the overflight. In June 2005, airborne hyperspectral data were obtained over the small area with the HyMap sensor (Baulkham Hills, NSW, Australia) [150]. It provided spectral images after processing to geocoded reflectance covering the spectral range of 400 to 2450 nm with a spectral resolution of 12 to 17 nm [43]. The average flight altitude of 2645 m above sea level resulted in a spatial resolution of 5 m. The raw HyMap data were corrected to at-sensor-radiance based on calibration coefficients obtained during laboratory calibration by HyVista. The atmospheric correction was performed with ATCOR4 software. A mask was applied to the airborne data to keep pixels of bare soil surface only. The soil mask (Figure 2B)was created following the approach provided by ENSOMAP software, which is an open source tool for quantitative soil properties mapping based on hyperspectral imagery [151].

32 soil samples were randomly taken from the upper soil surface (0–2 cm) in the study area and the corresponding locations can be seen in Figure 4.2A. Samples were air dried and passed through a 2 mm sieve before laboratory analysis. The particle size distribution was determined by wet sieving the sand fraction and using the pipette method for silt and clay fractions after the removal of organic matter with H2O2 and dispersion with Na-hexametaphosphate. The clay content values of field samples vary between 8.4% and 63.4%. Collected soil samples were randomly divided into two subsets with a ratio of 1:1 to calibrate and validate the fine-tuned model. A brief statistical summary can be seen in Table 4.1.

Table 4.1. Statistics of soil clay content for the calibration and validation dataset.

Dataset	Number	Mean (%)	Standard Deviation (%)	Min (%)	Max (%)
Calibration	16	30.2	14.1	10.8	63.4
Validation	16	27.7	13.6	8.4	50.2

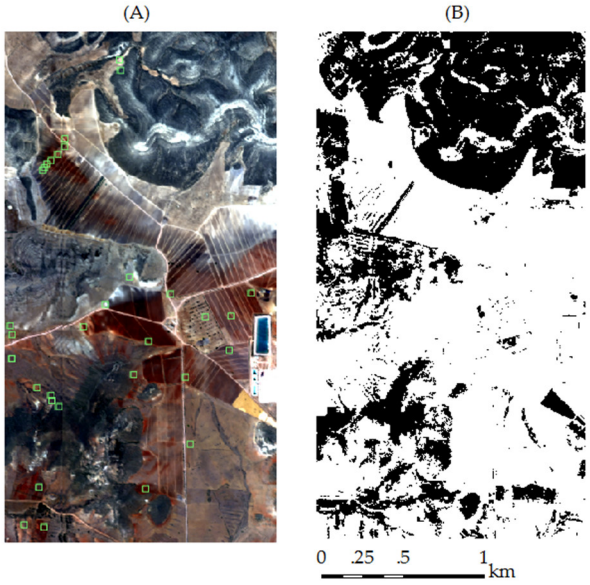

Figure 4.2 HyMap imagery (A) and the soil mask (B) in the study area Cabo de Gata-Nijar. The locations of field samples were shown in green squares.

4.3.2 Methods

The proposed workflow was shown in Figure 4.3. Spectral measurements of soil samples were acquired in the well-controlled laboratory and the corresponding soil properties were also retrieved by conventional chemical/physical analysis. A 1D-CNN model as mentioned before was developed based on the soil spectral library and will be used as the base model for further analysis. Sixteen field samples collected in the study area were used to fine-tune the pre-trained 1D-CNN model and the others 16 were for the independent validation. It is pointed out that normalized spectral indices have the potential to be transferred between sensors. Therefore, a spectral index for soil clay is also explored on the large-scale soil spectral library.

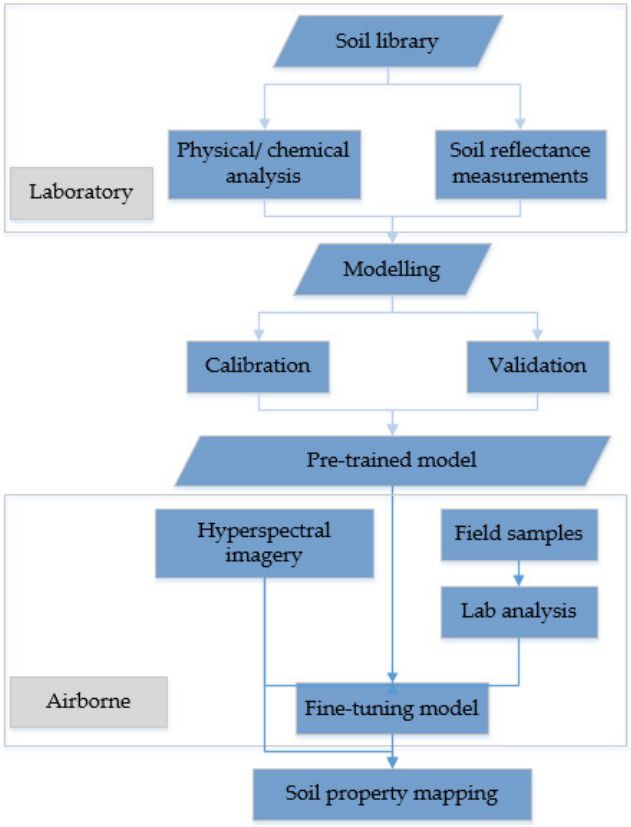

Figure 4.3 Schematic diagram of proposed workflow on transfer learning for soil property mapping.

4.3.2.1 Convolutional neural networks

The CNN is composed of multiple feature generation stages, each of which includes a convolutional layer, a nonlinearity layer and a pooling layer. After several feature generation stages, the CNN is often followed by one or more fully-connected layers and a final classifier layer for classification tasks. In this study, we adopt the CNN for the estimation of soil clay content, which is continuous data instead of categorical data. For example, the clay content values for LUCAS mineral soils range from 0.0 to 79.0%. Therefore, we use a regression layer to replace the final classifier layer. The architecture can be seen in Figure 4.4.

A soil spectrum can be regarded as a 2D image whose height is equal to 1 [152]. Therefore, the size of input layer can be viewed as $n \times 1$, and n is the number of bands. Each convolutional layer contains a number of 1D filter kernels with the size of $k \times 1$, which generate feature maps when applied to the input spectral data. The number of layers, the kernel size and the number of kernels in the convolutional layers are hyperparameters that set manually. In this study, we use four convolutional layers and the number of filter kernel

was set to (32,32,64,64). The size of filter kernel is 3. The weights in the kernels are learnt using the back propagation (BP) algorithm with labelled training dataset. The main benefit is that feature maps used in the classification or regression are learnt from data without any manual feature extraction [153].

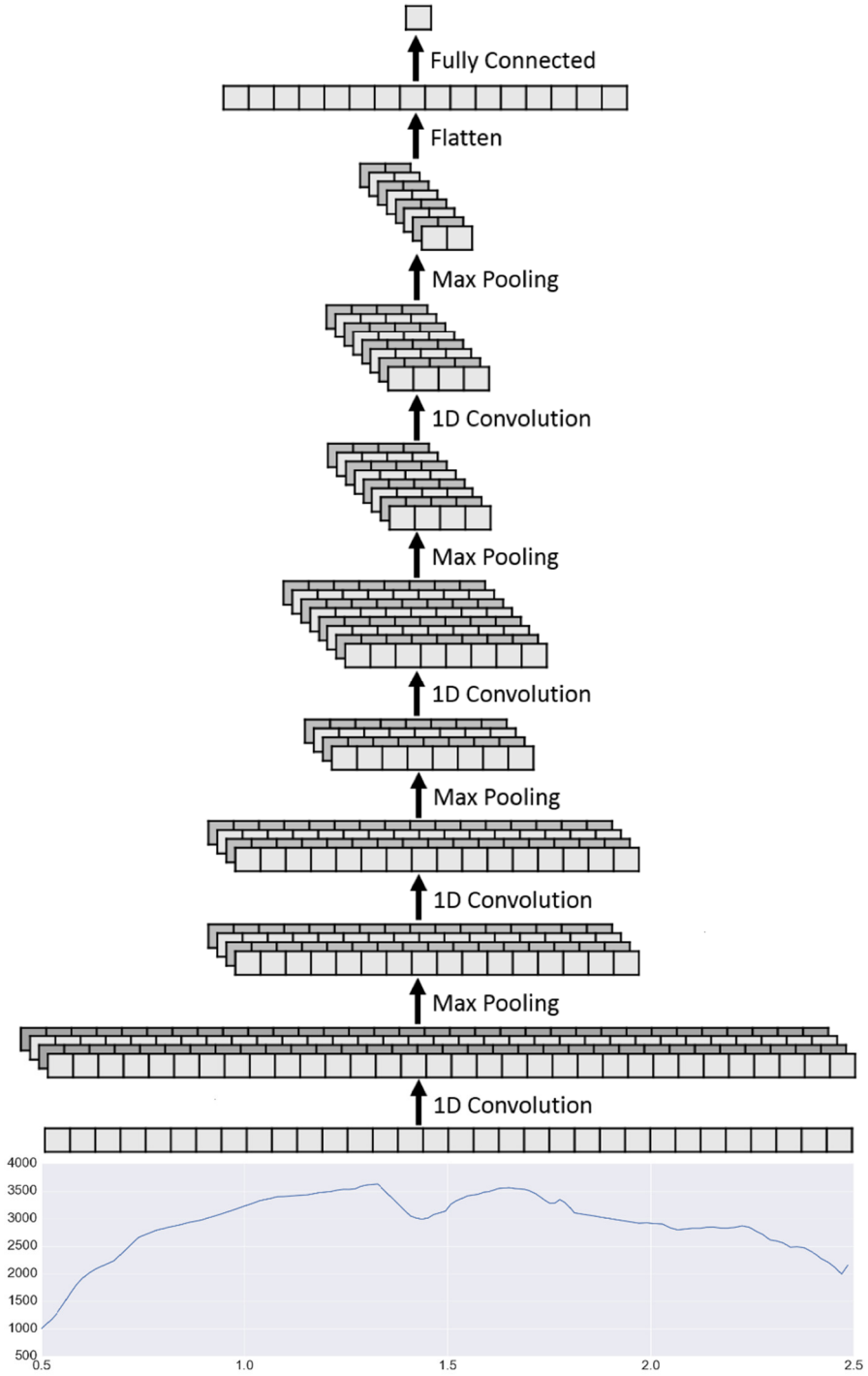

Figure 4.4 The architecture of the CNN for hyperspectral data classification (modified from [154]).

4.3.2.2 Transfer learning based on the pre-trained 1D-CNN model

It is pointed out that there are two ways to apply transfer learning with deep networks [155]. One possibility is to utilize the pre-trained network with the learned weights to obtain features that would be subsequently used in the new problem as shown in Figure 4.6. Feature generation layers, prior to the last fully-connected layer, are frozen and the outputs of the CNN constitute learnt features. Another option is to fine-tune the pre-trained network weights by training the network with the new dataset. As we are trying to make model transferrable between different sensors, the second method was adopted to fine-tune the whole pre-trained CNN model.

The LUCAS data is classified into two categories: mineral and organic soils. We first use mineral soils to build a CNN model as described before and the CNNs typically have a large number of parameters and require a significant amount of training data. We use the spectra extracted from hyperspectral data at the location of field samples and the corresponding soil clay content values to fine-tune the pre-trained CNN model. Finally, the fine-tuned model is applied to the whole hyperspectral image so as to obtain the soil clay content map in the study area.

4.3.2.3 Spectral index for soil clay content

Clay minerals are characterised by absorption features near 2200-2300 nm. The location of the clay absorption peak was identified at 2209 nm with the following two bands representing the shoulders of the absorption peak: 2133 nm and 2225 nm. Using these bands, a short-wave infrared fine particle index (SWIR FI), as shown in Equation (4.3), was proposed by [31] and implemented in ENSOMAP software.

$$\text{SWIR FI} = \frac{(b2133\ nm)^2}{b2225\ nm\ \times (b2209\ nm)^3} \tag{4.1}$$

4.3.3 Assessment

The performance of calibration models for soil clay content was assessed by *RMSE*, R^2 and the ratio of percent deviation (*RPD*), which were calculated by the following equations:

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2} \tag{4.2}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2} \tag{4.3}$$

$$RPD = \frac{SD}{RMSE} \tag{4.4}$$

where, n is the number of validation samples, y is the measured value, $\overline{y}$ is the mean of the measured value, and $\hat{y}$ is the estimated value. *RPD* is the ratio of the standard deviation (*SD*) of the calibration data to the *RMSE* of the validation data. It is commonly used to investigate the prediction error with variation in the data.

4.4 Results and Discussion

4.4.1 Interpretation of mineral and organic soils from LUCAS dataset

The LUCAS dataset contains about 16,000 soil samples classified as mineral soils that were used to train the one-dimensional CNN model. About 660 organic soil samples containing clay information were used to test if CNN model developed by mineral soils is transferrable for organic soils. The histograms of soil clay content distributions of mineral and organic soils were shown in Figure 4.5A,B. Clay contents for mineral and organic soils were skewed forming long tails with only a few samples having values higher than 60%. The average clay content value for organic soils is 15% while for mineral soils is 17%. Organic soils have generally lower clay content as pointed out in Reference [86].

The mean soil reflectance spectra and standard deviations for mineral and organic soils were plotted in Figure 4.5C,D. The mean spectra of both mineral and organic soils have a similar curve shape whose reflectance values increase with increasing wavelength in the range of 500–1300 nm. The main spectral difference is that the mean reflectance spectrum for mineral soils demonstrates a higher albedo than spectra for organic soils as mineral soils have a lower level of SOC content. It is well known that higher levels of organic material lead to darker soils, and soil reflectance decreases with increasing SOC content especially in the spectral range of 600–750 nm as observed in References [30,156].

The mean soil continuum-removal (CR) spectra and standard deviations for mineral and organic soils were also shown in Figure 4.5E,F. CR spectra can be used to isolate and identify

characteristic absorptions of minerals, organic compounds, and water in soils [8]. Both mineral and organic soils showed absorption peaks near 600, 1400, 1900 and 2300 nm. The absorption depths near 600 and 2300 nm for organic soils are much deeper than mineral soils. The highest correlation between double square-root of the SOC content (SOC1/4) and reflectance is found in the visible region, with a maximum around 600 nm [30]. Around 2300 nm (2309 and 2347 nm) are combinations and overtones of the C-H group, which is characteristic of different organic substances [157]. Mineral soils also have an absorption peak near 2200 nm, which is correlated with clay content [158]. Organic soils have absorption peaks near 1720 nm, which correlated with SOC.

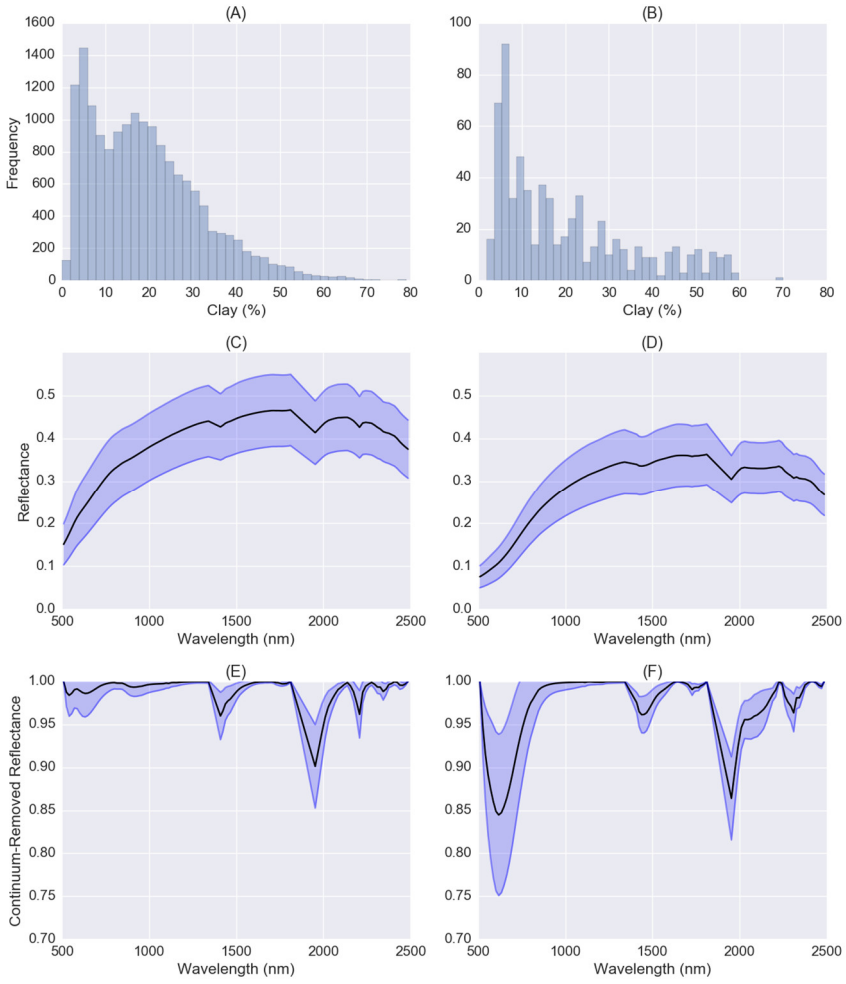

Figure 4.5 (A-B) are histograms of soil clay content distribution of mineral and organic soils; **(C-D)** are mean soil reflectance spectra (black lines) and standard deviations (blue lines, lower and upper boundaries) for mineral and organic soils; **(E-F)** are mean soil continuum-removal spectra (black lines) and standard deviations (blue lines, lower and upper boundaries) for mineral and organic soils. Values are given in reflectance **(C-D)** and normalised continuum-removal values **(E-F)**.

