Vis-NIR and MIR Spectroscopy for Prediction and Mapping of Soil Organic Carbon and Clay

YI PENG
PhD THESIS. SCIENCE AND TECHNOLOGY. Aug 2014

Faculty of Science and Technology
Dept. of Agroecology
Aarhus University
Blichers Allé 20,
P. O. Box 50
8830 Tjele
Denmark
Main Supervisor
Senior Scientist Mogens Humlekrog Greve
Department of Agroecology
Aarhus University

Co-Supervisors
Associate Professor René Gislum
Senior Scientist Kirsten Schelde
Postdoc Maria Knadel
Department of Agroecology
Aarhus University

Assessment Committee
Senior Scientist Mogens Nicolaisen (chair)
Department of Agroecology, Aarhus University, Denmark
Professor Eyal Ben-Dor
Department of Geography, Tel-Aviv University, Israel
Professor Zhang Ganlin
Institute of Soil Science, Chinese Academy of Sciences, China
Preface

This dissertation is submitted in partial fulfilment of the requirements for the Doctor of Philosophy (PhD) degree at the Faculty of Science and Technology, Aarhus University, Denmark. The research was conducted in the period from August 2011 until August 2014 at the Department of Agroecology, Aarhus University. The study was partly funded by the China Scholarship Council (CSC), partly by the HOBE project Phase 2 “Hydrological Observatory and Exploratorium”, funded by the Villum Foundation and partly by Aarhus University. This work has been supervised by Mogens Humlekrog Greve, René Gisulum, Kirsten Schelde and Maria Knadel from the Department of Agroecology, Aarhus University. This dissertation is based on the following three journal papers, which are enclosed. The titles of the papers are:

1. Predicting soil organic carbon at field scale using a national soil spectral library
2. Quantification of SOC and Clay content Using Vis-NIR-MIR Spectroscopy with Jack-knifing Partial Least Squares Regression
3. Modeling SOC at regional scale by combing multi-spectral images with laboratory spectra

I would like to acknowledge all my supervisors for their constant support during my study over last three years. I am grateful to my external supervisors Professor Sabine Grunwald and Dr Xiongxiong from the University of Florida for hosting me during my stay abroad and for their help and guidance with the remote sensing research. I would like to thank all technicians from our department, especially Stig T. Rasmussen, Henrik Nørgaard, René Larsen, Dennis Christensen, Karin Dyrberg, and Mette Balslev Greve for their technical assistance.

I appreciate daily interaction with my fellow students, my former office mate Kabindra Adhikari and other colleagues, all the members of our journal club for sharing their knowledge to make a helpful research environment, especially Emmanuel Arthur, Marcos Paradelo Pérez have helped me in various research issues. A special thanks goes to my friend Ibrahim Karaman for his help with chemometrics, Smitha A. Ankinakatte for her help with coding to make my life easier. Many thanks go to all Norresø Kollegiet residents for a wonderful life as a family.

Last, but not least, I am thankful to my family for their constant support and encouragement which helped me to go through the whole journey.

Yi Peng/ 彭懿

August 2014, Foulum
Summary

Soils are crucial to life on earth in their support of plant growth. The sustained provision of this function depends on proper use and management of soil resources. Soil organic carbon (SOC) and clay contents are two major soil quality parameters that directly link to crop production. To achieve sustainable crop production, it is essential to understand, determine and visualise the variation of SOC and clay content within space and time. However, traditional field sampling and wet chemistry soil analysis are expensive and time-consuming. Thus, finding a more efficient and low-cost method for SOC and clay content determination is becoming a big challenge for agriculture, land use management and other purposes.

As part of the HOBE project phase 2, this thesis explored optimal methods for field-scale predictions of SOC and clay content using Visible Near-Infrared (Vis-NIR) and Mid-Infrared (MIR) diffuse reflectance spectroscopy in order to improve digital soil mapping at regional scale. The objectives of the work presented in this dissertation were to use a national spectral library to predict SOC at field scale; to use different chemometric tools to select important spectral features for SOC and clay calibration/prediction; to use laboratory Vis-NIR spectra with remote sensing (RS) spectra for SOC mapping at regional scale; to evaluate Vis-NIR and MIR techniques to replace standard wet chemistry soil analysis, and to determine the acceptable prediction accuracy of SOC and clay contents for needed for liming and nitrogen fertilization.

To achieve these objectives we used, soil samples from Denmark at national, regional and field scale, across a range of SOC and clay contents, and with different management histories. Laboratory Vis-NIR (LabSpec5100, NIRSTM 2500) was applied to all the soils, MIR (Fourier transform IR MB3000) was applied to soils from one field, and two different satellite images (Landsat8, SPOT5) were applied to soils from a catchment area. Additionally, different chemometric methods such as principle component analysis (PCA), partial least squares regression (PLSR), jack-knifing (JK) and Cubist regression rules were used for multivariate data analysis.

The first study investigated the possibilities of using a national soil spectral library for SOC prediction at field scale. To this end, 35 samples were taken from the target site for validation, and three different re-sampling strategies were applied to create a library subset and to develop different calibrations for each strategy. The three strategies were: geographically closest, same landscape and parent material, and the most alike spectra. Meanwhile, the full library was used for calibration and the spiked library with local samples was used for recalibration. The validation results from the different calibration sets showed that of all the models, the geographically closest strategy was the best approach for creating a subset for field SOC prediction. The results also confirmed that the spiking strategy outperformed the
model without spiking. This study suggested that the soil parent material, landscape and geographical location highly affect the accuracy of the model developed using a subset from a spectral library.

The second study was done on a heterogenous field and dealt with the prediction of SOC and clay contents using Vis-NIR and MIR, and using JK for variable selection. Fifty samples from the target site were used in this study and cross validation was used for model evaluation. The uncertainty estimation method called root mean square deviation (RMSD) was used. Firstly, this study found that repeated MIR measurements can effectively improve prediction accuracy by 20%. The results suggested that three and four replications can be considered the optimal number of repeated MIR measurements for SOC and clay content using Fourier Transform IR MB3000 spectrometer. Secondly, a comparison of all the prediction results found that the combination of MIR and Vis-NIR with the PLSR-JK technique yielded the lowest prediction errors. Thus, we concluded that the PLSR-JK technique simplified and enhanced the interpretation of the developed models due to a reduction in the number of variables used in the models. Finally, the RMSD value can be used as an indicator of the uncertainty in the lab analyses of SOC and clay content.