4.4.2 1D-CNN and spectral index for LUCAS soil clay content estimation

In the architecture of 1D-CNN, four convolutional layers were adopted with weights initialised by a uniform distribution. The optimiser is adamax [159] and loss function is mean squared error (MSE) to train the model. (R^2 = 0.834, *RMSE* = 5.31 and *RPD* = 2.42).

Before fine-tuning the pre-trained 1D-CNN model using organic soils, the number of neurons in the fully-connected layer was reduced from 32 to 16 so as to reduce the training parameters. The result is (R^2 = 0.756, *RMSE* = 7.07 and *RPD* = 2.26). We also tried to directly apply the pre-trained 1D-CNN model without fine-tuning and achieved a comparatively poor accuracy (R^2 = 0.378, *RMSE* = 11.29 and *RPD* = 1.42), as shown in Figure 4.6B.

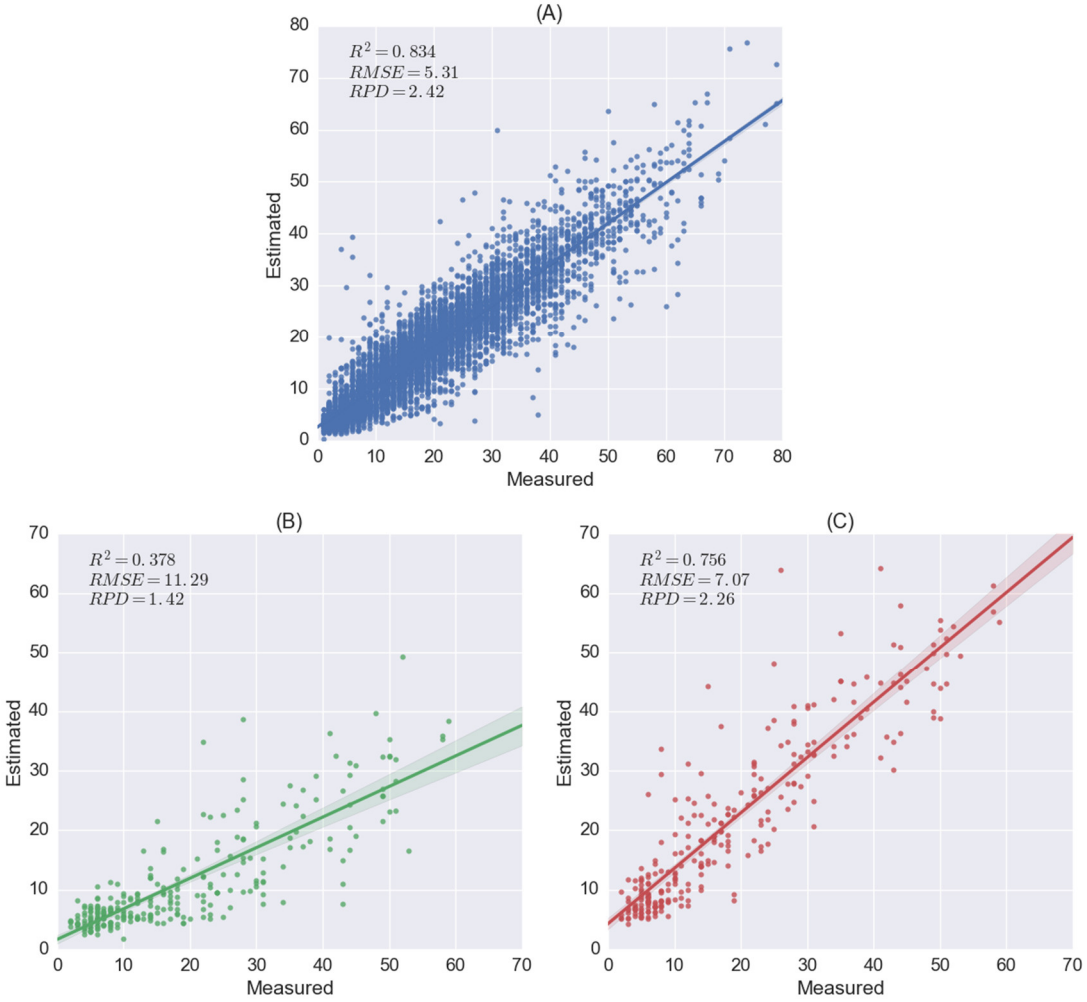

Figure 4.6 Results of soil clay content estimation for LUCAS mineral and organic soils using 1D-CNN and transfer learning. **(A)** is the scatter plot of measured and estimated clay content for mineral soils obtained by 1D-CNN model. **(B)** is for organic soils using the pre-trained 1D-CNN model developed by mineral soils without fine-tuning. **(C)** is for organic soils by fine-tuning the pre-trained 1D-CNN model.

The absorption feature near 2200 nm for the mean spectrum of mineral soils was shown in Figure 4.7A. For mineral, the absorption peak is at 2207 nm which is very close to 2209 nm as adopted in the spectral index of SWIR FI. The depth is 0.971 and the full-width at half-maximum (FWHM) is 30 nm. However, there is no observed absorption feature near 2200 nm for organic soils. Spectral index failed on both mineral and organic test dataset as shown by the scatter plots between SWIR FI and soil clay content values in Figure 4.7B,C, especially for soil samples having clay content values greater than 20%. We also tried to adopt the equation for SWIR FI with bands at 2207, 2140 and 2225 nm for mineral soils but didn't achieve much improvement. Therefore, we only consider transfer learning based on 1D-CNN for the following application with hyperspectral imagery.

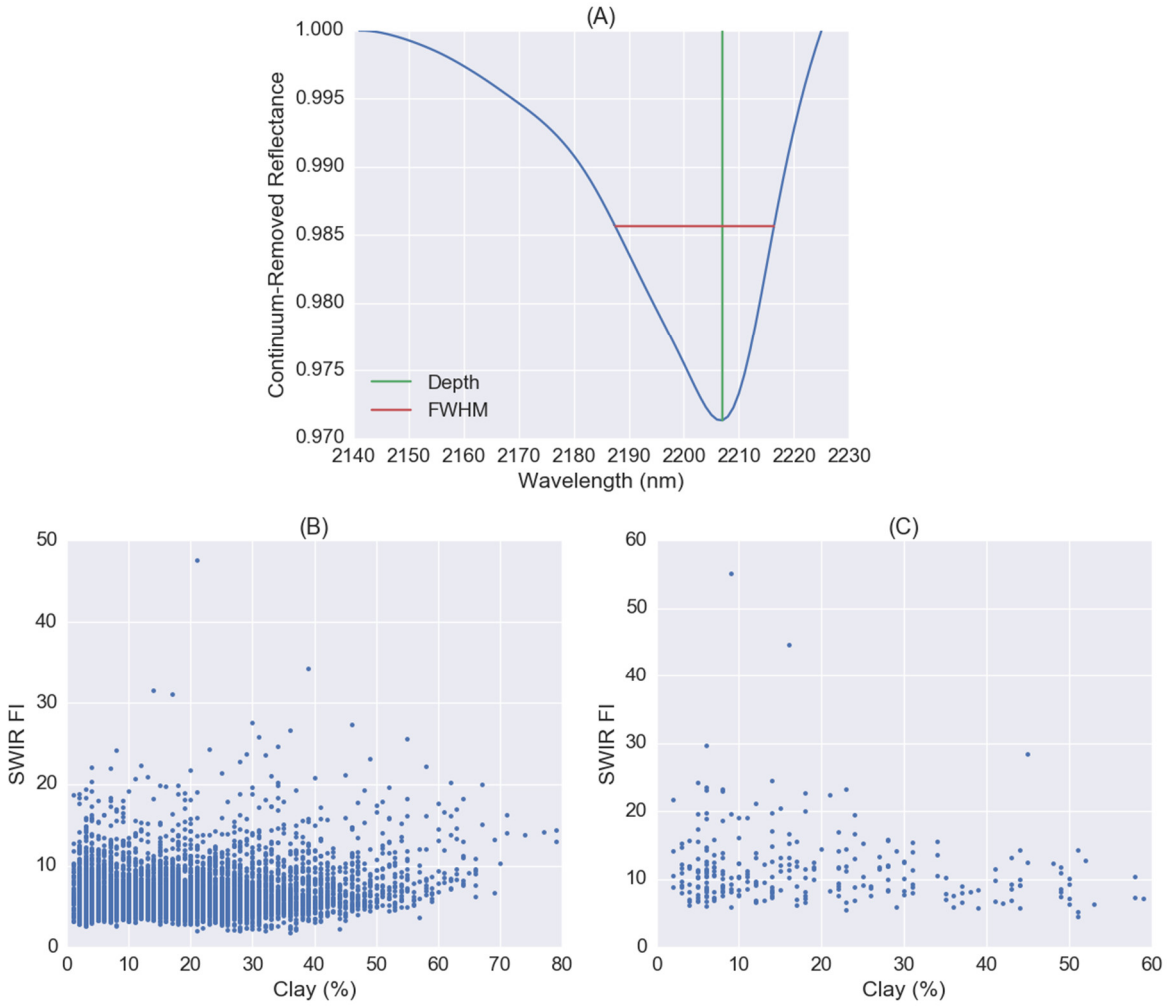

Figure 4.7. Absorption feature near 2200 nm for the mean spectrum of mineral soils **(A)** and scatter plots between soil clay contents and the corresponding SWIR FI values for mineral **(B)** and organic soils **(C)**

4.4.3 Application of transfer learning for soil clay content mapping using the pre-trained 1D-CNN model

The clay content values of field samples vary between 8.4% and 63.4%. The mean soil reflectance spectrum (black line) and standard deviation for spectra extracted from hyperspectral imagery at the locations of field samples were shown in Figure 4.8B. The overall albedo is lower compared to LUCAS mineral or organic soil spectra measured in the laboratory. The mean soil reflectance spectrum (black line) and standard deviation for spectra for CR spectra were shown in Figure 4.8C. The absorption depth near 1400 nm is much deeper than LUCAS soil spectra measured in the laboratory, which is caused by water absorption.

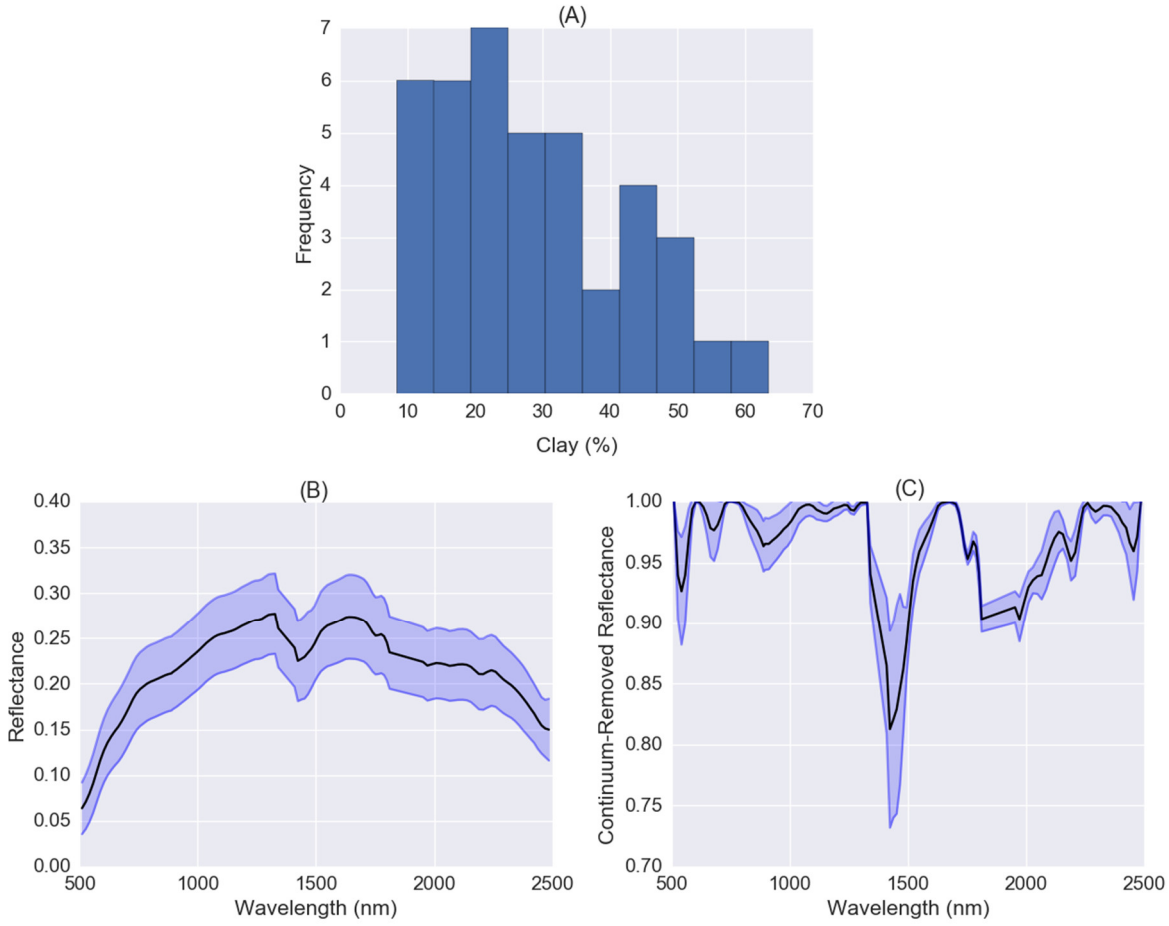

Figure 4.8 (A) histogram of soil clay content distribution of soil samples collected from study area Cabo de Gata-Nijar; **(B)** mean soil reflectance spectrum (black line) and standard deviation (blue lines, lower and upper boundaries) derived from the hyperspectral image; **(C)** mean soil continuum-removal spectrum (black line) and standard deviation (blue lines, lower and upper boundaries).

The pre-trained CNN model was fine-tuned by field samples collected in the study area. The accuracy (R^2=0.601, *RMSE*=8.62 and *RPD*=1.54) is lower than the result obtained from LUCAS organic soils. The fine-tuned model was applied to the whole hyperspectral image except for the non-bared soil pixels. From the histogram of clay content (Figure 4.9B), it can be seen the distribution of soil clay content was also skewed forming long tails and the majority of soil clay values fallen in the range from 10% to 40%. For clay content map (Figure 4.9C), non-bared soil pixel values were set to 0 and clay content values greater than 50% were set to 50%.

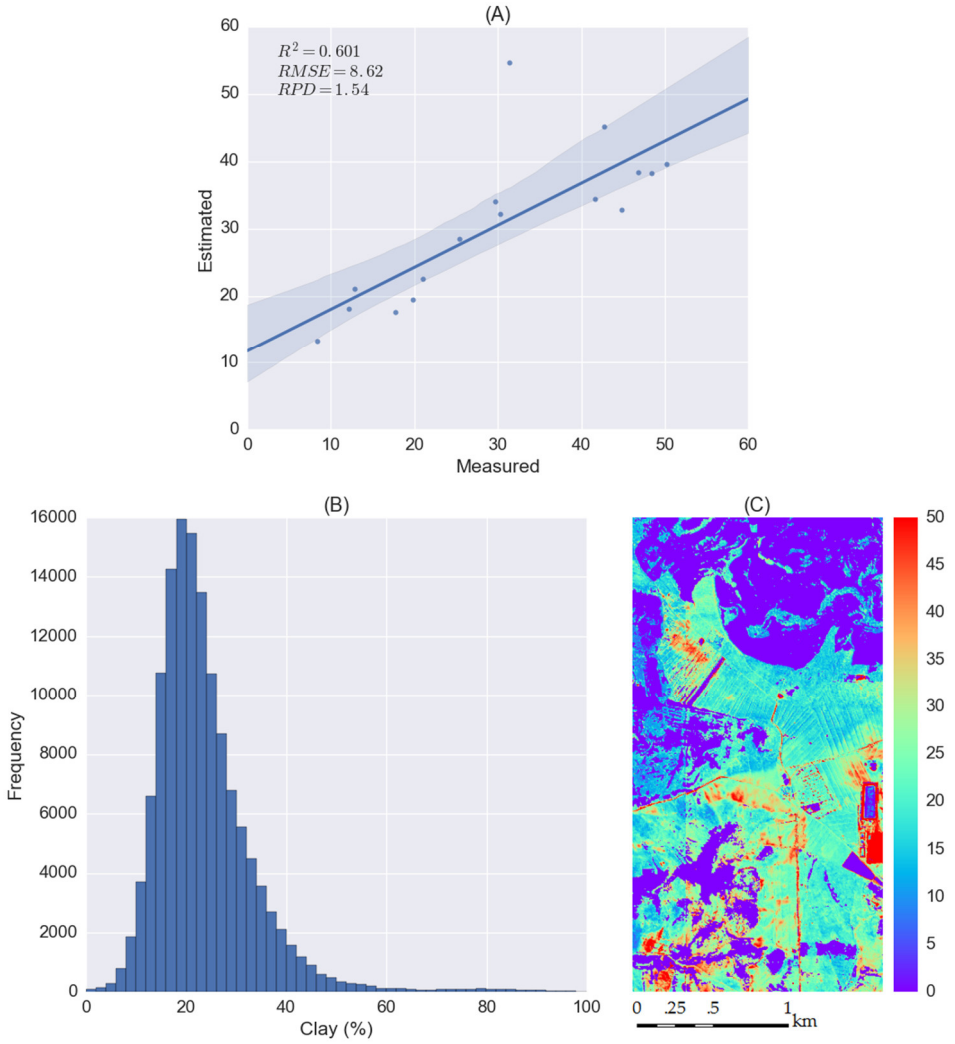

Figure 4.9 Results of transfer learning for soil clay mapping using hyperspectral imagery and the pre-trained CNN model. **(A)** is the scatter plot between measured and estimated clay contents for testing data; **(B)** is the histogram of soil clay content distribution of derived soil clay map without considering masked non-bared soil pixels; **(C)** is the soil clay map in the study area with masked non-bared soil pixel values set to 0.

4.4.4 Comparison between spectral index and transfer learning

Spectral index is a simple and easy implemented algorithm that often only use few bands rather than the full visible near-infrared spectral range. It is particularly efficient in deriving information that relies on the specific spectral response of the targeted object [160]. Although it is suggested spectral index is transferable from one sensor to another, SWIR FI proposed by [31] showed little correlation with the clay content of LUCAS soils, especially for soil samples having clay content higher than 20%. The absorption peak around 2200 nm for mineral soils is slightly different from what observed by [31]. It is pointed out that indices obtained using one instrument could be significantly different from the same indices obtained using other instruments [161]. For organic soils, there is no absorption peak around 2200 nm because of extremely spectral diverse compared with mineral soils. Therefore, it is still difficult to directly use the spectral index for the transferable study of soil properties, especially for different soil categories.

Transfer learning is proposed based on deep learning (DL). With LUCAS mineral soils, the 1D-CNN obtained an accuracy (R^2) of 0.834. Organic and mineral soils from LUCAS data were measured by the same instrument and in well-controlled laboratory. The main difference is the diversity of spectra. For the CNN model, it means the input domain is different. When trying to use the pre-trained 1D-CNN model developed from mineral soils, fine-tuning is required to make the model transferrable from source domain to target domain. By doing that, the R^2 value improved from 0.378 to 0.756. DL provides an end-to-end learning approach with no need for feature engineering. Unlike many prior regression approaches, DL models can be trained on additional data without restarting from scratch, making them viable for continuous learning. Therefore, it is possible to reuse a DL model trained from the large-scale spectral library for local-scale soil property quantification, which makes DL applicable to fairly small datasets. The transferred calibration model obtained an accuracy of 0.601 for soil clay content mapping, which was comparatively lower than achieved by the spectral library. It is pointed out that surface spectral data are generally affected by the confounding effects of soil moisture and soil roughness [162]. Water absorption contributed to the spectral difference between laboratory and airborne hyperspectral data, as shown in Figure 4.8. Soil moisture has a strong influence on the amount and composition of reflected and emitted energy from the soil surface. Most importantly, water absorption features near 1400 and 1950 nm will mask important

spectral information associated with soil variables, including clay [130,163]. A direct standardization (DS) method was proposed to correct the difference between instruments [164] and successfully utilised to reduce the effects of soil moisture and other environmental factors on field Vis–NIR–SWIR spectra [10,165]. For the CNN model, choosing the optimal architecture and training it optimally are still open questions. It is hard to comprehend what is going on under the hood of DL algorithms [40], which could be a problem for non-experts to develop effective DL algorithms or adopt it to different study areas. Besides, it is difficult for CNN to directly incorporate spectral information with other soil properties and location information like support vector machine, random forest and spectrum-based learner, which are very important to improve the estimation accuracies of soil properties. It should be pointed out that the proposed method for soil clay content mapping was only validated on very few samples, because of the limited available dataset, which constrains the generalizability and thus should be further explored by incorporating more soil samples.

4.4.5 Large-scale soil spectral library for digital soil mapping at the local scale using hyperspectral imagery

There are some studies relate to the retrieval of soil properties by taking advantage of large-scale spectral data. The potential of the LUCAS database for the SOC estimation in Belgium and Luxembourg was investigated in Reference [134]. The LUCAS dataset was divided into several classed using a cluster analysis. PLS regression models were calibrated for each class and then adopted to estimate the SOC content on the soil spectra of the calibration datasets of the same class. Soil samples were scanned by the same instrument that used for the LUCAS dataset. The achieved RPD values for the proposed methods were between 1.41 to 2.24. A bottom-up approach was further developed to estimate SOC using hyperspectral imagery [4] and achieved RPD values of 1.7 for Luxembourg data and 1.4 for Belgium data. The PLS regression models developed using the LUCAS dataset were applied to field soil spectra measured in the laboratory instead of hyperspectral imagery. Besides, this approach requires that the large-scale spectral library should contain spectra that closely match those of the local soil samples. For transfer learning, it does not have such a limitation, but it requires a few soil samples to fine-tune the pre-trained model, as demonstrated in the study of transferring the classification model developed using ImageNet to remotely sensed

images [147]. The soil clay content map was generated from airborne hyperspectral data by transferring laboratory regression models with methods of model updating, Repfile, Transfer by Orthogonal Projection (TOP) and Piecewise Direct Standardization (PDS) [6]. Transferred models showed better performance than the laboratory model calibrated without transfer. These methods are used to address factors that cause spectral distortions resulting from the different measurement conditions, while transfer learning is a more general approach to develop a transferrable model instead of aiming to solve the spectra standardization problem. However, the above-mentioned methods, including transfer learning are limited to bare fields as the presence of the vegetation may contribute to the spectral confusion with soil reflectance [166]. Spectral mixture analysis was adopted in Reference [5], to extend the mapping capability up to a vegetation coverage of 40% using a feature-based multiple linear regression model.

Soil property model is often calibrated using field samples collected in the same area, which generally yields the best prediction accuracy. This is because the samples used for calibration are geographically close to the target site and thus are expected to have soil properties and spectral responses that are similar to the target samples [19]. However, it often requires large amounts of field work and many hours or days processing the data. It would be great if the model can take advantage of available existing soil libraries. However, it is pointed out that there are still few studies combing the use of laboratory, proximal, and remote spectroscopic sensing research. One reason might be that there are significant challenges posed by the inherent differences between the standardised laboratory measurements and those made under natural conditions [8]. The signal-to-noise ratio of air- or space-borne hyperspectral data is relatively low compared to laboratory data, due to a low integration time over the target area [158]. The application of imaging spectroscopy is also restricted by atmosphere attenuation, revisiting time, sensor radiometric and spatial resolutions, and BRDF effects. While the effort is putting on reducing the effect of water and other environmental factors, the soil community should also be aware of advancements like DL. Although the model for airborne hyperspectral data was less accurate than the laboratory model, it demonstrated the potential of utilising laboratory spectra and hyperspectral imagery for soil property mapping, and it will continuously benefit from the advancement of DL research.

4.5 Conclusions

In this paper, we investigated the potential of using a pre-trained CNN model for the estimation of soil clay content. The success of DL provides a promising approach to mapping soil properties using hyperspectral data with large-scale soil spectral libraries. A 1D-CNN approach was proposed to the estimation of soil clay content and achieved an accuracy (R^2) of 0.834 with LUCAS mineral soil dataset. The 1D-CNN model was further fine-tuned by soil samples collected in the field with spectra extracted from the hyperspectral imagery. The transferred model obtained an accuracy (R^2) of 0.601 for regional soil clay content mapping. To the best of our knowledge, this is the first case study adopting CNN-based transfer learning for soil spectroscopy. However, the proposed approach was tested only on a limited area, and its application to practice is still open, especially to areas with different soil conditions. Besides, the proposed method is limited to bare soils, and the influence of external factors, including vegetation coverage and soil moisture should be further studied. Although the result obtained by the hyperspectral imagery is still not compatible to laboratory spectroscopy, the CNN-based transfer learning provides a new way to make use of both large-scale spectral libraries and hyperspectral data to map soil properties.

Acknowledgments

The first author wants to express their acknowledgments to the China Scholarship Council (CSC) for providing financial support to study at TU Dresden. The LUCAS topsoil dataset in this work was made available by the European Commission through the European Soil Data Centre and managed by the Joint Research Centre (JRC) http://esdac.jrc.europa.edu/. We acknowledge Sabine Chabrillat and the Helmholtz-Zentrum Potsdam Deutsches GeoForschungsZentrum for providing the Cabo de Gata HyMAP imagery. We acknowledge the University of Almeria and Paula Escribano for providing the Cabo de Gata soil data.

Chapter 5

A Case Study of the Forced Invariance Approach for Soil Salinity Estimation in Vegetation-Covered Terrain Using Airborne Hyperspectral Imagery

Lanfa Liu, Min Ji and Manfred Buchroithner

Contributions:

Lanfa Liu conceived and performed the research and wrote the manuscript.

Min Ji contributed to the design of the research and data analysis.

Manfred Buchroithner supervised the study at all stages and reviewed the manuscript.

Citation:

Liu, L.; Ji, M.; Buchroithner, M. A Case Study of the Forced Invariance Approach for Soil Salinity Estimation in Vegetation-Covered Terrain Using Airborne Hyperspectral Imagery. *ISPRS Int. J. Geo-Inf.* 2018, *7*, 48.

5.1 Abstract

Soil spectroscopy is a promising technique for soil analysis, and has been successfully utilised in the laboratory. When it comes to space, the presence of vegetation significantly affects the performance of imaging spectroscopy or hyperspectral imaging on the retrieval of topsoil properties. The Forced Invariance Approach has been proven able to effectively suppress the vegetation contribution to the mixed image pixel. It takes advantage of scene statistics and requires no specific a priori knowledge of the referenced spectra. However, the approach is still mainly limited to lithological mapping. In this case study, the objective was to test the performance of the Forced Invariance Approach to improve the estimation accuracy of soil salinity for an agricultural area located in the semi-arid region of Northwest China using airborne hyperspectral data. The ground truth data has been obtained from an eco-hydrological wireless sensing network. The relationship between Normalized Difference Vegetation Index (NDVI) and soil salinity is discussed. The results demonstrate that the Forced Invariance Approach is able to improve the retrieval accuracy of soil salinity at a depth of 10 cm, as indicated by a higher value for the coefficient of determination (R^2). Consequently, the vegetation suppression method has the potential to improve quantitative estimation of soil properties with multivariate statistical methods.

5.2 Introduction

In arid and semi-arid areas, soil salinization is one of the major threats to agricultural production, which could be caused by incorrect or careless irrigation [167]. The significant impacts of soil salinity on the soil-water-plant system can reduce the nutrient absorption and lead to a considerable decrease of crop productivity [168,169]. Remote sensing has been shown to be a particularly valuable tool for monitoring soil conditions frequently and spatially [3,170,171]. The presence of salts can be detected directly on bare soils with salt crust via the variation of spectral reflectance, and the spectral behaviour of salt has been studied in detail [172,173]. However, the ability to map soil salinity using the direct approach is limited, especially in agricultural areas [174,175]. The biophysical characteristics of vegetation can serve as an indirect sign of soil salinity, as plants subjected to salinity stress typically have

lower photosynthetic activity, causing increased visible reflectance and reduced near-infrared reflectance from the vegetation. Therefore, various indices have been proposed for assessing and mapping soil salinity, such as the Soil Adjusted Vegetation Index (SAVI), Normalized Difference Salinity Index (NDSI) and Salinity Index (SI) [176–178]. Al-Khaier [179] achieved an accurate detection of soil salinity by a normalized salinity index in bare agricultural soils using ASTER bands 4 (near-infrared) and 5 (short-wave infrared). Additionally, Khan [180] successfully used NDSI with the near-infrared and red bands of the Indian Remote Sensing LISS-II sensor to map soil salinity.

Soil salinity indices usually only take advantage of a few bands, and are suitable for multispectral remote sensing images. A lot of success has been achieved in mapping severely saline areas or differentiating between saline and non-saline soils, but it is still difficult to quantitatively retrieve soil salinity [181]. Hyperspectral remote sensing or imaging spectroscopy provides high-resolution data that contains detailed spectral information of soils, and makes it possible to establish models for quantitative estimation of soil salinity. Imaging spectroscopy can not only be used for geology, water and vegetation applications, but also provide a promising method for obtaining soil properties at the large scale, especially with the new hyperspectral sensors, such as EnMAP, HSUI, and HyspIRI [3,11,34,131,182].

Many factors are constraining the application of imaging spectroscopy in the field or from space, such as low signal-to-noise ratio, atmosphere attenuation, sensor resolution and Bidirectional Reflectance Distribution Functional (BRDF) effects, especially for the thin upper soil layer. Thus, optical remote sensing of soils from large distances is a significant challenge [16,77]. In the agricultural area, one of the main problems is spectral mixing. The vegetation coverage and remains might be presented in the image pixel and contribute to creating spectral confusion with soil reflectance [183,184]. Additionally, spectral absorption and reflection vary according to the type of vegetation. Therefore, removing the effects of vegetation on the soil reflectance spectra is an important research topic.

Spectral Mixture Analysis (SMA) is one of the most common techniques used to reduce the contribution of vegetation and to derive quantitative endmember abundance from hyperspectral data [14]. The HyMap hyperspectral imagery was utilised to characterise and map irrigation-induced soil salinization, and a mixture-tuned matched filter (MTMF) approach was assessed to extract and map spectral endmembers from HyMap imagery [185].

The spectral capabilities of upcoming EnMAP were also evaluated to extract quality endmember classes that contain spectral features related to active photosynthetic vegetation (PV), non-photosynthetic active vegetation (NPV) and bare soil (BS). The estimated spectral cover can be integrated into soil erosion models using the linear unmixing method [186]. Franceschini [14] assessed pixel-fractional cover corresponding to bare soil using the linear unmixing method, and applied it to the prediction of soil properties. The model without taking into account the bare soil fractional cover showed a lower accuracy. SMA approaches often assume that endmember cover fractions contained in image pixels are linearly summed. The sub-pixel cover fraction of each land-cover endmember may be plants, bare soil or other constituents. Therefore, it is required that the observations contain enough information to solve a set of linear equations. These endmembers are usually selected either from the image data or existing spectral libraries [187]. The problem is that referenced spectra for soils are often considered to be stable or unique, and the effects of soil properties on the spectra are not included in the models because they are unknown [188]. The Forced Invariance Approach was proposed by [189] to overcome the effects of vegetation on spectral discrimination of the underlying lithological substrate. It utilises scene statistics and requires no detailed knowledge of the reference spectra of endmembers nor any complex mixing models, and has been successfully applied in archaeology and geological mapping using multispectral and hyperspectral data [189–191]. However, to date, there exist few studies that have analysed whether the Forced Invariance Approach is suitable for soil spectroscopy. The accuracies of soil property estimation in the agricultural area are expected to be improved by vegetation-suppressed spectra without requiring extra field work.

The Forced Invariance Approach is focused on the production of contrast-enhanced colour composite images, which are generally used for further visual analysis and identification of lithological or urban features. Its performance on soil analysis has not been tested yet. The objective of this paper is to explore its feasibility to improve soil salinity estimation in the agricultural area. The data source was limited to airborne hyperspectral images. For the first time, the Forced Invariance Approach was adopted to improve the quantitative estimation of soil salinity at a depth of 4 cm and 10 cm by integrating eco-hydrological wireless sensor network data in an experimental agricultural area [192,193]. The possibility and the performance of vegetation suppression using the Forced Invariance

Approach were discussed, and the results demonstrated that the accuracy of the determination of soil salinity at a depth of 10 cm had been improved. The vegetation suppression method is not only suitable for qualitative analysis, as used in lithological mapping, but also has the potential to improve quantitative estimation of soil properties.