The final study upscaled point laboratory Vis-NIR spectra to image scale using the Bayesian empirical kriging method, and integrated them with remote sensing (RS) images to map top soil SOC in the Skjern river catchment, Denmark. A model stratification strategy was also applied to run separate models for upland and wetland samples. The Cubist regression rules were used for SOC modelling and spectral feature selection. Finally, 28 variables were selected for spatial modelling; these variables included ancillary environmental data, vegetation indices such as Normalized Difference Vegetation Index (NDVI) and Enhanced Vegetation Index (EVI) derived from different RS images and a laboratory spectral feature indicative of SOC (1930nm). The results showed that the prediction accuracy was significantly improved after adding the laboratory Vis-NIR spectral feature. Furthermore, calibrating the samples separately for upland and wetland improved the prediction accuracy of the upland model, but not the wetland model.

At the end of this thesis, there is a general discussion on the accuracy of using Vis-NIR and MIR to predict SOC and clay contents. In terms of liming and nitrogen fertilizer application in agriculture, the possibilities of using spectral measurement to replace laboratory wet chemistry soil analysis were evaluated. This work shows the promising potential of using Vis-NIR, MIR spectroscopy and RS images to determine SOC and clay contents and to facilitate digital soil mapping (DSM) for precision agriculture and land use management. The Danish national soil spectral library can be used as very important information to determine SOC
and clay contents at both field and regional scale. Proper chemometric tools must be chosen for multivariate data analysis to simplify the model and obtain accurate prediction results.
Sammendrag (Danish summary)

Jordbunden er afgørende for livet på jorden som grundlaget for plantevækst. For at fastholde jordens frugtbarhed kræver det korrekt forvaltning af jordressourcerne. Jordens indhold af organisk kulstof (SOC) og indhold af ler er de to vigtigste jordkvalitetsparametre i Danmark, og jordens frugtbarhed er i høj grad en funktion af disse. For at opnå bæredygtig planteproduktion er det vigtigt at kende variation af SOC og ler i tid og rum. Traditionel prøvetagning i felt og vådkemiske jordbundsanalyser er dyre og tidskrævende. At finde en mere effektiv og billig metode til bestemmelse SOC og lerindhold er en stor udfordring for landbruget.

Som en del af HOBE projektets fase 2, undersøges i denne afhandling muligheden for at kortlægge SOC og lerindhold ved hjælp af Visible (Vis-NIR) og Mid-Infrarød (MIR) diffus reflektans spektroskopi med henblik på at forbedre den digitale jordkortlægning på regionalt plan.

Målsætningerne for arbejdet var at anvende et nationalt spektralt bibliotek til bestemmelse af SOC på markskala; at bruge forskellige kemometriske værktøjer til udvælgelse af vigtige spektrale egenskaber til SOC- og lerkalibrering; at bruge Vis-NIR spektre målt i laboratoriet sammen med remote sensing (RS) spektre målt på marken til SOC kortlægning på regionalt plan; at evaluere Vis-NIR og MIR teknikker til at erstatte standard vådkemiske jordbundsanalyser; og at bestemme den acceptable prediktionsnøjagtighed af SOC og ler ved beregning af kalkmængder til kalkudbring samt bestemmelse af JB nummer som grundlag for kvælstoftildeling.

For at nå disse mål brugte vi jordprøver fra Danmark på national, regional og markskala. Spektrometret LabSpec5100, NIRSTM 2500 blev anvendt på alle jorder, MIR (Fourier transformation IR MB3000) blev anvendt på jord fra én mark, og to forskellige satellitbilleder (Landsat8, SPOT5) blev anvendt til kortlægningen af et opland. Derudover blev forskellige kemometriske metoder såsom PCA, PLSR anvendt til multivariat dataanalyse.

prøver udkonkurrerede modellen uden supplering. Denne undersøgelse viser at underopdeling efter de ovenævnte metoder forbedrer nøjagtigheden af kalibreringsmodellerne sammenlignet med anvendelse af modeller udviklet på hele det spektrale bibliotek.


5. Perspectives .................................................................................................................. 40
6. References ................................................................................................................... 41
7. Supporting papers ....................................................................................................... 47
List of Figures

Figure 1. The region of the electromagnetic spectrum ................................................................. 4
Figure 2. Different type of molecular vibration model ................................................................. 5
Figure 3. 1st derivative regression coefficient plot of the first factor from the PLSR for the
geographically closest .................................................................................................................. 6
Figure 4. Illustration of remote sensing ........................................................................................ 7
Figure 5. Map of Danish landscape types ..................................................................................... 14
Figure 6. Distribution of soil profiles for the Danish soil spectral library .................................... 15
Figure 7. Study area of Skjern river catchment and distribution of sampling points ............... 16
Figure 8. (1) Silstrup sampling site with 65 sampling points (black dots). (2) Søravad site with
125 samples; blue dots represent 50 reference samples. ......................................................... 17
Figure 9. Schematic diagram of a interferometer, configured for FT-NIR/IR .......................... 19
List of Tables

Table 1. NIR and MIR detectors ................................................................. 20
Table 2. General information of Landsat 8 Operational Land Imager (OLI) and Thermal Infrared Sensor (TIRS) ........................................................................................................... 21
Table 3. Triplicate SOC measurements from different laboratories and operators on different dates ........................................................................................................................................ 28
Table 4. Triplicate SOC measurements from the same laboratory and operator on different dates ........................................................................................................................................ 29
Table 5. General statistics on soil characteristics of the Danish soil library ............................................ 33
Table 6. Lime requirement for soil with different SOM contents based on the same actual pH value and soil type ........................................................................................................................................ 34
Table 7. Definition of soil types from the Danish soil classification ......................................................... 36
Table 8. Nitrogen fertilisation standards for Danish agricultural crops .................................................... 37
### Abbreviations/Acronyms