5.3 Material and methods

5.3.1 Study area of Zhangye Oasis

The study area is located in Zhangye Oasis in the middle stream of the Chinese Heihe River Basin (100°04′ E, 39°15′ N). The oasis is located in the Gobi Desert, situated in the arid and semi-arid region of Northwest China (Figure 5.1) [194]. The mean annual precipitation and temperature are 121.5 mm and 6 °C, respectively. Most of the precipitation occurs between July and September. The average annual precipitation varies from 100 to 250 mm, whereas potential annual evaporation ranges from 1200 to 1800 mm, which is ten times higher than the average annual precipitation [195]. Land cover types include wetland, grassland, and farmland. Corn is the main plant in the study area. Irrigation water in the study area is mainly supplied from the middle reaches of the Heihe River. Soil properties (bulk density, texture, and organic content) vary in the study area, and soil samples have been determined to be silt-loam with sand (9–36%), silt (56–81%), and clay (5–19%) [196].

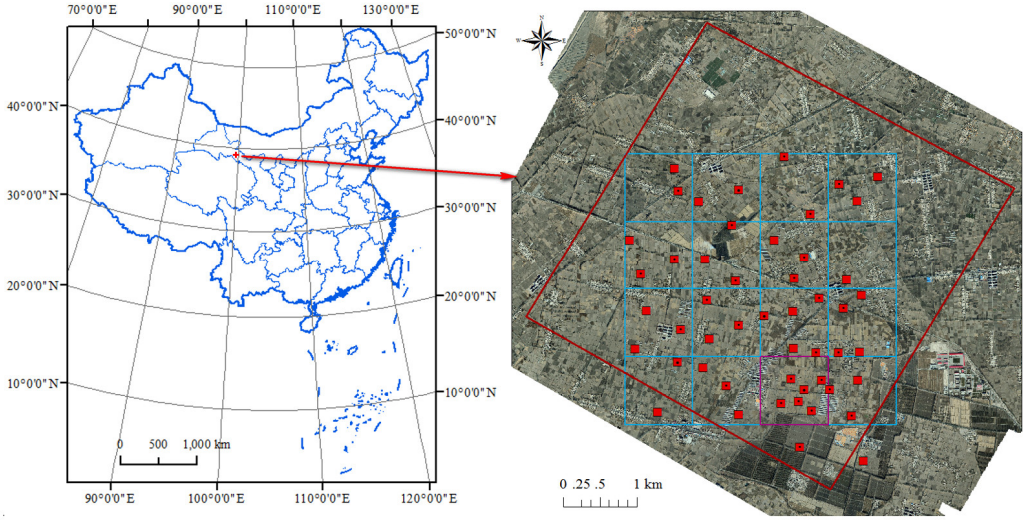

Figure 5.1 Location of study area and the distribution of wireless sensor network nodes.

5.3.2 Data description

5.3.2.1 Eco-Hydrological wireless sensor network data

As part of the eco-hydrological wireless sensor network (WSN), in 2012 48 nodes were installed in the middle stream of the Heihe River Basin, covering both the Yingke and Daman irrigation districts of Zhangye Oasis (Figure 5.2). Data was recorded from the Hydro Probe II sensors [193] every 10 min at two different depths: 4 cm and 10 cm. Recorded information included date and time of reading, soil temperature, soil moisture, electrical conductivity (EC, soil salinity) and soil conductivity. Salinity can be viewed as the total concentration of all dissolved salts in water. Salinity can be measured by a complete chemical analysis called total dissolved solids (TDS), which is difficult and time-consuming. More often, salinity is not measured directly, but is instead derived from the conductivity measurement. There is a high correlation between electrical conductivity (EC) and total dissolved solids (TDS). In this study, we mainly use the EC values from the eco-hydrological wireless network database. The data corresponding to the date of the flight campaign was used to test the performance of the forced invariance method for the estimation of soil salinity in the agricultural area.

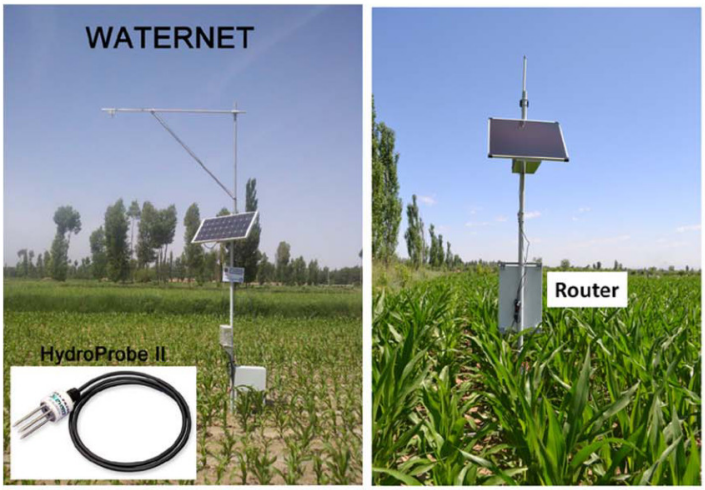

Figure 5.2 Sensor node and router of the wireless sensor network.

5.3.2.2 CASI Airborne hyperspectral data

The flight across the Heihe River Basin was conducted on 29 June 2012 at an altitude of 2000 m above, as part of the Heihe Watershed Allied Telemetry Experimental Research (HiWATER). The Compact Airborne Spectrographic Imager (CASI) 1500 developed by Itres Research Ltd. [197] was used to collect electromagnetic reflectance data. CASI 1500 is a visible

and near-infrared push-broom hyperspectral sensor with 48 spectral bands covering the spectral range from 380 nm to 1050 nm. It has a field of view (FOV) of 40° with 1500 across-track imaging pixels, and the ground spatial resolution is 1.0 m. The radiometric parameter was calibrated in the calibration laboratory of the Institute of Remote Sensing and Digital Earth, Chinese Academy of Sciences, using an integrating sphere as the light source, which was developed by the Labsphere Corporation [198]. The raw data was converted from digital numbers after spectral and radiance calibration and geometrically corrected to a standard earth-centred coordinate system.

5.3.3 Methods

5.3.3.1 Vegetation suppression using the Forced Invariance Method

When transferring soil spectroscopy from laboratory to nature, one of the most significant issues affecting the imaging capability of space-borne and airborne instruments is the presence of vegetation. It can obscure or even completely mask the spectral signatures of the underlying soil information. The Forced Invariance Method was originally developed by Robert Crippen and Ronald Blom (2001) [189]. It is supposed to de-correlate the vegetative component of the total signal on a pixel-by-pixel basis for each band by calculating the relationship of each input band with the vegetation index to overcome the effects of vegetation on spectral discrimination of the underlying lithological substrate. It takes advantage of information from red and near-infrared bands without requiring any specific priori knowledge of the scene. It has been successfully used in many fields using multispectral and hyperspectral data.

In general, the idea is to fit a smooth curve to represent the relationship between the vegetation index and each band's pixel value. By flattening these curves to a target value (such as the mean digital number value of each band), one can expect to remove the correlation with vegetation. The method can be implemented in the following sequential steps [190]: (1) dark pixel correction; (2) vegetation index calculation; (3) estimation of statistical relationship between vegetation index (VI) and digital number (DN) values for each band (Figure 5.3A); (4) calculation of a smooth best-fit curve for the above relationships (Figure 5.3B); and finally, (5) selection of a target average DN value P_{target} and scaling all pixels at each vegetation index

level by an amount that shifts the curve to the target DN. After curve flattening, the new value will be defined by the following equation [199]:

$$P_{new} = P_{original} \times \frac{P_{target}}{P_{NDVI}} \tag{5.1}$$

where P_{new} is the vegetation-suppressed value, $P_{original}$ is the original pixel value and P_{NDVI} is the NDVI corresponding value. By suppressing the vegetation component, it has the potential to reveal not only the underlying geological and archaeological features, but also soil characteristics.

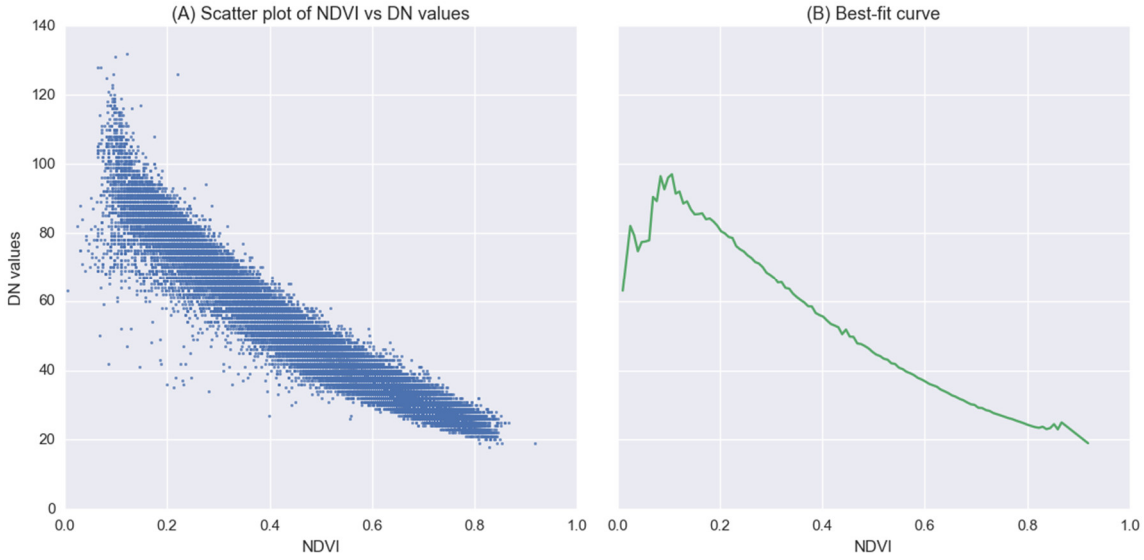

Figure 5.3 Scatter plot **(A)** and best-fit curve **(B)** of Normalized Difference Vegetation Index (NDVI) and digital number (DN) values.

The Forced Invariance Approach is based on the assumptions that 1) the distribution of vegetation across the terrain is independent of rock type and 2) rock albedo is not substantially correlated with the vegetation amount. In our case, this means that soil properties should have no or little correlation with the vegetation index, and that is why this approach has the potential to separate the contribution of vegetation from the target pixels. NDVI was chosen as the vegetation index in the Forced Invariance Approach because it varies much more with vegetation vitality than with variations in lithological variables. Therefore, to check if the approach can be applied to soil analysis, the correlation between NDVI and soil salinity should be examined. Soil moisture is also a major concern for agriculture. Engstrom (2008) [200] already pointed out that the correlation between soil moisture and NDVI was not significant in areas with little to no relief.

5.3.3.2 Spectral modelling of soil properties

The relationship between spectra extracted from the hyperspectral image and soil properties was analysed using the Generalized Linear Model (GLM). The GLM is a flexible generalisation of ordinary linear regression that allows for response variables that have error distribution models other than a normal distribution. The study results from Yuan Huang [201] show that soil moisture, EC and clay content were log-normally distributed, while organic carbon, sand and silt content were normally distributed. Therefore, the Logit Link Function was chosen to model the correlation between spectral data and soil salinity in this study.

Each pixel spectrum of the hyperspectral image comprehends a total of 48 bands, which would cause redundancy of information. Minimum Noise Fraction (MNF) is one of the most common methods to extract features from hyperspectral data, and can effectively reduce a large dataset into a smaller number of components that contain the majority of information. Therefore, MNF transform was performed to the mosaicked and subtracted airborne hyperspectral data using the ENvironment for Visualizing Images (ENVI) software. The data acquired by the vegetation suppression method was also transformed by MNF. The first 14 MNF components were retained as the input variables.

5.3.4 Model performance assessment

For each soil property, the soil spectral quantitative model was developed on a random sample of two-thirds of the data using the GLM. The calibrations were tested by predicting the soil salinity (EC) on validation data sets composed of the remaining one-third of samples. The model accuracies were evaluated on estimated and measured soil salinity using *RMSE* and R^2.

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2}{\sum_{i=1}^{n}(Y_i - \overline{Y})^2} \tag{5.2}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2} \tag{5.3}$$

where n is the number of validation samples, y represents the measured values, $\overline{y}$ is the mean of the measured values, and $\hat{y}$ is the estimated values.

5.4. Results and Discussion

5.4.1 Correlation between NDVI and soil salinity

The terrain in Zhangye Oasis is relatively flat. The correlation between NDVI and soil salinity at the depths of 4 cm and 10 cm are shown in Figure 5.4. The deviation from the fitted line demonstrated that NDVI basically has little correlation with soil salinity either at the depth of 10 cm or 4 cm. The Pearson values between NDVI and soil salinity were also calculated. The correlation at the depth of 4 cm has a slightly higher value (r=0.042) than at the depth of 10 cm (r=0.032).

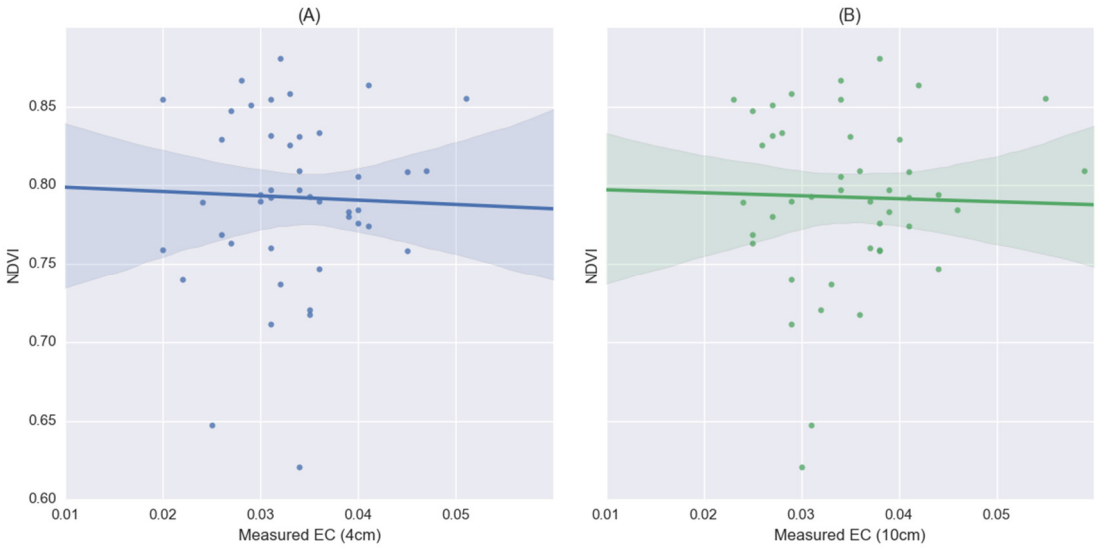

Figure 5.4 Correlation between NDVI and soil salinity (EC) at the depths of 4 cm **(A)** and 10 cm **(B)**.

5.4.2 Vegetation suppression performance using the Forced Invariance Approach

The vegetation cover, which is mainly corn in the study area, could hinder the acquisition of spectral signatures of the underlying soil information. The Forced Invariance Method is assumed to be applicable to the suppression of vegetation. From Section 3.1, we know that it is possible to take advantage of this method to enhance the soil information from the mixed spectra. To check the performance of the vegetation suppression method, the easiest way is to check the true colour image (false colour image is an alternative way) with the naked eye. It can be seen that, while the original image (Figure 5.5A) is dominated by vegetation, the green hue is not so obvious in the processed image (Figure 5.5B), and the latter one also shows some bare soil spots. Another approach is to take advantage of the NDVI which is one of the most

useful vegetation indexes. By comparing Figure 5.6A and Figure 5.6B it can be seen that the NDVI values are also significantly reduced.

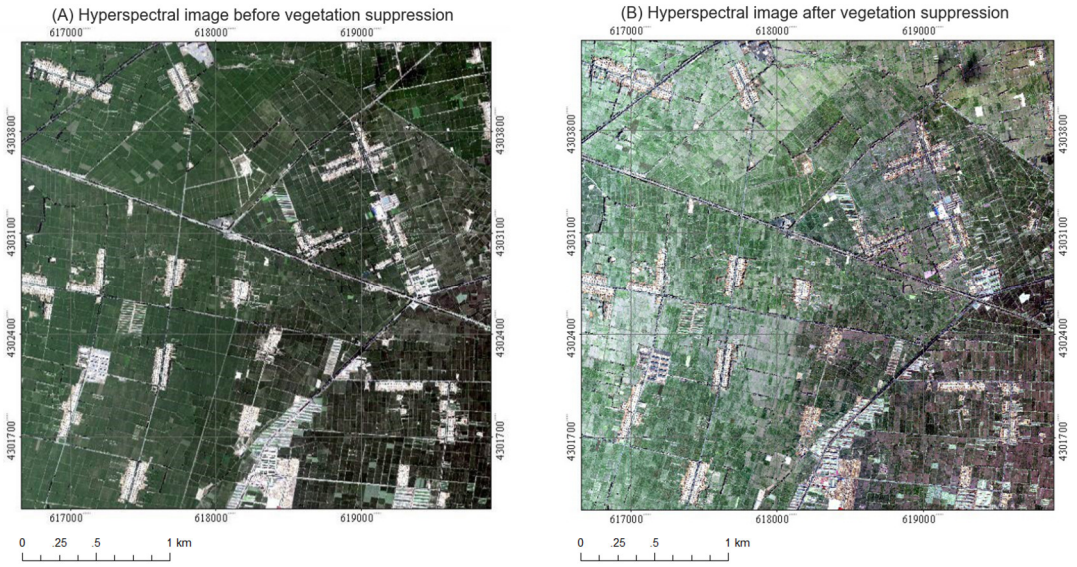

Figure 5.5 Comparison of airborne hyperspectral true colour images (R: 640.5 nm, G: 554.7 nm B: 468.7 nm) before **(A)** and after vegetation suppression **(B)**.

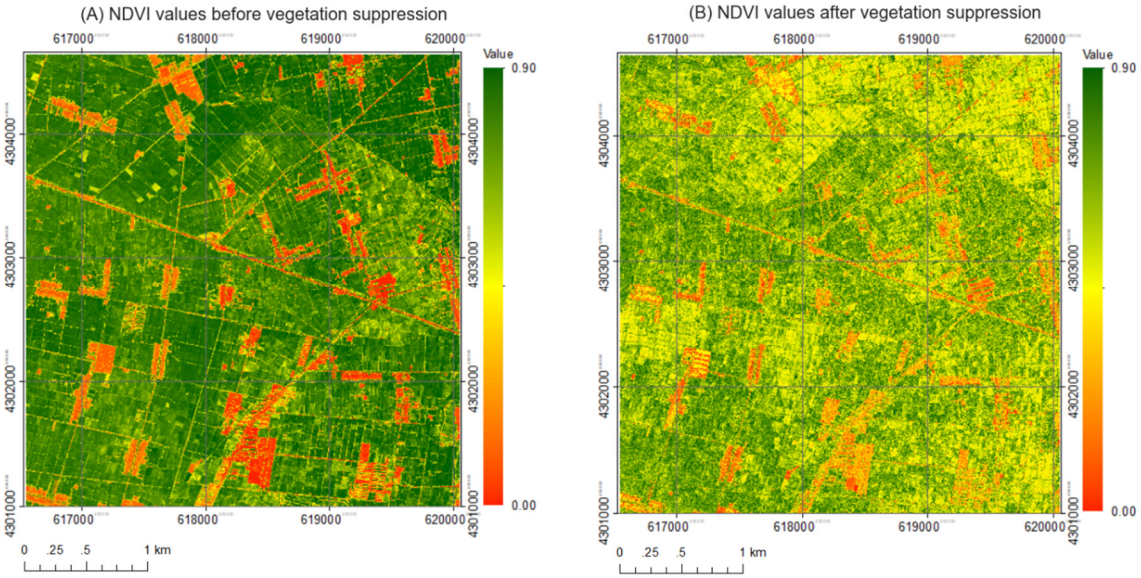

Figure 5.6 Comparison of NDVI values of hyperspectral data before **(A)** and after vegetation suppression **(B)**.

As the dataset used in this study represents airborne hyperspectral imagery, we can further examine the effect of the Forced Invariance Approach using spectral lines. The corn and bare soil spectra measured by ASD Field Spec3 (obtained from Heihe Plan Science Data Centre) were taken as pure endmembers. The acquired spectra were compared to the spectra extracted from hyperspectral images at the pixel corresponding to sensor node 06 before and after vegetation suppression. The spectra comparison is shown in Figure 5.7. The soil

spectrum has no obvious absorption features. Although the spectrum from hyperspectral imagery at the specified pixel after vegetation suppression still has a similar shape with corn spectrum, the slop of "red edge" was reduced in height.

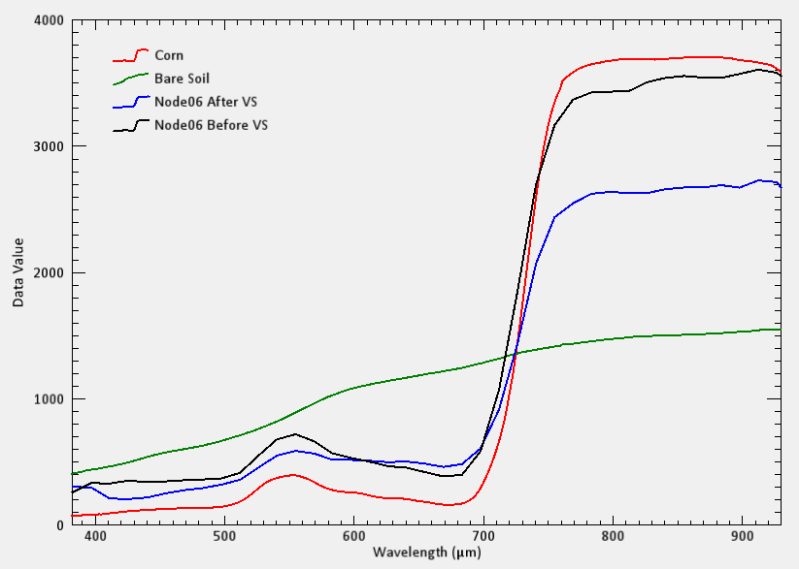

Figure 5.7 Comparison between measured corn and bare soil spectra and the spectra at the location of the specified sensor node from hyperspectral images before and after vegetation suppression.

5.4.3 Estimation of soil properties using airborne hyperspectral data

To quantitatively evaluate the performance of the Forced Invariance Approach using airborne hyperspectral data for agriculture, the relationship between soil spectra and soil salinity were modelled using the GLM. The accuracy for soil salinity at the depth of 10 cm (R^2=0.458) is slightly higher than at the depth of 4 cm (R^2=0.445) using hyperspectral data without vegetation suppression (Figure 5.8), which is more obvious for results obtained from data with vegetation suppression (Figure 5.9). The reason is that surface soil is significantly influenced by exterior factors like irrigation and wind, and landscape fragmentation and complicated cultivation structure also contribute to the high spatial heterogeneity of the soil properties. Therefore, it is less stable and more heterogeneous at the depth of 4 cm than soil at 10 cm.

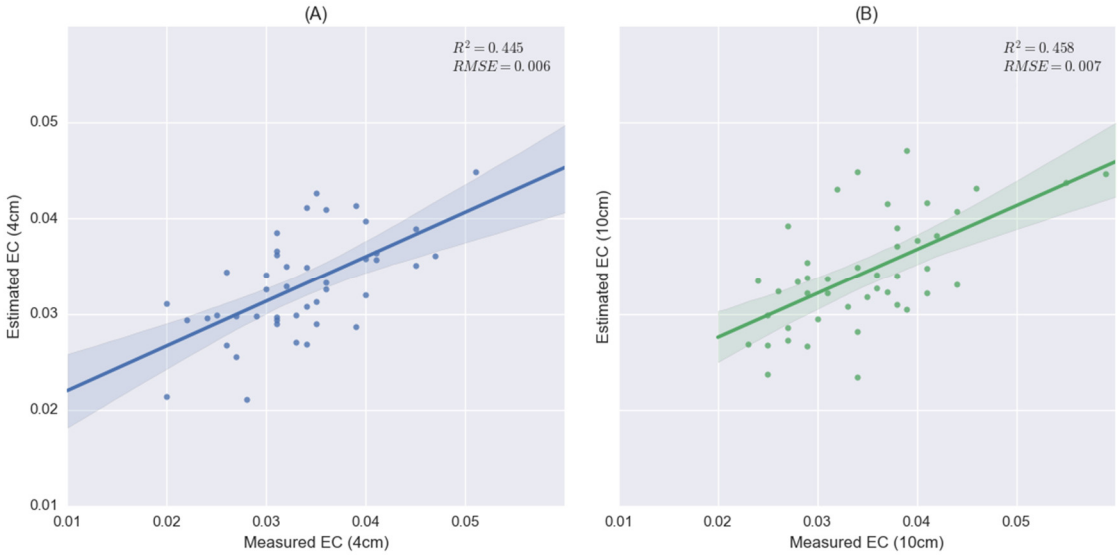

Figure 5.8 Regression plots between measured target values and estimated values before vegetation suppression for soil salinity at the depth of 4 cm **(A)** and 10 cm **(B)**.

By comparing Figure 5.8 and Figure 5.9, it can be seen that the accuracies for the estimation of soil salinity at the depth of 10 cm (R^2=0.538) improved significantly after applying the Forced Invariance Approach, but not like at the depth of 4 cm (R^2=0.43). Apart from the high spatial heterogeneity of surface soil properties, it might also be caused by the correlation to the NDVI. Although the correlation of soil salinity to NDVI was not significant, as revealed by Figure 5.4, soil properties at a depth of 4 cm still show a higher correlation value than at a depth of 10 cm. The modelling results showed that this approach performed better for soil salinity at the depth of 10 cm, which is in agreement with the assumption that the target property should have no or little correlation with the vegetation index. However, it does not guarantee that the model's accuracy will be improved with the increase of soil depth due to the limited effective penetration depth of optical sensors.

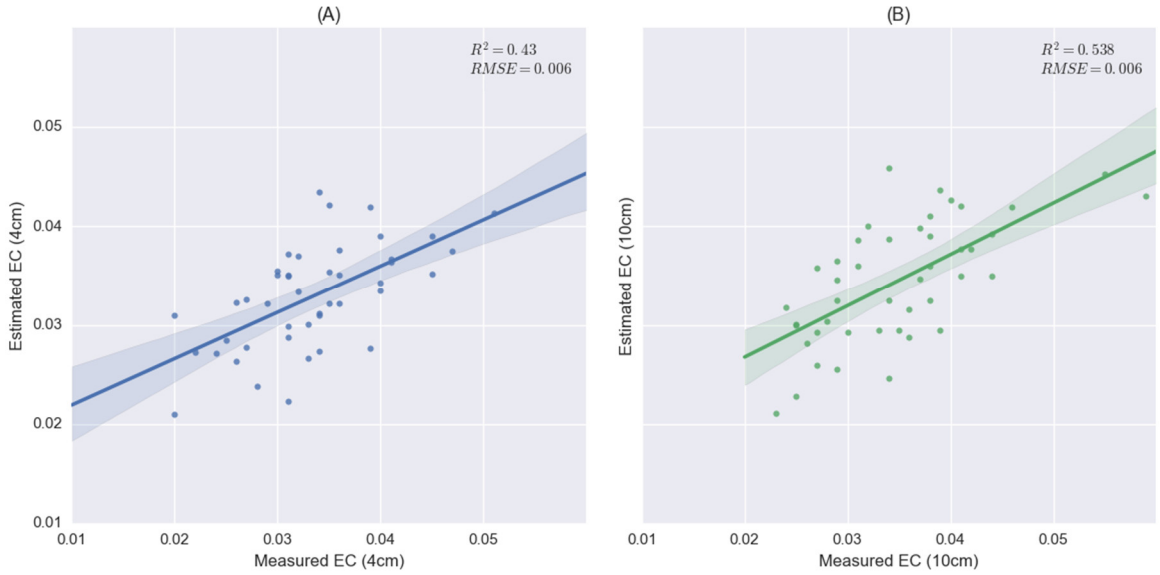

Figure 5.9 Regression plots between measured target values and estimated values after vegetation suppression for soil salinity at the depth of 4 cm **(A)** and 10 cm **(B)**.

5.5. Conclusion

The spatial distribution of soil salinity has important implications for soil and water resource management in arid and semi-arid agricultural regions. The present study examines the possibility to improve the estimation accuracy of soil salinity at different soil depths using imaging spectroscopy and vegetation suppression based on the Forced Invariance Approach which calculates images that are invariant relative to a specific spectral index, and where features represented by that spectral index will not appear in the resulting images because those features will not contribute to the variance.

The relationship between NDVI and soil salinity in the study area indicates that there exists no significant correlation. The GLM developed using wireless network data and airborne hyperspectral data shows a better performance for soil salinity estimation at the depth of 10 cm than at 4 cm, and to the estimation accuracy (R^2=0.538) for soils at the depth of 10 cm after vegetation suppression improved when compared to the result (R^2=0.458) obtained from the model built using hyperspectral data without vegetation suppression. However, the approach failed for soils at the depth of 4 cm. Hence, one should check carefully before applying the Forced Invariance Approach to improve quantitative soil analysis. Besides, the main drawback of the vegetation suppression algorithm is a severe distortion of the spectral

values in non-vegetated areas. The masking technique should be considered in the mapping procedure to keep pixel values from bare soil or sparse vegetation unchanged. The presence of vegetation restrains the application of hyperspectral imagery in retrieving underlying soil properties. The Forced Invariance Approach cannot only produce contrast-enhanced colour composite images for lithological mapping but also has the potential to contribute to the retrieval of soil properties with multivariate statistical methods.

Acknowledgments

The first author acknowledges the Chinese Scholarship Council (CSC) for funding his study at TU Dresden. The dataset has been provided by "Heihe Plan Science Data Centre, National Natural Science Foundation of China" (http://www.heihedata.org). We acknowledge support by the German Research Foundation and the Open Access Publication Funds of the TU Dresden.

Chapter 6

Conclusions and Outlook

Recently significant advances have been made in the application of visible and near-infrared shortwave infrared spectroscopy applied to soil analysis. It has been demonstrated to be a fast and cheap tool for estimating various soil chemical and physical properties. Many efforts have been put on the development of regional, continental and even global soil spectral libraries and memory-based approaches were studied for large-scale data along with common approaches. In this thesis, several methods for extracting features from reflectance spectra were presented and transfer learning was proposed to make laboratory data useful for soil clay content mapping using the hyperspectral imagery collected under natural conditions by fine-tuning a pre-trained 1D-CNN model.

Previous studies by various authors showed that PLS regression is a valid statistical approach for the soil spectral analysis. However, its role in soil spectral feature extraction has long been ignored. In this study, PLS-derived components performed well with three soil categories of LUCAS data (woodland, grassland, and cropland). The combined PLS-GBDT approach yielded a better performance than PLS or GBDT alone. GBDT is a well-known machine learning algorithm that uses the decision tree as the weak learner. However, its capability to handle high-dimensional data is limited. Both PLS and GBDT have the capability to estimate the contributions of input variables. The determination of the varying importance of spectral bands as demonstrated by the PLS method turned out to be a useful tool to retain

target-related information to quantitatively retrieve soil properties like SOC, N and clay in this study.

A new fractal-based feature extraction method was proposed and performed well with LUCAS organic soils. The variogram estimator showed a slightly better performance than the other two estimators (madogram and rodogram) when applied to fractal feature generation for soil property estimation. Step–window pairs had a significant impact on estimation accuracies of soil properties and a hyper-parameter optimisation method was suggested to tune the parameters. Fractal analysis can be used as an approach to characterise statistical self-similarity and further quantify the irregularity of soil spectra. Fractal features, by taking advantage of fractal information encoded in the form of soil spectral curves, can reflect the impact of various properties on soil spectra except when the properties have less direct spectral response. Besides, the proposed fractal method cannot only reduce the dimensionality in the original space but also simultaneously maintain the spectral shape.

Deep learning provides a promising approach to map soil properties using hyperspectral data with existing large-scale soil spectral libraries. A 1D-CNN model for soil clay content estimation was developed using LUCAS mineral soils with an accuracy of R^2=0.834, *RMSE*=5.31 and *RPD*=2.42, which demonstrated that 1D-CNN is an effective method for soil property estimation. The pre-trained model was fine-tuned by field samples collected in the study area with spectra extracted from HyMap imagery, which achieved an accuracy of R^2=0.601, *RMSE*=8.62 and *RPD*=1.54. The fine-tuned model was then applied to bare soil pixels of the imagery resulting in a soil clay map. Although the results are still not yet comparable with laboratory spectroscopy, it provides a way to make use of both large-scale spectral libraries and hyperspectral data.

With feature extraction, the models directly using the whole large-scale dataset achieved good performance on the quantification of multiple soil properties. However, it should be mentioned that memory-based methods are comparatively better suitable for such large-scale soil spectral libraries than global approaches as pointed out by L. Ramirez-Lopez etc. [24]. For each unknown soil spectrum, it is possible to sample a desired number of spectra from the library to build a local model. Furthermore, the sampled small dataset can be used to fine-tune the pre-trained CNN model built using the whole dataset, in which way, the CNN is able

to be combined with MBL so as to take advantage of the information contained within not only local but also global data.