**ASD** - Inc Analytical Spectral Devices  
**C** - Carbon  
**CEC** - Cation Exchange Capacity  
**cLHs** - Latin Hypercube Sampling  
**DSM** - Digital Soil Mapping  
**EC** - Electrical Conductivity  
**FT** - Fourier-Transform  
**GPS** - Global Positioning System  
**JK** - Jack-Knifing  
**LBC** - Lime Buffer Capacity  
**MIR** - Middle Infrared  
**MSC** - Multiplicative Scatter Correction  
**PC** - Principal Component  
**PCA** - Principal Component Analysis  
**PLSR** - Partial Least Squares Regression  
**RS** - Remote Sensing  
**RMSEC** - Root Mean Square Error of Calibration  
**RMSECV** - Root Mean Square Error of Cross-Validation  
**RMSEP** - Root Mean Square Error of Prediction  
**RPD** - Ration of Performance to Deviation  
**RMSD** - Root Mean Square Deviation  
**RPIQ** - Ratio of Performance to Interquartile Distance  
**SEP** - Standard error of performance  
**SD** - Standard Deviation
**SDD** - Standard Deviation of Difference

**SNV** - Standard Normal Variate

**SOC** - Soil Organic Carbon

**SOM** - Soil Organic Matter

**N** - Nitrogen

**UV** - Ultraviolet

**Vis-NIR** - Visible Near Infrared
List of supporting papers

This dissertation is based on one published research paper, one accepted research paper and one manuscript, which are referred to in the text by their roman numerals.


III. **Peng Yi**, Xiong Xiong, Kabindra Adhikari, Maria Knadel, Sabine Grunwald, and Mogens Humlekrog Greve. Modeling SOC at Regional Scale by Combining Multi-spectral Images with Laboratory Spectra. Working manuscript include.
1. Introduction

1.1 Importance of soil

Soil is defined as the top layer of the Earth’s terrestrial surface. It is a mixture formed by physical and chemical processes of minerals, organic matter, air, water and living organisms (Hillel, 1998). Soil consists of a solid phase (minerals and organic matter) as well as a porous phase that holds gases and water (Paul, 2006). Soil provides a medium for plant growth, water processes, storage, filtration and transformation. Soil also acts as a modifier of the atmosphere and supplies most of the nutritional requirements for organisms (Hillel, 1998, Hillel, 2007). Because of these functions, soil is linked to all the global ecosystems and plays an extremely important role in food production, climate change and carbon storage.

Soil is essentially a non-renewable resource (Glanz, 1995). The main functions of soil are worthy of protection because of their socio-economic as well as agricultural and environmental importance. However, soil is an extremely complex, variable and living medium. Thus, understanding soil characteristics is key to being able to provide ecosystem services and apply sustainable land management, which are essential for human existence and economic prosperity.

1.2 SOC and clay

Soil carbon comprises the major land surface carbon pool, which is the largest carbon pool for terrestrial carbon sequestration (Batjes, 1996). Soil carbon can be present in both organic and inorganic forms. In most soils (with the exception of calcareous soils) the majority of carbon (C) is in the form of soil organic carbon (SOC) (Miegroet and Olsson, 2011). The content of SOC in soils ranges from less than 1% in sandy soils to almost 100% in wetland soils. Soil organic carbon is also the major component of soil organic matter (SOM), which is extremely important in all soil processes. Soil organic matter is essentially derived from residual plant and animal material or, synthesised by microbes and decomposed under the influence of temperature, moisture and ambient soil conditions (Brady and Weil, 2010). Consequently, almost all SOC comes from plants, with amounts originating from photosynthetic soil bacteria.

In order to understand soil behaviour and management, knowledge of the proportions of different-sized particles (e.g. soil texture) in a soil is needed (Brady and Weil, 2010). According to the standard of the International Society of Soil Science, clay is defined as a particle size smaller than 0.002 mm in diameter. In Denmark soils with more than 10% clay are considered to be fine textures or clayey. Clay also refers to a particular group of minerals. It consists of secondary minerals formed in the soil itself by the decomposition of primary minerals and their re-composition into new ones. The most prevalent minerals in the clay
fraction of temperate region soils are alumina-silicates, whereas in tropical regions hydrated oxides of iron and aluminium dominate (Hillel, 1998). Clay minerals can be classified into two types, 1:1 or 2:1, which are based on the number and arrangement of the tetrahedral silicate sheets and the octahedral hydroxide sheets present. The most common mineral in the 1:1 type is kaolinite, which consists of one tetrahedral sheet and one octahedral sheet. Montmorillonite is a typical 2:1 clay mineral type, which has an octahedral sheet sandwiched between two tetrahedral sheets (Brady and Weil, 2010, Hillel, 1998). In Denmark the mineralogy is dominated by illite.

1.3 Why is SOC and clay so important?
On the global scale, soils contain approximately 1500 Pg of carbon, which is around three times the amount of carbon in the terrestrial plant system (IPCC, 2000). Soil organic carbon plays a key role in the global carbon cycle and thus it is an important factor in climate change mitigation strategies (Schlesinger, 1999). In agriculture, SOC is one of the most important constituents of the soil and directly affects crop cultivation. In general, SOC performs the following two major functions for soil fertility: enhancing nutrient availability (plant residues provide a source of energy for plant growth and nutrients for soil microorganisms) and improving soil structure and soil physical properties (smaller plant residues promote soil structure by holding the soil particles together as stable aggregates, thus improving soil physical properties, and by increasing the cation exchange capacity (CEC) and water-holding capacity of sandy soil) (Leeper and Uren, 1993). Soils poor in organic matter have a smaller microbial biomass, activity, and nutrient mineralisation due to a shortage of energy sources. Additionally, SOM is used to estimate the soil pH buffer capacity, which is a very important parameter for calculating the lime requirement for crop production (Brady and Weil, 2010, Hillel, 1998, Hillel, 2007). Soil organic carbon, can be utilised as both an indicator of soil quality and as a strategy to offset CO2 emission by C sequestration.