The re-use of existing laboratory soil spectral databases in model development would certainly save time and money. However, there are still limited studies using models built from laboratory spectra to estimate soil properties on hyperspectral imagery. Many factors including instrument properties, experimental conditions and target characteristics restrict laboratory spectroscopic models to be adaptive to air- or space-borne spectral data. Vegetation is also a significant issue in non-bare soil regions. With the rapid development of deep learning, it is possible to transfer models from laboratory data to hyperspectral imagery with transfer learning. Only few studies have so far focused on deep learning applications in soil spectroscopy. Besides, efforts should also be put on reducing the spectral differences between image and laboratory data so that the model can be easily transferred from one sensor to another.

Bibliography

1. Soriano-Disla, J. M.; Janik, L. J.; Viscarra Rossel, R. A.; MacDonald, L. M.; McLaughlin, M. J. The performance of visible, near-, and mid-infrared reflectance spectroscopy for prediction of soil physical, chemical, and biological properties. *Appl. Spectrosc. Rev.* **2014**, *49*, 139–186.

2. Ben-Gear, I.; Norris, K. H. Direct spectrophotometric determination of fat and moisture in meat products. *J. Food Sci.* **1968**, *33*, 64–67.

3. Nocita, M.; Stevens, A.; van Wesemael, B.; Aitkenhead, M.; Bachmann, M.; Barthès, B.; Ben-Dor, E.; Brown, D. J.; Clairotte, M.; Csorba, A.; Dardenne, P.; Demattê, J. A. M.; Genot, V.; Guerrero, C.; Knadel, M.; Montanarella, L.; Noon, C.; Ramirez-lopez, L.; Robertson, J.; Sakai, H.; Soriano-Disla, J. M.; Shepherd, K. D.; Stenberg, B.; Towett, E. K.; Vargas, R.; Wetterlind, J. Soil spectroscopy: An alternative to wet chemistry for soil monitoring. *Adv. Agron.* **2015**, *132*, 139–159.

4. Castaldi, F.; Chabrillat, S.; Jones, A.; Vreys, K.; Bomans, B.; van Wesemael, B. Soil organic carbon estimation in croplands by hyperspectral remote APEX data using the LUCAS topsoil database. *Remote Sens.* **2018**, *10*, 153.

5. Bayer, A. D.; Bachmann, M.; Rogge, D.; Andreas, M.; Kaufmann, H. Combining field and imaging spectroscopy to map soil organic carbon in a semiarid environment. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2016**, *9*, 3997–4010.

6. Nouri, M.; Gomez, C.; Gorretta, N.; Roger, J. M. Clay content mapping from airborne hyperspectral Vis-NIR data by transferring a laboratory regression model. *Geoderma* **2017**, *298*, 54–66.

7. Ben-Dor, E. Quantitative remote sensing of soil properties. *Adv. Agron.* **2002**, *75*, 173–243.

8. Viscarra Rossel, R. A.; Behrens, T.; Ben-Dor, E.; Brown, D. J.; Demattê, J. A. M.; Shepherd, K. D.; Shi, Z.; Stenberg, B.; Stevens, A.; Adamchuk, V.; Aïchi, H.; Barthès, B. G.; Bartholomeus, H. M.; Bayer, A. D.; Bernoux, M.; Böttcher, K.; Brodský, L.; Du, C. W.; Chappell, A.; Fouad, Y.; Genot, V.; Gomez, C.; Grunwald, S.; Gubler, A.; Guerrero, C.;

Hedley, C. B.; Knadel, M.; Morrás, H. J. M.; Nocita, M.; Ramirez-lopez, L.; Roudier, P.; Campos, E. M. R.; Sanborn, P.; Sellitto, V. M.; Sudduth, K. A.; Rawlins, B. G.; Walter, C.; Winowiecki, L. A.; Hong, S. Y.; Ji, W. A global spectral library to characterize the world's soil. *Earth-Science Rev.* **2016**, *155*, 198–230.

9. Mouazen, A. M.; Maleki, M. R.; De Baerdemaeker, J.; Ramon, H. On-line measurement of some selected soil properties using a VIS-NIR sensor. *Soil Tillage Res.* **2007**, *93*, 13–27.

10. Ji, W.; Li, S.; Chen, S.; Shi, Z.; Viscarra Rossel, R. A.; Mouazen, A. M. Prediction of soil attributes using the Chinese soil spectral library and standardized spectra recorded at field conditions. *Soil Tillage Res.* **2016**, *155*, 492–500.

11. Guanter, L.; Kaufmann, H.; Segl, K.; Foerster, S.; Rogass, C.; Chabrillat, S.; Kuester, T.; Hollstein, A.; Rossner, G.; Chlebek, C.; Straif, C.; Fischer, S.; Schrader, S.; Storch, T.; Heiden, U.; Mueller, A.; Bachmann, M.; Mühle, H.; Müller, R.; Habermeyer, M.; Ohndorf, A.; Hill, J.; Buddenbaum, H.; Hostert, P.; van der Linden, S.; Leitão, P.; Rabe, A.; Doerffer, R.; Krasemann, H.; Xi, H.; Mauser, W.; Hank, T.; Locherer, M.; Rast, M.; Staenz, K.; Sang, B. The EnMAP spaceborne imaging spectroscopy mission for earth observation. *Remote Sens.* **2015**, *7*, 8830–8857.

12. Goetz, A. F.; Vane, G.; Solomon, J. E.; Rock, B. N. Imaging spectrometry for earth remote sensing. *Science* **1985**, *228*, 1147–1153.

13. Green, R. O.; Eastwood, M. L.; Sarture, C. M.; Chrien, T. G.; Aronsson, M.; Chippendale, B. J.; Faust, J. A.; Pavri, B. E.; Chovit, C. J.; Solis, M.; Olah, M. R.; Williams, O. Imaging spectroscopy and the Airborne Visible/Infrared Imaging Spectrometer (AVIRIS). *Remote Sens. Environ.* **1998**, *65*, 227–248.

14. Franceschini, M. H. D.; Demattê, J. A. M.; da Silva Terra, F.; Vicente, L. E.; Bartholomeus, H.; de Souza Filho, C. R. Prediction of soil properties using imaging spectroscopy: Considering fractional vegetation cover to improve accuracy. *Int. J. Appl. Earth Obs. Geoinf.* **2015**, *38*, 358–370.

15. Steinberg, A.; Chabrillat, S.; Stevens, A.; Segl, K.; Foerster, S. Prediction of common surface soil properties based on Vis-NIR airborne and simulated EnMAP imaging

spectroscopy data: Prediction accuracy and influence of spatial resolution. *Remote Sens.* **2016**, *8*, 613.

16. Ben-Dor, E.; Chabrillat, S.; Demattê, J. A. M.; Taylor, G. R.; Hill, J.; Whiting, M. L.; Sommer, S. Using imaging spectroscopy to study soil properties. *Remote Sens. Environ.* **2009**, *113*, S38–S55.

17. Nocita, M.; Stevens, A.; van Wesemael, B.; Brown, D. J.; Shepherd, K. D.; Towett, E.; Vargas, R.; Montanarella, L. Soil spectroscopy: An opportunity to be seized. *Glob. Chang. Biol.* **2014**, 10–11.

18. Zeng, R.; Zhao, Y.; Li, D.; Wu, D.; Wei, C.; Zhang, G. Selection of "local" models for prediction of soil organic matter using a regional soil vis-nir spectral library. *Soil Sci.* **2016**, *181*, 13–19.

19. Orgiazzi, A.; Ballabio, C.; Panagos, P.; Jones, A.; Fernández-Ugalde, O. LUCAS Soil, the largest expandable soil dataset for Europe: A review. *Eur. J. Soil Sci.* **2017**.

20. Viscarra Rossel, R. A.; Webster, R. Predicting soil properties from the Australian soil visible-near infrared spectroscopic database. *Eur. J. Soil Sci.* **2012**, *63*, 848–860.

21. Brodský, L.; Klement, A.; Penížek, V.; Kodešová, R.; Boruvka, L. Building soil spectral library of the Czech soils for quantitative digital soil mapping. *Soil Water Res.* **2011**, *6*, 165–172.

22. Romero, D. J.; Ben-Dor, E.; Demattê, J. A. M.; Souza, A. B. e; Vicente, L. E.; Tavares, T. R.; Martello, M.; Strabeli, T. F.; da Silva Barros, P. P.; Fiorio, P. R.; Gallo, B. C.; Sato, M. V.; Eitelwein, M. T. Internal soil standard method for the Brazilian soil spectral library: Performance and proximate analysis. *Geoderma* **2018**, *312*, 95–103.

23. Masri, D.; Masri, D.; Woon, W. L.; Aung, Z. Soil property prediction: An extreme learning machine approach. In *International Conference on Neural Information Processing*; 2015; pp. 18–27.

24. Ramirez-Lopez, L.; Behrens, T.; Schmidt, K.; Stevens, A.; Demattê, J. A. M.; Scholten, T. The spectrum-based learner: A new local approach for modeling soil vis-NIR spectra of complex datasets. *Geoderma* **2013**, *195*, 268–279.

25. Brown, D. J.; Shepherd, K. D.; Walsh, M. G.; Dewayne Mays, M.; Reinsch, T. G. Global soil characterization with VNIR diffuse reflectance spectroscopy. *Geoderma* **2006**, *132*, 273–290.

26. Stenberg, B.; Viscarra Rossel, R. A.; Mouazen, A. M.; Wetterlind, J. Visible and near infrared spectroscopy in soil science. *Adv. Agron.* **2010**, *107*, 163–215.

27. Gholizadeh, A.; Borůvka, L.; Saberioon, M.; Vašát, R. A memory-based learning approach as compared to other data mining algorithms for the prediction of soil texture using diffuse reflectance spectra. *Remote Sens.* **2016**, *8*, 341.

28. Veres, M.; Lacey, G.; Taylor, G. W. Deep learning architectures for soil property prediction. *Proc. -2015 12th Conf. Comput. Robot Vis.* **2015**, 8–15.

29. Xue, J.; Su, B. Significant remote sensing vegetation indices: A review of developments and applications. *J. Sensors* **2017**, *2017*.

30. Bartholomeus, H. M.; Schaepman, M. E.; Kooistra, L.; Stevens, A.; Hoogmoed, W. B.; Spaargaren, O. S. P. Spectral reflectance based indices for soil organic carbon quantification. *Geoderma* **2008**, *145*, 28–36.

31. Levin, N.; Kidron, G. J.; Ben-Dor, E. Surface properties of stabilizing coastal dunes: Combining spectral and field analyses. *Sedimentology* **2007**, *54*, 771–788.

32. Mukherjee, K.; Ghosh, J. K.; Mittal, R. C. Dimensionality reduction of hyperspectral data using spectral fractal feature. *Geocarto Int.* **2012**, *27*, 515–531.

33. Huang, H.; Luo, F.; Liu, J.; Yang, Y. Dimensionality reduction of hyperspectral images based on sparse discriminant manifold embedding. *ISPRS J. Photogramm. Remote Sens.* **2015**, *106*, 42–54.

34. Liu, L.; Ji, M.; Dong, Y.; Zhang, R.; Buchroithner, M. Quantitative retrieval of organic soil properties from visible near-infrared shortwave infrared (Vis-NIR-SWIR) spectroscopy using fractal-based feature extraction. *Remote Sens.* **2016**, *8*, 1035.

35. Viscarra Rossel, R. A.; Chappell, A.; De Caritat, P.; Mckenzie, N. J. On the soil information content of visible-near infrared reflectance spectra. *Eur. J. Soil Sci.* **2011**, *62*, 442–453.

36. Viscarra Rossel, R. A.; Chen, C. Digitally mapping the information content of visible-near infrared spectra of surficial Australian soils. *Remote Sens. Environ.* **2011**, *115*, 1443–1455.

37. Bengio, Y.; Courville, A.; Vincent, P. Representation learning: A review and new perspectives. *Pattern Anal. Mach. Intell. IEEE Trans.* **2013**, *35*, 1798–1828.

38. Roweis, S. Nonlinear dimensionality reduction by locally linear embedding. *Science (80-.).* **2000**, *290*, 2323–2326.

39. Ramirez-lopez, L.; Behrens, T.; Schmidt, K.; Viscarra Rossel, R. A.; Demattê, J. A. M.; Scholten, T. Distance and similarity-search metrics for use with soil vis-NIR spectra. *Geoderma* **2013**, *199*, 43–53.

40. Ball, J. E.; Anderson, D. T.; Chan, C. S. A comprehensive survey of deep learning in remote sensing: Theories, tools and challenges for the community. *J. Appl. Remote Sens.* **2017**, *11*.

41. Xing, C.; Ma, L.; Yang, X. Stacked denoise autoencoder based feature extraction and classification for hyperspectral images. *J. Sensors* **2015**, *2016*.

42. Dertat, A. Applied deep learning-part 3: Autoencoders https://towardsdatascience.com/applied-deep-learning-part-3-autoencoders-1c083af4d798 (accessed Jan 10, 2018).

43. Wang, Y.; Huang, T.; Liu, J.; Lin, Z.; Li, S.; Wang, R.; Ge, Y. Soil pH value, organic matter and macronutrients contents prediction using optical diffuse reflectance spectroscopy. *Comput. Electron. Agric.* **2015**, *111*, 69–77.

44. Shi, Z.; Wang, Q. L.; Peng, J.; Ji, W.; Liu, H. J.; Li, X.; Viscarra Rossel, R. A. Development of a national VNIR soil-spectral library for soil classification and prediction of organic matter concentrations. *Sci. China Earth Sci.* **2014**, *57*, 1671–1680.

45. Ben-Dor, E.; Banin, A. Near-Infrared analysis as a rapid method to simultaneously evaluate several Soil properties. *Soil Sci. Soc. Am. J.* **1995**, *59*, 364–372.

46. Wang, J.; Cui, L.; Gao, W.; Shi, T.; Chen, Y.; Gao, Y. Prediction of low heavy metal concentrations in agricultural soils using visible and near-infrared reflectance spectroscopy. *Geoderma* **2014**, *216*, 1–9.

47. Ben-Dor, E.; Patkin, K.; Richter, R.; Mueller, A.; Kaufmann, H. Mapping of several soil properties using DAIS-7915. In *A Decade of Trans-European Remote Sensing Cooperation*; Buchroithner, M., Ed.; CRC Press: Dresden, 2001; pp. 385–390.

48. Kopačková, V.; Ben-Dor, E.; Carmon, N.; Notesco, G. Modelling diverse soil attributes with visible to longwave infrared spectroscopy using PLSR employed by an automatic modelling engine. *Remote Sens.* **2017**, *9*, 134.

49. Leone, A.; Viscarra-Rossel, R. A.; Amenta, P.; Buondonno, A. Prediction of soil properties with PLSR and vis-NIR spectroscopy: Application to mediterranean soils from Southern Italy. *Curr. Anal. Chem.* **2012**, *8*, 283–299.

50. Gholizadeh, A.; Carmon, N.; Klement, A.; Ben-Dor, E.; Borůvka, L. Agricultural soil spectral response and properties assessment: Effects of measurement orotocol and data mining technique. *Remote Sens.* **2017**, *9*, 1078.

51. Tran, T. N.; Afanador, N. L.; Buydens, L. M. C.; Blanchet, L. Interpretation of variable importance in partial least squares with significance multivariate correlation (sMC). *Chemom. Intell. Lab. Syst.* **2014**, *138*, 153–160.

52. Li, X.; Zhang, Y.; Bao, Y.; Luo, J.; Jin, X.; Xu, X.; Song, X.; Yang, G. Exploring the best hyperspectral features for LAI estimation using partial least squares regression. *Remote Sens.* **2014**, *6*, 6221–6241.

53. Mehmood, T.; Liland, K. H.; Snipen, L.; Sæbø, S. A review of variable selection methods in partial least squares regression. *Chemom. Intell. Lab. Syst.* **2012**, *118*, 62–69.

54. Norgaard, L.; Wagner, J.; Nielsen, J. P.; Munc, L.; Engelsen, S. B. Interval partial least-squares regression (iPLS): A comparative chemometric study with an example from near-infrared spectroscopy. *Appl. Spectrosc.* **2000**, *54*, 413–419.

55. Vohland, M.; Besold, J.; Hill, J.; Fründ, H.-C. Comparing different multivariate calibration methods for the determination of soil organic carbon pools with visible to near infrared spectroscopy. *Geoderma* **2011**, *166*, 198–205.

56. Christy, C. D.; Dyer, S. A. Estimation of soil properties using a combination of spectral and scalar sensor data. In *2006 IEEE Instrumentation and Measurement Technology Conference Proceedings*; 2006; pp. 729–734.

57. Gogé, F.; Joffre, R.; Jolivet, C.; Ross, I.; Ranjard, L. Optimization criteria in sample selection step of local regression for quantitative analysis of large soil NIRS database. *Chemom. Intell. Lab. Syst.* **2012**, *110*, 168–176.

58. Bu, H. L.; Li, G. Z.; Zeng, X. Q.; Yang, J. Y.; Yang, M. Q. Feature selection and partial least squares based dimension reduction for tumor classification. In *Proceedings of the 7th IEEE International Conference on Bioinformatics and Bioengineering, BIBE*; 2007; pp. 967–973.

59. Boulesteix, A.-L. PLS dimension reduction for classification with microarray data. *Stat. Appl. Genet. Mol. Biol.* **2004**, *3*, 1–30.

60. Liu, Y.; Rayens, W. PLS and dimension reduction for classification. *Comput. Stat.* **2007**, *22*, 189–208.

61. Tang, L.; Peng, S.; Bi, Y.; Shan, P.; Hu, X. A new method combining LDA and PLS for dimension reduction. *PLoS One* **2014**, *9*, e96944.

62. Rosipal, R.; Krämer, N. Overview and recent advances in partial least squares. In *Subspace, Latent Structure and Feature Selection*; Springer Berlin Heidelberg, 2006; pp. 34–51.

63. Höskuldsson, A. PLS regression methods. *J. Chemom.* **1988**, *2*, 211–228.

64. Chen, T.; Guestrin, C. XGBoost: Reliable large-scale tree boosting system. In *22nd SIGKDD Conference on Knowledge Discovery and Data Mining*; 2016.

65. Agrawal, R. J.; Shanahan, J. G. Location disambiguation in local searches using gradient boosted decision trees. *Proc. 18th SIGSPATIAL Int. Conf. Adv. Geogr. Inf. Syst. - GIS '10* **2010**, 129–136.

66. Mitchell, R.; Frank, E. Accelerating the XGBoost algorithm using GPU computing. *PeerJ Prepr.* **2017**, *5*, e2911v1.

67. Tóth, G.; Jones, A.; Montanarella, L. *LUCAS topsoil survey: Methodology, data, and results*; 2013.

68. Tóth, G.; Jones, A.; Montanarella, L. The LUCAS topsoil database and derived information on the regional variability of cropland topsoil properties in the European Union. *Environ. Monit. Assess.* **2013**, *185*, 7409–7425.

69. Friedman, J. H. Greedy function approximation: A gradient boosting machine. *Ann. Stat.* **2001**, *29*, 1189–1232.

70. Friedman, J. H. Stochastic Gradient Boosting. *Comput. Stat. Data Anal.* **2002**, *38*, 367–378.

71. Chopra, T.; Vajpai, J. Fault diagnosis in benchmark process control system using stochastic gradient boosted decision trees. *Int. J. Soft Comput. Eng.* **2011**, *1*, 98–101.

72. Ke, G.; Meng, Q.; Finley, T.; Wang, T.; Chen, W.; Ma, W.; Ye, Q.; Liu, T. LightGBM®: A highly efficient gradient boosting decision tree. In *Advances in Neural Information Processing Systems*; 2017; pp. 3148–3156.

73. Pedregosa, F.; Weiss, R.; Brucher, M. Scikit-learn®: Machine learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.

74. LightGBM https://github.com/Microsoft/LightGBM/ (accessed Dec 10, 2017).

75. Zhu, J.; Shan, Y.; Mao, J.; Yu, D.; Rahmanian, H.; Zhang, Y. Deep Embedding Forest: Forest-based Serving with Deep Embedding Features. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*; 2017; p. 1703.05291.

76. Viscarra Rossel, R. A.; McGlynn, R. N.; McBratney, A. B. Determining the composition of mineral-organic mixes using UV-vis-NIR diffuse reflectance spectroscopy. *Geoderma* **2006**, *137*, 70–82.

77. Ben-Dor, E.; Taylor, R. G.; Hill, J.; Demattê, J. A. M.; Whiting, M. L.; Chabrillat, S.; Sommer, S. Imaging spectrometry for soil applications. *Adv. Agron.* **2008**, *97*, 321–392.

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

79. Peng, X.; Shi, T.; Song, A.; Chen, Y.; Gao, W. Estimating soil organic carbon using VIS/NIR spectroscopy with SVMR and SPA methods. *Remote Sens.* **2014**, *6*, 2699–2717.

80. Vohland, M.; Ludwig, M.; Thiele-bruhn, S.; Ludwig, B. Determination of soil properties with visible to near- and mid-infrared spectroscopy⊙: Effects of spectral variable selection. *Geoderma* **2014**, *223–225*, 88–96.

81. Zhang, L.; Zhang, L.; Kumar, V. Deep learning for remote sensing data: A technical tutorial on the state of the art. *IEEE Geosci. Remote Sens. Mag.* **2016**, *4*, 22–40.

82. Vincent, P.; Larochelle, H.; Lajoie, I.; Bengio, Y.; Pierre-AntoineManzagol Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *J. Mach. Learn. Res.* **2010**, *11*, 3371–3408.

83. Caruana, R.; Niculescu-Mizil, A. An empirical comparison of supervised learning algorithms. *Proc. 23rd Int. Conf. Mach. Learn.* **2006**, *C*, 161–168.

84. Caruana, R.; Karampatziakis, N.; Yessenalina, A. An empirical evaluation of supervised learning in high dimensions. *Proc. 25th Int. Conf. Mach. Learn.* **2008**, 96–103.

85. Stevens, A.; Nocita, M.; Tóth, G.; Montanarella, L.; van Wesemael, B. Prediction of soil organic carbon at the European scale by visible and near infraRed reflectance spectroscopy. *PLoS One* **2013**, *8*, e66409.

86. Nocita, M.; Stevens, A.; Toth, G.; Panagos, P.; van Wesemael, B.; Montanarella, L. Prediction of soil organic carbon content by diffuse reflectance spectroscopy using a local partial least square regression approach. *Soil Biol. Biochem.* **2014**, *68*, 337–347.

87. Chabrillat, S.; Ben-Dor, E.; Viscarra Rossel, R. A.; Demattê, J. A. M. Quantitative soil spectroscopy. *Appl. Environ. Soil Sci.* **2013**.

88. Rossel, R. A. V.; Behrens, T. Using data mining to model and interpret soil diffuse reflectance spectra. *Geoderma* **2010**, *158*, 46–54.

89. Epema, G. F.; Kooistra, L.; Wanders, J. Spectroscopy for the assessment of soil properties in reconstructed river floodplains. In *3rd EARSeL Workshop on Imaging Spectroscopy*; 2003.

90. Udelhoven, T.; Emmerling, C.; Jarmer, T. Quantitative analysis of soil chemical properties with diffuse refectance spectrometry and partial least-square regression: A feasibility study. *Plant Soil* **2003**, *251*, 319–329.

91. McBratney, A. B.; Minasny, B.; Viscarra Rossel, R. A. Spectral soil analysis and inference systems: A powerful combination for solving the soil data crisis. *Geoderma* **2006**, *136*, 272–278.

92. Shepherd, K. D.; Walsh, M. G. Infrared spectroscopy-enabling an evidence-based diagnostic surveillance approach to agricultural and environmental management in developing countries. *J. Near Infrared Spectrosc.* **2007**, *15*, 1–19.

93. Tóth, G.; Hermann, T.; Da Silva, M. R.; Montanarella, L. Heavy metals in agricultural soils of the European Union with implications for food safety. *Environ. Int.* **2016**, *88*, 299–309.

94. Vågen, T.-G.; Shepherd, K. D.; Walsh, M. G.; Winowiecki, L.; Desta, L. T.; Tondoh, J. E. *AfSIS technical specifications: Soil health surveillance*; 2010.

95. Qiao, T.; Ren, J.; Craigie, C.; Zabalza, J.; Maltin, C.; Marshall, S. Quantitative prediction of beef quality using Visible and NIR spectroscopy with large data samples under industry conditions. *J. Appl. Spectrosc.* **2015**, *82*, 137–144.

96. Li, F.; Xu, L.; Wong, A.; Clausi, D. A. Feature extraction for hyperspectral imagery via ensemble localized manifold learning. *IEEE Geosci. Remote Sens. Lett.* **2015**, *12*, 2486–2490.

97. Bakir, C. Nonlinear feature extraction for hyperspectral images. *Int. J. Appl. Math. Electron. Comput.* **2015**, *3*, 244–248.

98. Lunga, D.; Prasad, S.; Crawford, M. M.; Ersoy, O. Manifold-learning-based feature extraction for classification of hyperspectral data: A review of advances in manifold learning. *Signal Process. Mag. IEEE* **2014**, *31*, 55–66.

99. Zheng, L.; Li, M.; An, X.; Pan, L.; Sun, H. Spectral feature extraction and modeling of soil total nitrogen content based on NIR technology and wavelet packet analysis. In *SPIE Asia-Pacific Remote Sensing*; 2010; Vol. 7857, p. 78571M1-8.

100. Kalousis, A.; Prados, J. Feature extraction from mass spectra for classification. In *6th European Conference on Principles and Practice of Knowledge Discovery in Databases*; 2005.

101. Ghosh, J. K.; Somvanshi, A. Fractal-based dimensionality reduction of hyperspectral images. *J. Indian Soc. Remote Sens.* **2008**, *36*, 235–241.

102. Mukherjee, K.; Bhattacharya, A.; Ghosh, J. K.; Arora, M. K. Comparative performance of fractal based and conventional methods for dimensionality reduction of hyperspectral data. *Opt. Lasers Eng.* **2014**, *55*, 267–274.

103. Ballabio, C.; Panagos, P.; Monatanarella, L. Mapping topsoil physical properties at European scale using the LUCAS database. *Geoderma* **2016**, *261*, 110–123.

104. Reljin, I. S.; Reljin, B. D.; Avramov-Ivić, M. L.; Jovanović, D. V.; Plavec, G. I.; Petrović, S. D.; Bogdanović, G. M. Multifractal analysis of the UV/VIS spectra of malignant ascites: Confirmation of the diagnostic validity of a clinically evaluated spectral analysis. *Phys. A Stat. Mech. its Appl.* **2008**, *387*, 3563–3573.

105. Hall, P.; Wood, A. On the performance of box-counting estimators of fractal dimension. *Biometrika* **1993**, *80*, 246–251.

106. Constantine, A. G.; Hall., P. Characterizing surface smoothness via estimation of effective fractal dimension. *J. R. Stat. Soc. Ser. B* **1994**, *56*, 97–113.

107. Chan, G.; Hall, P.; Poskitt, D. Periodogram-based estimators of fractal properties. *Ann. Stat.* **1995**, 1684–1711.

108. Klinkenberg, B.; Although, I.; Columbia, B. A review of methods used to determine the fractal dimension of linear features. *Math. Geol.* **1994**, *26*, 23–46.

109. Gneiting, T.; Sevcikova, H.; Percival, D. B. Estimators of fractal dimension: assessing the roughness of time series and spatial data. *Stat. Sci.* **2011**, *27*, 247–277.

110. Mukherjee, K.; Ghosh, J. K.; Mittal, R. C. Variogram fractal dimension based features for hyperspectral data dimensionality reduction. *J. Indian Soc. Remote Sens.* **2013**, *41*, 249–258.

111. Song, R.; Chen, S.; Deng, B.; B, L. L. Extreme gradient boosting for dentifying individual users across different digital devices. In *International Conference on Web-Age Information Management*; 2016; pp. 43–54.

112. Babajide Mustapha, I.; Saeed, F. Bioactive molecule prediction using extreme gradient boosting. *Molecules* **2016**, *21*, 983.

113. Kopačková, V. Using multiple spectral feature analysis for quantitative pH mapping in a mining environment. *Int. J. Appl. Earth Obs. Geoinf.* **2014**, *28*, 28–42.

114. Wijaya, A.; Marpu, P. R.; Gloaguen, R. Geostatistical texture classification of tropical rainforest in Indonesia. In *ISPRS International Symposium on Spatial Data Quality*; 2007.

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

116. He, T.; Wang, J.; Lin, Z.; Cheng, Y. Spectral features of soil organic matter. *Geo-Spatial Inf. Sci.* **2009**, *12*, 33–40.

117. Zhang, T.; Li, L.; Zheng, B. Estimation of agricultural soil properties with imaging and laboratory spectroscopy. *J. Appl. Remote Sens.* **2013**, *7*, 073587.

118. Ciampalini, A.; André, F.; Garfagnoli, F.; Grandjean, G.; Lambot, S.; Chiarantini, L.; Moretti, S. Improved estimation of soil clay content by the fusion of remote hyperspectral and proximal geophysical sensing. *J. Appl. Geophys.* **2015**, *116*, 135–145.

119. Adeline, K. R. M.; Gomez, C.; Gorretta, N.; Roger, J. M. Predictive ability of soil properties to spectral degradation from laboratory Vis-NIR spectroscopy data. *Geoderma* **2017**, *288*, 143–153.

120. Gomez, C.; Drost, A. P. A.; Roger, J. M. Analysis of the uncertainties affecting predictions of clay contents from VNIR/SWIR hyperspectral data. *Remote Sens. Environ.* **2015**, *156*, 58–70.

121. Clark, R. N. Spectroscopy of rocks and minerals, and principles of spectroscopy. In *Manual of remote sensing*; 1999; Vol. 3, pp. 3–58.

122. Erlei, Z.; Xiangrong, Z.; Shuyuan, Y.; Shuang, W. Improving hyperspectral image classification using spectral information divergence. *Geosci. Remote Sens. Lett. IEEE* **2014**, *11*, 249–253.

123. Gomez, C.; Viscarra Rossel, R. A.; McBratney, A. B. Soil organic carbon prediction by hyperspectral remote sensing and field vis-NIR spectroscopy: An Australian case study. *Geoderma* **2008**, *146*, 403–411.

124. Lagacherie, P.; Baret, F.; Feret, J. B.; Madeira Netto, J.; Robbez-Masson, J. M. Estimation of soil clay and calcium carbonate using laboratory, field and airborne hyperspectral measurements. *Remote Sens. Environ.* **2008**, *112*, 825–835.

125. Gomez, C.; Lagacherie, P.; Coulouma, G. Continuum removal versus PLSR method for clay and calcium carbonate content estimation from laboratory and airborne hyperspectral measurements. *Geoderma* **2008**, *148*, 141–148.

126. Bartholomeus, H.; Epema, G.; Schaepman, M. Determining iron content in Mediterranean soils in partly vegetated areas, using spectral reflectance and imaging spectroscopy. *Int. J. Appl. Earth Obs. Geoinf.* **2007**, *9*, 194–203.

127. Jiang, Q.; Chen, Y.; Guo, L.; Fei, T.; Qi, K. Estimating soil organic carbon of cropland soil at different levels of soil moisture using vis-NIR spectroscopy. *Remote Sens.* **2016**, *8*, 755.

128. Pearlman, J.; Carman, S.; Segal, C.; Jarecke, P.; Clancy, P.; Browne, W. Overview of the Hyperion imaging spectrometer for the NASA EO-1 mission. In *IEEE International Symposium on Geoscience and Remote Sensing Symposium*; 2001; pp. 3036–3038.

129. Manzo, C.; Valentini, E.; Taramelli, A.; Filipponi, F.; Disperati, L. Spectral characterization of coastal sediments using field spectral libraries, airborne hyperspectral images and topographic LiDAR data (FHyL). *Int. J. Appl. Earth Obs. Geoinf.* **2015**, *36*, 54–68.

130. Lobell, D. B.; Asner, G. P. Moisture effects on soil reflectance. *Soil Sci. Soc. Am. J.* **2002**, *66*, 722.

131. Liu, L.; Ji, M.; Buchroithner, M. Combining partial least squares and the gradient-boosting method for soil property retrieval using visible near-infrared shortwave infrared spectra. *Remote Sens.* **2017**, *9*, 1299.

132. Kopačková, V.; Ben-Dor, E. Normalizing reflectance from different spectrometers and protocols with an internal soil standard. *Int. J. Remote Sens.* **2016**, *37*, 1276–1290.

133. Notesco, G.; Ogen, Y.; Ben-Dor, E. Mineral classification of makhtesh ramon in israel using hyperspectral longwave infrared (LWIR) remote-sensing data. *Remote Sens.* **2015**, *7*, 12282–12296.

134. Castaldi, F.; Chabrillat, S.; Chartin, C.; Genot, V.; Jones, A. R.; van Wesemael, B. Estimation of soil organic carbon in arable soil in Belgium and Luxembourg with the LUCAS topsoil database. *Eur. J. Soil Sci.* **2018**, *69*, 592–603.

135. Schwanghart, W.; Jarmer, T. Linking spatial patterns of soil organic carbon to topography - A case study from south-eastern Spain. *Geomorphology* **2011**, *126*, 252–263.

136. Werban, U.; Bartholomeus, H. M. .; Dietrich, P.; Grandjean, G.; Zacharias, S. Digital soil mapping: Approaches to integrate sensing techniques to the prediction of key soil properties. *Vadose Zo. J.* **2013**, *12*, 1–4.

137. Li, D.; Chen, X.; Peng, Z.; Chen, S.; Chen, W.; Han, L.; Li, Y. Prediction of soil organic matter content in a litchi orchard of South China using spectral indices. *Soil Tillage Res.* **2012**, *123*, 78–86.