Clay is another essential component which has a profound influence on agricultural soils. It contributes many benefits to the physical, chemical and biological properties of soils and supports plant growth. Due to its extremely small particle size, clay has very large specific surface areas and high number of pores between clay particles, which results in a high capacity to retain nutrients and water better than other soil types (Hillier, 1995, Brady and Weil, 2010). Additionally, clay particles carry a negative charge and will be naturally attracted to iron since most of these are positively charged (Brady and Weil, 2010). This characteristic is important because many positively charged ions are plant nutrients, such as calcium, magnesium and potassium. Although clay holds a large amount of water, much of it is not available to plants. On the other hand, the strong water-holding capacity of clay prevents nutrients (particularly nitrogen, potassium and sulphur) loss when rainfall percolates through the soil. Therefore, clay also plays an important role in the nitrogen
leaching process. Overall, clay plays a vital role in crop cultivation by retaining plant nutrients and moisture. Without clay, it would be very difficult to achieve successful crop production.

From the above description, it can be seen that the practice of precision agriculture and sustainable land management strongly depends on the variation of clay and SOC. In order to calculate the proper amount of lime and nitrogen fertiliser for crop production and minimise negative effects on the environment under different field conditions, the information on these two soil properties is critical for farmers. Consequently, understanding, monitoring and visualising variation of clay and SOC are needed for constructing proper decision-support systems for precision agriculture and sustainable land management. However, insufficient recognition and quantification of spatial and temporal variation of soil properties is a major stumbling block for implementing precision decisions (McBratney et al., 2005). In order to create precise and fine-resolution soil maps, more details of the variation in certain soil properties for site-specific agriculture applications, such as fertilisation and crop management are needed. In the past few decades, many soil and geoscientists have developed various methods to create digital soil maps, but these approaches need to be based on a large number of observations, which makes them costly and time-consuming (McBratney et al., 2003, Grunwald et al., 2011, Arrouays et al., 2014). Therefore, unlike the traditional wet chemistry approaches, a rapid and low-cost method for analysis of soil properties is highly required.

1.4 Spectral soil sensing

1.4.1 Vis-NIR and MIR spectra

Visible Near-Infrared (Vis-NIR) and Mid-Infrared (MIR) diffuse reflectance spectroscopy is becoming a robust analytical methodology, also in agriculture (Roberts et al., 2004). This technique is relatively low-cost, accurate and non-destructive. Also, no specific sample preparation is required prior to analysis, making it more efficient than many other methods. In the past twenty years, Vis-NIR and MIR spectroscopy methods have been developed for gathering quantitative and qualitative information on soil characteristics (Rossel et al., 2006, Du and Zhou, 2009, Stenberg et al., 2010, Bellon-Maurel and McBratney, 2011). By utilising this technique, all types of organic and many types of inorganic soil component information can be collected as optical data by the amount of radiation reflected from the soil sample. If soil samples contain chemical compounds such as C-H, N-H or O-H, this method can be used to quantify/qualify soil properties based on the relationship between spectra and reference data (Roberts et al., 2004, Williams and Norris, 2001).

The wavelength range of Vis-NIR and MIR spectroscopy is between the ultraviolet (UV) and the microwave region of the electromagnetic spectrum. It is commonly defined as the region
from 400 nm to 25000 nm (Fig. 1). By convention, the visible light region is from 400 nm to 780 nm, the NIR region is from 780 nm to 2498 nm, and the MIR region is from 2498 nm to 25000 nm. Many spectroscopy researchers also use the wavenumber (cm⁻¹) instead of length in nanometer to refer to wavelength. The formula to convert wavelength in nanometers to wavenumber is shown in Equation.1. This thesis will only refer to wavelength in nanometer.

\[
\text{Wavenumber} = \frac{10^7}{\text{nanometer}} \quad \text{(Eq. 1)}
\]

Infrared spectroscopy is the study of the interaction of radiated energy with matter. This technique is based on vibrations of atoms in a molecule (Stuart, 2004). Molecules with different structures vibrate at specific frequencies. The main molecular vibrations are bending and stretching (Fig. 2). When a molecule is irradiated with infrared light it can absorb radiation if the radiation has the same energy as the vibrational or rotational transitions of the molecule. In the NIR region, mainly multiples of the vibrational energy (overtones) are absorbed, whereas rotational and vibrational energy are absorbed in the MIR region (Marel and Beutelspacher, 1976). Additionally, light energy can be directly linked to the frequency or wavelength of the light (Roberts et al., 2004). This is why infrared spectroscopy offers the possibility to measure different types of interatomic bond vibrations at different frequencies.

Figure 1. The region of the electromagnetic spectrum after (Rossel et al., 2006)
As in the theory presented above, the first assumption when using spectroscopy as a technique to quantify the concentration of components in soil samples is the Beer-Lambert Law (Williams and Norris, 2001). The Beer-Lambert Law states that there is a linear relationship between absorbance and concentration of an absorbing species. The Beer-Lambert Law can be written as:

\[ A = \varepsilon \cdot b \cdot c \]  
(Eq. 2)

Where \( A \) is the value of absorbance, \( \varepsilon \) is the molar absorptivity, \( b \) is the path length of the sample and \( c \) is the certain concentration of the absorbing analyte. Based on this assumption, multivariate calibrations are required to mathematically extract spectral information and to correlate with soil properties (Martens, 1989). The details of multivariate data analysis will be described in Chapter 2.

Soil organic matter has broad absorptions in the NIR range of the electromagnetic spectrum, which is generally from the overtones and combination absorptions of O-H, C-H, and N-H (Clark, 1999, Rossel and Behrens, 2010). In Danish agricultural soil, we found that bands around 1449, 2037 and 2304 nm could be assigned to SOM (Figure 3) (Peng et al., 2013). Wavelengths around 1400 nm and 1900nm show very strong water absorption features in soil NIR spectra (Workman Jr and Weyer, 2007). Clay minerals such as kaolin and illite show absorptions in the Vis-NIR region due to metal-OH bend plus O-H stretch combinations (Hunt, 1970). Although the visible spectrum characterises only a small part of the electromagnetic spectrum (blue: 450-515 nm, green: 525-605 nm and red 630-690 nm), it is an essential part of the spectrum for measuring soil colour, which reflects the soil organic
matter and iron contents of soils (Fernandez and Schulze, 1987, Islam et al., 2004). Thus, visible absorption spectroscopy is often combined with the NIR region to gather information on SOM and clay minerals (Stenberg et al., 2010, Rossel et al., 2006). The absorption values for soil compounds detected in the Vis-NIR region often overlap, making it more difficult to interpret when comparing to the MIR region. However, the Vis-NIR region still contains useful information for SOC quantitative analysis (Reeves, 2010, Bellon-Maurel and McBratney, 2011, Gholizadeh et al., 2013).

Figure 3. 1st derivative regression coefficient plot of the first factor from the PLSR for the geographically closest model (Paper 1)

The MIR spectrum can be classified into four regions. Generally, these regions are approximately split as follows: the X-H stretching region (2500–4000 nm), the triple-bond region (4000-5000 nm), the double-bond region (5000-6666 nm) and the fingerprint region (6666-16666 nm). The absorptions in MIR spectroscopy are associated with fundamental bonds of molecular vibrations. The fundamental vibrations in the stretching region are generally due to O-H, C-H and N-H stretching. Triple and double bonds such C=C, C=O and C≡C can be easily recognised in the MIR spectrum (Marel and Beutelspacher, 1976, Smith, 1999, Stuart, 2004). It has been assumed that each band in a MIR spectrum can be assigned to a specific deformation of the molecule, the movement of a group of atoms, or the bending or stretching of a particular bond. Thus, in organic chemistry the analysis of MIR absorption spectra shows what types of bonds are present in the sample, and this method can be alternatively applied for qualitative analysis on soil under the laboratory condition.
1.4.2 Remote sensing spectra

Remote sensing (RS) is the art and science of obtaining information about an object, area or phenomenon from a distance (Jensen, 2009, Lillesand et al., 2004). Modern RS as a technology started with the first photographs in the early nineteenth century. The output of a RS system is usually an image representing the scene being observed. Nowadays, RS is applied to observe important biophysical characteristics and human activities on Earth. One of the biggest advantages of this technique is to provide images of areas in a fast and cost-efficient manner and facilitate sampling at large-scale study areas. Using the same fundamentals as the Vis-NIR technique described above, the RS technique uses various aerial remote sensors to detect and collect data by means of electromagnetic radiation. This data source provides important coverage, mapping and classification of land cover such as vegetation, soil, rock, water and urban infrastructure (Schott, 2007, Schowengerdt, 2006). Remote sensing can be commonly divided into active and passive RS (Liu and Mason, 2009) (Figure. 4). In this thesis, we only focus on passive RS. In passive RS, optical sensors detect solar radiation reflected or scattered from the earth, and because different materials have different colours and brightness under the sun, different types of land cover reflect visible and infrared light at different wavelengths. The wavelength region usually extends from the visible and near infrared (VNIR 400-1100 nm) to the short-wave infrared (SWIR 1100-2500 nm).

Figure 4. Illustration of remote sensing
As discussed in the previous section, Vis-NIR and MIR techniques are capable of estimating soil properties under laboratory conditions. However, one of the drawbacks of this method is that it only obtains soil spectral information from a specific sampling point, therefore usually called point spectrometry. In order to obtain spatial spectral information at field, regional or national scale, soil science needs the RS technique, which has the advantage over point spectrometry of a larger spatial domain and a readily applicability under field conditions. Additionally, RS images can be gathered in a short time and provide more detailed information of the target field. This work cannot be achieved by any traditional soil sampling work or laboratory spectral measurements (Ben-Dor et al., 2009, Ben - Dor et al., 2008). In the past few decades, soil and geoscientists have used RS information as a first or second data resource to observe, quantify or map soil properties such as SOC, SOM, soil moisture and soil erosion (Hill et al., 1993, Hill et al., 1994, McKenzie and Ryan, 1999, Ben-Dor et al., 2003, Whiting et al., 2004, Fernandez-Buces et al., 2006, Stevens et al., 2006, Rivero et al., 2007, Ben-Dor et al., 2008, Haubrock et al., 2008, Schmid et al., 2008, Stevens et al., 2008, Rivero et al., 2009, Ge et al., 2011, Grunwald et al., 2011, Mulder et al., 2011).

Although RS has attracted a lot of attention in many disciplines, this technology is not well developed and more research is still required. This is because RS spectra have limitations compared to laboratory spectra, such as a low signal-to-noise ratio, atmosphere attenuation, a low integration time for a given pixel, a spectral mixing problem, optical shifts from one pixel to another, and bidirectional reflectance distribution functional effects (Ben-Dor et al., 2008). Moreover, this technique cannot be regularly used by farmers because the application of RS for soil analysis is relatively costly and time-consuming due to image correction procedures (Ben-Dor et al., 2009). Thus, McBratney et al. (2003) proposed the simultaneous application of multiple technologies for soil mapping for optimal use of the digital soil mapping (DSM) system. Many researchers therefore suggested the creation of a large database consisting of high-quality soil spectral information covering a wide range of different soil types at different geographic scales (Brown, 2007, Shepherd and Walsh, 2002). This database is commonly referred to as a soil spectral library.

1.5 Soil spectral library

Visible near-infrared and mid-infrared spectroscopy are some of the promising analytical techniques to replace routine laboratory wet chemistry soil analysis. Because spectroscopy technique still requires reference data from wet chemistry analysis for model development, it is still costly. Therefore, to minimise the number of calibration samples required for local applications, the construction of a soil spectral library to cover all the variations in different soil types at regional, national and global scale is recommended (McCarty et al., 2002, Brown, 2007, Brown et al., 2006, Shepherd and Walsh, 2002). In the last ten years, different scales of soil spectral libraries have been built and become a classical tool for studies as a basic

The first global soil spectral library was constructed by Brown et al. (2006), which included 4184 soil samples from 37 countries, but only 416 samples from outside of the U.S. Based on this database, they found a strong relationship between Vis-NIR spectra and soil properties such as clay, CEC and SOC. However, in terms of a global spectral library for calibration purposes, they estimated that a complete global soil spectral library would require approximately $10^9$ soil samples to represent all types of soil around the world. Therefore, the strategy of constructing national soil spectral libraries from different countries is an alternative to a complete global library.