138. Peón, J.; Recondo, C.; Fernández, S.; F. Calleja, J.; De Miguel, E.; Carretero, L. Prediction of topsoil organic carbon using airborne and satellite hyperspectral imagery. *Remote Sens.* **2017**, *9*, 1211.

139. Rivero, R. G.; Grunwald, S.; Binford, M. W.; Osborne, T. Z. Integrating spectral indices into prediction models of soil phosphorus in a subtropical wetland. *Remote Sens. Environ.* **2009**, *113*, 2389–2402.

140. Yuan, Y.; Member, S.; Zheng, X.; Lu, X.; Member, S. Hyperspectral image superresolution by transfer learning. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2017**, *10*, 1963–1974.

141. Sameen, M. I.; Pradhan, B. A novel road segmentation technique from orthophotos using deep convolutional autoencoders. *Korean J. Remote Sens.* **2017**, *33*, 423–436.

142. Sameen, M. I.; Pradhan, B.; Aziz, O. S. Classification of very high resolution aerial photos using spectral-spatial convolutional neural networks. *J. Sensors* **2018**, *2018*.

143. Lyu, H.; Lu, H.; Mou, L. Learning a transferable change rule from a recurrent neural network for land cover change detection. *Remote Sens.* **2016**, *8*, 1–22.

144. Ye, F.; Su, Y.; Xiao, H.; Zhao, X.; Min, W. Remote sensing image registration using convolutional neural network features. *IEEE Geosci. Remote Sens. Lett.* **2018**, *15*, 232–236.

145. Nahhas, F. H.; Shafri, H. Z. M.; Sameen, M. I.; Pradhan, B.; Mansor, S. Deep learning approach for building detection using liDAR-orthophoto fusion. *J. Sensors* **2018**, *2018*, 7212307.

146. Zhu, X. X.; Tuia, D.; Mou, L.; Xia, G. S.; Zhang, L.; Xu, F.; Fraundorfer, F. Deep learning in remote sensing: A comprehensive review and list of resources. *IEEE Geosci. Remote Sens. Mag.* **2017**, *5*, 8–36.

147. Hu, F.; Xia, G.-S.; Hu, J.; Zhang, L. Transferring deep convolutional neural networks for the scene classification of high-resolution remote sensing imagery. *Remote Sens.* **2015**, *7*, 14680–14707.

148. Zhao, B.; Huang, B.; Zhong, Y. Transfer learning with fully pretrained deep convolution networks for land-use classification. *IEEE Geosci. Remote Sens. Lett.* **2017**, *14*, 1436–1440.

149. Huang, Z.; Pan, Z.; Lei, B. Transfer learning with deep convolutional neural network for SAR target classification with limited labeled data. *Remote Sens.* **2017**, *9*, 1–21.

150. Chabrillat, S.; Naumann, N.; Escribano, P.; Bachmann, M.; Spengler, D.; Holzwarth, S.; Palacios-Orueta, A. Oyonarte, C. Cabo de Gata-Nijar natural park 2003-2005 - A multitemporal hyperspectral flight campaign for EnMAP science preparatory activities. EnMAP flight campaigns technical report. *GFZ Data Serv.* **2016**.

151. Chabrillat, S.; Guillaso, S.; Rabe, A.; Foerster, S.; Guanter, L. From HYSOMA to ENSOMAP - A new open source tool for quantitative soil properties mapping based on hyperspectral imagery from airborne to spaceborne applications. In *EGU General Assembly Conference Abstracts*; 2016; Vol. 18, p. 14697.

152. Lazebnik, S. Deep convolutional neural networks for hyperspectral image classification. *J. Sensors* **2015**, 258619.

153. Petersson, H.; Gustafsson, D.; Bergstr, D. Hyperspectral image analysis using deep learning - A review. In *2016 6th International Conference on Image Processing Theory Tools and Applications (IPTA)*; 2016; pp. 1–6.

154. Wu, H.; Prasad, S. Convolutional recurrent neural networks for hyperspectral data classification. *Remote Sens.* **2017**, *9*, 298.

155. Mehdipour Ghazi, M.; Yanikoglu, B.; Aptoula, E. Plant identification using deep neural networks via optimization of transfer learning parameters. *Neurocomputing* **2017**, *235*, 228–235.

156. Nocita, M.; Kooistra, L.; Bachmann, M.; Müller, A.; Powell, M.; Weel, S. Predictions of soil surface and topsoil organic carbon content through the use of laboratory and field spectroscopy in the Albany Thicket Biome of Eastern Cape Province of South Africa. *Geoderma* **2011**, *167–168*, 295–302.

157. Vindušková, O.; Dvořáček, V.; Prohasková, A.; Frouz, J. Distinguishing recent and fossil organic matter - A critical step in evaluation of post-mining soil development - using near infrared spectroscopy. *Ecol. Eng.* **2014**, *73*, 643–648.

158. Vohland, M.; Ludwig, M.; Thiele-Bruhn, S.; Ludwig, B.; Mechanisms, P.; Vohland, M.; Ludwig, M.; Thiele-Bruhn, S.; Ludwig, B. Quantification of soil properties with hyperspectral data®: Selecting spectral variables with different methods to improve accuracies and analyze prediction mechanisms. *Remote Sens.* **2017**, *9*, 1–24.

159. Kingma, D. P.; Ba, J. Adam: A method for stochastic optimization. *arXiv Prepr. arXiv1412.6980* **2014**, 1–15.

160. Calleja, J. F.; Hellmann, C.; Mendiguren, G.; Punalekar, S.; Peón, J.; MacArthur, A.; Alonso, L. Relating hyperspectral airborne data to ground measurements in a complex and discontinuous canopy. *Acta Geophys.* **2015**, *63*, 1499–1515.

161. Castro-Esau, K. L.; Sánchez-Azofeifa, G. A.; Rivard, B. Comparison of spectral indices obtained using multiple spectroradiometers. *Remote Sens. Environ.* **2006**, *103*, 276–288.

162. Castaldi, F.; Palombo, A.; Pascucci, S.; Pignatti, S.; Santini, F.; Casa, R. Reducing the influence of soil moisture on the estimation of clay from hyperspectral data: A case study using simulated PRISMA data. *Remote Sens.* **2015**, *7*, 15561–15582.

163. Wu, C.; Jacobson, A.; Laba, M.; Baveye, P. C. Alleviating moisture content effects on the visible near-infrared diffuse-reflectance sensing of soils. *Soil Sci.* **2009**, *174*, 456–465.

164. Wang, Y.; Veltkamp, D. J.; Kowalski, B. R. Multivariate instrument standardization. *Anal. Chem.* **1991**, *63*, 2750–2756.

165. Ji, W.; Viscarra Rossel, R. A.; Shi, Z. Accounting for the effects of water and the environment on proximally sensed vis-NIR soil spectra and their calibrations. *Eur. J. Soil Sci.* **2015**, *66*, 555–565.

166. Liu, L.; Ji, M.; Buchroithner, M. A case study of the forced invariance approach for Soil salinity estimation in vegetation-covered terrain using airborne hyperspectral imagery. *ISPRS Int. J. Geo-Information* **2018**, *7*, 48.

167. Douaoui, A. E. K.; Nicolas, H.; Walter, C. Detecting salinity hazards within a semiarid context by means of combining soil and remote-sensing data. *Geoderma* **2006**, *134*, 217–230.

168. Alexakis, D.; Gotsis, D.; Giakoumakis, S. Evaluation of soil salinization in a Mediterranean site (Agoulinitsa district-West Greece). *Arab. J. Geosci.* **2014**, *8*, 1373–1383.

169. Fourati, H. T.; Bouaziz, M.; Benzina, M.; Bouaziz, S. Modeling of soil salinity within a semi-arid region using spectral analysis. *Arab. J. Geosci.* **2015**, *8*, 11175–11182.

170. Biro, K.; Pradhan, B.; Buchroithner, M.; Makeschin, F. Land use/land cover change analysis and its impact on soil properties in the northern part of Gadarif region, Sudan. *L. Degrad. Dev.* **2013**, *24*, 90–102.

171. Biro, K.; Pradhan, B.; Sulieman, H.; Buchroithner, M. Exploitation of TerraSAR-X data for land use/land cover analysis using object-oriented classification approach in the African Sahel Area, Sudan. *J. Indian Soc. Remote Sens.* **2013**, *41*, 539–553.

172. Asfaw, E.; Suryabhagavan, K. V.; Argaw, M. Soil salinity modeling and mapping using remote sensing and GIS: The case of Wonji sugar cane irrigation farm, Ethiopia. *J. Saudi Soc. Agric. Sci.* **2016**.

173. Metternicht, G. I.; Zinck, J. a. Remote sensing of soil salinity: Potentials and constraints. *Remote Sens. Environ.* **2003**, *85*, 1–20.

174. Allbed, A.; Kumar, L. Soil salinity mapping and monitoring in arid and Semi-arid regions using remote sensing technology: A review. *Adv. Remote Sens.* **2013**, *2*, 373–385.

175. Zhang, T.; Qi, J.; Gao, Y.; Ouyang, Z.; Zeng, S.; Zhao, B. Detecting soil salinity with MODIS time series VI data. *Ecol. Indic.* **2015**, *52*, 480–489.

176. Tilley, D. R.; Ahmed, M.; Son, J. H.; Badrinarayanan, H. Hyperspectral reflectance response of freshwater macrophytes to salinity in a brackish subtropical marsh. *J. Environ. Qual.* **2007**, *36*, 780.

177. Huete, A. . A soil-adjusted vegetation index (SAVI). *Remote Sens. Environ.* **1988**, *25*, 295–309.

178. Bannari, A.; Guedon, A. M.; El-Harti, A.; Cherkaoui, F. Z.; El-Ghmari, A. Characterization of slightly and moderately saline and sodic soils in irrigated agricultural land using simulated data of advanced land imaging (EO-1) sensor. *Commun. Soil Sci. Plant Anal.* **2008**, *39*, 2795–2811.

179. Al-Khaier, F. Soil salinity detection using satellite remote sensing, International Institute for Geo-information Science and Earth Observation, 2003.

180. Khan, N. M.; Rastoskuev, V. V.; Sato, Y.; Shiozawa, S. Assessment of hydrosaline land degradation by using a simple approach of remote sensing indicators. *Agric. Water Manag.* **2005**, *77*, 96–109.

181. Weng, Y.-L.; Gong, P.; Zhu, Z.-L. A spectral index for estimating soil salinity in the Yellow River Delta region of China using EO-1 Hyperion data. *Pedosphere* **2010**, *20*, 378–388.

182. Hochberg, E. J.; Roberts, D. a.; Dennison, P. E.; Hulley, G. C. Special issue on the Hyperspectral Infrared Imager (HyspIRI): Emerging science in terrestrial and aquatic ecology, radiation balance and hazards. *Remote Sens. Environ.* **2015**, *167*, 1–5.

183. Palacios-Orueta, A.; Pinzon, J. E.; Ustin, S. L.; Roberts, D. A. Remote sensing of soils in the Santa Monica Mountains: II. Hierarchical foreground and background analysis. *Remote Sens. Environ.* **1999**, *68*, 138–151.

184. Mashimbye, Z. E.; Cho, M. a.; Nell, J. P.; De Clercq, W. P.; Van Niekerk, A.; Turner, D. P. Model-Based integrated methods for quantitative estimation of soil salinity from hyperspectral remote sensing data: A case study of selected south African soils. *Pedosphere* **2012**, *22*, 640–649.

185. Dehaan, R.; Taylor, G. R. Image-derived spectral endmembers as indicators of salinisation. *Int. J. Remote Sens.* **2003**, *24*, 775–794.

186. Malec, S.; Rogge, D.; Heiden, U.; Sanchez-Azofeifa, A.; Bachmann, M.; Wegmann, M. Capability of spaceborne hyperspectral EnMAP mission for mapping fractional cover for soil erosion modeling. *Remote Sens.* **2015**, *7*, 11776–11800.

187. Asner, G. P.; Heidebrecht, K. B. Spectral unmixing of vegetation, soil and dry carbon in arid regions: Comparing multispectral and hyperspectral observations. *Int. J. Remote Sens.* **2001**, *23*, 3939–3958.

188. Muller, E.; Décamps, H. Modeling soil moisture–reflectance. *Remote Sens. Environ.* **2000**, *76*, 173–180.

189. Crippen, R. E.; Blom, R. G. Unveiling the lithology of vegetated terrains in remotely sensed imagery. *Photogramm. Eng. Remote Sensing* **2001**, *91109*, 935–943.

190. Yu, L.; Porwal, A.; Holden, E.-J.; Dentith, M. C. Suppression of vegetation in multispectral remote sensing images. *Int. J. Remote Sens.* **2011**, *32*, 7343–7357.

191. Evans, D.; Traviglia, A. Uncovering Angkor: Integrated remote sensing applications in the archaeology of early Cambodia. In *Satellite Remote Sensing: A New Tool for Archaeology*; Lasaponara, R.; Masini, N., Eds.; Springer Berlin Heidelberg, 2003; pp. 197–230.

192. Kang, J.; Li, X.; Jin, R.; Ge, Y.; Wang, J.; Wang, J. Hybrid optimal design of the eco-hydrological wireless sensor network in the middle reach of the Heihe River Basin, China. *Sensors* **2014**, *14*, 19095–19114.

193. Jin, R.; Li, X.; Yan, B.; Li, X. A nested ecohydrological wireless sensor network for capturing the surface heterogeneity in the midstream areas of the Heihe River Basin,China. *IEEE Geosci. Remote Sens. Lett.* **2015**, *11*, 2015–2019.

194. Meng, J.; Long, H. Water resources in oasis ecological balance: The case of Zhangye Oasis. *Chinese J. Arid L. Res.* **1998**, *11*, 255–262.

195. Li, X.; Cheng, G.; Liu, S.; Xiao, Q.; Ma, M.; Jin, R.; Che, T.; Liu, Q.; Wang, W.; Qi, Y.; Wen, J.; Li, H.; Zhu, G.; Guo, J.; Ran, Y.; Wang, S.; Zhu, Z.; Zhou, J.; Hu, X.; Xu, Z. Heihe watershed allied telemetry experimental research (HiWater) scientific objectives and experimental design. *Bull. Am. Meteorol. Soc.* **2013**, *94*, 1145–1160.

196. Xiaodong, S.; Ganlin, Z.; Feng, L. I. U.; Decheng, L. I.; Yuguo, Z.; Jinling, Y. Modeling spatio-temporal distribution of soil moisture by deep learning-based cellular automata model. *J. Arid Land* **2016**, *8*, 734–748.

197. Guanter, L.; Estellés, V.; Moreno, J. Spectral calibration and atmospheric correction of ultra-fine spectral and spatial resolution remote sensing data. Application to CASI-1500 data. *Remote Sens. Environ.* **2007**, *109*, 54–65.

198. Li, X.; Liu, S.; Xiao, Q.; Ma, M.; Jin, R.; Che, T.; Wang, W.; Hu, X.; Xu, Z.; Wen, J.; Wang, L. A multiscale dataset for understanding complex eco-hydrological processes in a heterogeneous oasis system. *Sci. Data* **2017**, *4*, 1–11.

199. Hede, A. N. H.; Koike, K.; Kashiwaya, K.; Sakurai, S.; Yamada, R.; Singer, D. A. How can satellite imagery be used for mineral exploration in thick vegetation areas? *Geochemistry, Geophys. Geosystems* **2017**, *18*, 584–596.

200. Engstrom, R.; Hope, A.; Kwon, H.; Stow, D. The relationship between soil moisture and NDVI near Barrow, Alaska. *Phys. Geogr.* **2008**, *29*, 38–53.

201. Huang, Y.; Wang, Y.; Zhao, Y.; Xu, X.; Zhang, J.; Li, C. Spatiotemporal distribution of soil moisture and salinity in the Taklimakan Desert highway shelterbelt. *Water* **2015**, *7*, 4343–4361.

Acknowledgements

The completion of this dissertation would not have been possible without the valuable support from various people to whom I would like to express my sincere gratitude.

First of all, I am grateful to my supervisor, Prof. Manfred F. Buchroithner, for his invaluable guidance during the research period. Discussions and comments as well as advices have been very helpful for me. His constant supports and encouragements helped me make continuous progress during the past few years. I have benefited a lot from his broad knowledge and scientific experience, which will also make a positive influence on my future research career.

I want to express my thanks to Prof. Dirk Burghardt and colleagues at the Institute of Cartography, especially to Mrs. Sharma for her great support during my PhD studies.

I am grateful to Min Ji for her assistance and endless support.

I would like to thank my very best friends, Pei Wang, Ye Liu, Tao Zhang, Bin Cai, Xiao Wang etc., who brightened my life outside the university in Dresden.

I acknowledge the Chinese Scholarship Council (CSC) for the financial support of my study in TU Dresden.

Finally, I would like to express my deepest and sincerest gratitude to my family for their incredible love. Their supports give me the greatest courage to proceed through each stage of my life.