A well-developed national soil spectral library would include samples either geographically covering the entire country or adequately describing the soil variability in the country in which the library is to be used (Rossel et al., 2008). According to this definition, to the best of the author’s knowledge, only Denmark, Sweden, France and Australia have so far completed their soil Vis-NIR spectral library (Goge et al., 2012, Peng et al., 2013, Rossel and Webster, 2012, Wetterlind and Stenberg, 2010). Based on these four national spectral libraries, from different studies, a total of 25 properties were predicted by different calibration methods. The results showed that the prediction of SOC and clay generally gave acceptable results. Of these four databases, the Danish and French spectral library samples were collected from regular, nationwide 7 km and 16 km grids, respectively. The Australian database consists of 21,493 samples from approximately 4000 profiles, which represent all of the orders in the Australian soil classification. Compared with these three soil spectral libraries, the Swedish library is relatively small, consisting of 396 samples representative of Swedish agricultural land. Constructing a national soil spectral library in a large country like Australia was a long process that took many years. In order to take full advantage of soil spectral libraries for DSM and precision agriculture as early as possible, soil spectral libraries based on the regional or local scale was suggested.

Various regional soil spectral libraries were constructed based on different study areas in the last 10 years. Cambule et al. (2012) developed a soil spectral library with 129 samples in Limpopo National park in Mozambique with an area measuring about one million hectares, a data- and access-limited area. To estimate SOC, they used two thirds of the samples for calibration, and the rest of them for validation. Meanwhile, 20 samples were randomly selected for duplicate analysis to quantify laboratory precision. The validation results gave a root mean error of prediction (RMSEP) of 0.32% SOC, about 2.5 times higher than laboratory uncertainty. This work could not yield accurate prediction results due to the limitations of sample size and lack of an established soil spectral library. However, the author still
suggested that the model was suitable for estimating soil properties in further mapping exercises in this area. Another regional soil spectral database was developed for the south-east of Spain by Guerrero et al. (2010) with 1040 soil samples. They used this library to predict nitrogen contents in unknown samples from a specific site, but all of the library samples were at least 20 km away from the target site. In this situation, the calibration model from this library might not contain sufficient spectral variation to directly predict unknown samples from a new target site. Thus, a method called “spiking” of the library and recalibrating the model has been proposed in recent soil science literature.

Spiking is the development of model functions that are calibrated on samples from a large or broader dataset in order to be used in a small or specific unknown dataset. Spiking is now commonly applied in soil spectroscopic modelling (Brown, 2007, Guerrero et al., 2010, Sankey et al., 2008, Goge et al., 2014). This method combines a limited number of local samples with the soil spectral library and subsequent re-calibration of the model for further predictions. Several studies have showed that spiking improves the final prediction accuracy compared with calibration from a large dataset, mainly by reducing the RMSEP value (Brown, 2007, Guerrero et al., 2010, Peng et al., 2013). However, Wetterlind and Stenberg (2010) spiked 25 local samples with a regional calibration, and the prediction results did not remarkably outperform local calibration with the same 25 samples. They explained that these 25 local samples covered all the soil variation at the study site in contrast to the local calibration samples in Sankey et al. (2008). For many areas in soil spectroscopic modelling, this method is still problematic, since spiking obviously does not work in all cases. The remaining questions relating to this method are: (1) Is it really necessary to spike? (2) If it is necessary to spike, how do you select samples and the optimal number of samples, and weight of spiking set? (3) Are there other ways of efficiently applying a soil spectral library to unknown samples?

An alternative way of applying the spectral library for field predictions could be to screen the samples from an existing soil spectral library for local calibrations. The choice of selection method depends on all available characteristic of target samples, and selected samples should be spectrally similar to the unknown samples. This can be done in many ways, such as using statistical and geographical approaches. Peng et al. (2013) made a geographical selection of around 100 samples from the Danish national soil spectral library to build a model, with validation samples from a Danish agricultural field. The prediction of SOC gave an acceptable error when compared with lab error. Nevertheless, due to the complexity of the soil matrix, the approach of creating subsets of samples spectrally similar to unknown samples is still facing challenges.
1.6 Error sources of modelling

In many soil wet chemistry analyses, chemical or physical properties may be determined imprecisely, in relative or absolute terms. For instance SOC and clay analyses could yield significant measurement errors and we may never be able to obtain the true value of the samples since even apparently identical samples will tend to produce different results. This may be caused by reference method error, error in the measurement process, or both. Therefore, all these errors could result in the signal-to-noise ratio reducing during the modelling process (Esbensen et al., 2002).

There are two types of common measurement error which can be defined in analytical science: systematic error and random error. Random error is an error in measurement that leads to measurement values differing when repeated measures of a same attribute or quantity are taken. This error is unpredictable, scattered about the true value. Random error can be estimated by comparing re-measurements, and reduced by averaging re-measurements. This error is closely related to the concept of precision. The precision of a measurement system is how closely multiple measurements of the same quantity agree with each other, which is related to reproducibility and repeatability (BIPM et al., 2008, Taylor, 1997).

Systematic error is mainly due to the measurement system, such as reference methods, or the instruments used. This error cannot be eliminated by repeated measurement, because it can be constant and it always pushes the results in the same direction. One of the most common methods to remove systematic error is through calibration of the measuring instrument. This error is closely related to the concept of accuracy, and the accuracy of a measurement system is how close the measurement is to the true value of the quantity being measured.

Based on these two major error sources, we can summaries that:

\[
\text{Observed score} = \text{True score} + \text{random error} + \text{systematic error}
\]

It is important to note that error sources accumulate throughout the whole process of sampling, preparation and measurement until the modelling stage. Consequently, these errors could result into model error and prediction error. During the modelling process, we expect to replace reference values with predictions based on spectral data. In this situation, the reference values are regarded as the “true” values. However, the prediction results from the model cannot be exactly the same as the reference values if there is noise, errors or inaccuracy in reference measurements. Thus, we have to be aware of the major error components in order to understand or evaluate the final results (Faber and Kowalski, 1997, Sørensen, 2002,
Sørensen and Dalsgaard, 2005, Mark et al., 1989, Aastveit and Marum, 1991). In general, the major error sources can originate from: (1) inhomogeneity of the sample; (2) sample preparation and the taking of sub-samples for analysis, uncertainty of lab measurements or sampling; (3) instrument inaccuracy, drifts from both spectrometers and laboratory wet chemistry soil analyses, which contribute to errors in spectral data and reference data, and (4) modelling errors (Esbensen et al., 2002).

1.7 Digital soil mapping

Conventional methods for mapping soils have plenty of limitations (Simonson, 1989, McKenzie et al., 2008, Hartemink et al., 2010). The DSM and modelling techniques have become ever more popular in recent decades, and have the potential to overcome some of the drawbacks of conventional soil mapping (McBratney et al., 2003, Hartemink et al., 2008, Boettinger et al., 2010, Grunwald et al., 2011, Arrouays et al., 2014). The international working group on digital soil mapping defined DSM as "the creation and the population of a geographically referenced soil database generated at a given resolution by using field and laboratory observation methods coupled with environmental data through quantitative relationships." (Lagacherie et al., 2006)

An empirical model called scorpan was proposed by McBratney et al. (2003) to quantitatively describe relationships between soil and environmental factors. The model can be written as:

\[ S=f(s, c, o, r, p, a, n) \]  \hspace{1cm} (Eq. 3)

Where:

\( S \) = soil classes or soil attributes

\( s \) = soils, other properties of the soil at a point

\( c \) = climate, climatic properties of the environment at a point

\( o \) = organisms, vegetation or fauna or human activity

\( r \) = topography, landscape attributes

\( p \) = parent material, lithology

\( a \) = age, the time factor

\( n \) = space, spatial position
Based on the above concept, various DSMs have been conducted in different parts of the world in the last few years. Recently, Adhikari et al. (2014) combined a Cubist data-mining tool with geo-statistics and used it to construct a soil class map of Denmark based on the FAO legend and high-resolution (30 m) 3-D Danish national soil texture maps (clay, silt fine sand and coarse sand content) were generated by Adhikari et al. (2013). Seventeen environmental variables were involved in this work and modelled the continuous texture distribution at six standard soil depths down to 2 m depth.

1.8 Objectives & hypothesis

In this thesis, we hypothesised that the techniques of laboratory Vis-NIR and MIR spectroscopy combined with different chemometric tools are able to effectively quantify SOC and clay content. The general aim of the thesis was to explore the optimal methods for field-scale predictions of SOC and clay content using Vis-NIR and MIR spectroscopy based on the Vis-NIR spectral library and different chemometric tools to improve digital soil mapping at regional scale.

Specific objectives were:

- To find a new approach for creating an optimal/representative subset of a soil spectral library for calibration and prediction of SOC at field scale.
- To test the effect of spiking a library with local/field samples for SOC predictions.
- To select important spectral features for SOC and clay calibration/prediction using different chemometric tools.
- To find the optimal number of replicates for Vis-NIR and MIR measurements.
- To develop a new approach for estimating the uncertainty of wet chemistry analyses at low cost.
- To evaluate Vis-NIR and MIR techniques as a replacement for standard wet chemistry soil analyses.
- To determine the acceptable range of prediction accuracy for SOC and clay contents for liming and nitrogen fertilisation.
- To combine laboratory Vis-NIR spectra with RS spectra to map SOC at regional scale.
2. Materials and Methods

2.1 Study location - Denmark

Denmark is a country in Northern Europe, it forms part of Scandinavia and comprises an area of approximately 43,000 km². It is a low-lying country with a mean elevation of about 32 m and the highest point about 171 m above mean sea level. Denmark has a temperate maritime climate with a mean temperature of 0°C in winter and 16°C in summer (Danmarks Meteorologiske Institut, 1996). The average annual precipitation varies between 500 to 800 mm from east to west. The precipitation exceeds evapotranspiration and water leaches through the soil during the winter and early spring season. Over 60% of the country is used for cultivation, and the main crops grown are wheat, maize, potato and barley. Only 10% of the land is covered by forest.

Figure 5 shows the landscape types of Denmark. In general, the landscape of the central and eastern parts of Denmark is dominated by a glacial (Weichselian) morainic landscape. In this area, soil types are mainly loamy soils on calcareous tills. The western part of the country is defined as a glacial flood plain with a coarse sandy soil, due to low-relief glaciofluvial sandy sediments from glacial meltwater. The northern, central and eastern parts of the country were covered by ice during the Weichselian glaciation. The northern part of Jutland was covered by a glacier advancing from Norway, while the soil in central and eastern Denmark was covered by a glacier advancing from the eastern Baltic Sea. (Danmarks Meteorologiske Institut, 1996).

Figure 5. Map of Danish landscape types
2.2 Soils and Sampling

2.2.1 National scale sampling (Danish national library)
The Danish soil profile database was used to build the Danish national spectral library, which was used for SOC prediction of samples from the field at Silstrup (Paper I). This library consists of 700 soil profiles gathered on a nationwide 7-km soil monitoring grid from 1987-1990 (Fig. 6) (Greve et al., 2007a). The entire library included 2851 soil samples from different horizons from the soil surface down to 1.8-2.5 m. Soil properties were measured for each sample such as SOC and texture. SOC was determined by dry combustion methods, and soil texture was analysed using hydrometer and wet sieving methods into four fraction sizes: clay (<2 μm), fine silt (2-20 μm), coarse silt (20-63μm), fine sand (63-200 μm) and coarse sand (200-2000 μm). In order to collect soil spectra and construct the soil spectral library, each sample was scanned by a LabSpec 5100 spectrophotometer (ASD Inc., Boulder, Colorado, USA); the scanning procedure was according to a laboratory protocol.

![Figure 6. Distribution of soil profiles for the Danish soil spectral library](image)

2.2.2 Regional scale sampling (Skjern river catchment)
Because groundwater resources provide more than 90% of all drinking water in Denmark, and an initiative to improve understanding of hydrological processes. The study area of Skjern river catchment was selected for a Hydrological Observatory (HOBE). Hydrologic parameters are highly affected by SOC and therefore explored the possibility of combining
laboratory spectra with RS spectra for SOC digital mapping at regional scale (Paper III). The study area is located on the west coast of Jutland (Fig. 7) and covers an area of approximately 2500 km². The climate in this region is temperate maritime with a mean annual precipitation of 990 mm, and mean annual temperature of 8.2°C (Jensen and Illangasekare, 2011). The predominant soil type is loamy sand (Greve et al., 2007a). The topography slopes gently from east to west, with maximum elevations around 125 m above sea level. Over 50% of land use is agriculture, growing cereals and crops with the remainder covered by grass (30%), forests (7%) and heathland (5%) (Jensen and Illangasekare, 2011).

A total of 328 topsoil samples were collected from the study area during 70’s. All samples were originally from the Danish soil classification that involved 36000 sites that were sampled (0-25 cm) for texture and SOC analysis (Greve et al., 2007b). Each sample was composed of 25 subsamples which were taken within a 70 × 70 m square. SOC was determined by combustion with a LECO IR-12 furnace (Krogh et al., 2003). Soil texture was determined only for samples with SOC values lower than 6% (in 296 samples).

Figure 7. Study area of Skjern river catchment and distribution of sampling points

2.2.3 Field scale sampling (Silstrup, Sørvad sites)

In order to apply the national soil spectral library to field-scale prediction, 65 cylindrical 20 × 20 cm topsoil samples were collected from a 15 × 15 m grid at the Silstrup site, on 10 October 2010 (Fig. 8). The field has been farmland since at least 1942 and measures 185 × 91 m. The most dominating geological feature is a salt dome called the Thisted structure, which considerably influences the local geology around the Silstrup area (Lindhardt et al., 2001). According to the USDA soil classification system, the soil type at the Silstrup site is classified
as Alfic Argiudoll and Typic Hapludoll for two pedological profiles (Lindhardt et al., 2001). The details of sample preparation can be found in Paper I.

Figure 8. (1) Silstrup sampling site with 65 sampling points (black dots). (2) Sørvad site with 125 samples; blue dots represent 50 reference samples.

The Sørvad site is a heterogeneous site with contrasting soil types and was selected for the HOBE II project. It represents the Salian Moraine landscapes characterised by mixed/heterogeneous soils. The site is located on a large elevated ‘island’ surrounded by outwash plains. Such elevated islands are the remains from the Saale ice age and mainly consist of sandy moraine deposits. The annual precipitation and the annual temperature are around 850 mm and 7.5°C, respectively. In the fine textured part of the field, the soil type is defined as a Hapludult, and in the coarse part of the field is defined as haplortod, according to the USDA classification. At this site, previous investigations have collected a wealth of valuable data such as soil horizon characterisations, pF curves and soil hydraulic conductivity. The
background data can be used for verifying the validity of the assessments of texture, soil water-holding capacity, etc.

A total of 125 sampling locations were selected using the conditioned Latin hypercube sampling (cLHS) strategy on soil electrical conductivity (EC) data and terrain attributes (Figure 8). This sampling strategy was suggested by Minasny and McBratney (2006) initially for DSM. Fifty representative samples were selected using the Kennard-Stone algorithm (Kennard and Stone, 1969). This re-sampling method is essentially a stepwise procedure that maximises the Euclidean distance according to the important number of principal components for the objects already chosen. The calculation was conducted using R with the soil.spec package (Terhoeven-Urselmans et al., 2010). The 50 samples were measured by Vis-NIR and MIR instruments. The details for the measurements can be found in Paper II.

2.3 Research instrumentation

2.3.1 Vis-NIR and MIR instrumentation

During the past few decades, great improvements have been made for Vis-NIR and MIR instruments and various spectrometers have been designed. Anything from simple filter spectrometers to high-speed Fourier transform spectrometers have been manufactured for many applications. On the basis of their properties, NIR instruments fall into one of two categories: dispersive or Fourier-Transform (FT) (McClure, 1994). The Vis-NIR and MIR instruments employed in this thesis were FT-NIR and FT-IR spectrometers, respectively.

The FT-NIR/IR spectrometer is an interferometer that produces a very high signal-to-noise ratio and speed. This type of instrument contains a beam splitter and two plane mirrors, one stationary and one moving. Light from the source is split into two equal power beams by a beam splitter. One is directed at the fixed mirror and the other at the movable mirror. After reflection from the plane mirrors, the beams are recombined and sent to the detector having first been subjected to a beam of polychromatic radiation (Fig. 9). The variations in the intensities of the combined beams can be measured as a function of path length differences (Griffiths and De Haseth, 2007).
The spectrometers are now more reliable and rugged than they used to be. A better quality of instrument gives a higher signal-to-noise ratio, which means the amount of signal produced by the instrument is higher than the noise in the instrument. The main significant improvement is the reduction in instrument noise and may be attributed to better manufacturing of detectors (McClure, 1994). Therefore, the detector is the most important part of the instrument, since its quality (noisy, sensitivity) can result in different qualities of data. There are two basic types of detector, which can be classified according to the principle of operation: thermal detectors and photon detectors (Roberts et al., 2004). Photon detectors are commonly used for NIR applications, because the response time of thermal detectors is too long. On the other hand, a unique advantage of thermal detectors is their uniform sensitivity at all wavelengths (Williams and Norris, 1987). Hence, the choice of photon detectors is usually decided by the wavelength range and other design parameters of the spectrometer, such as sample presentation. Furthermore, photo sensitivity (responsivity), noise equivalent power (NEP) and detectivity are also important characteristics of detector performance. The typical photon detectors used in Vis-NIR and MIR instruments are: lead sulphide (PbS) and lead selenide (PbSe), silicon, indium arsenide (InAs), indium gallium arsenide (InGaAs), indium antimonide (InSb) and lead telluride (PbTe) (Table 1). Although the photo sensitivity of these detectors depends on wavelength, they generally provide high detection performance (Roberts et al., 2004, Williams and Norris, 1987).
### Table 1. NIR and MIR detectors

<table>
<thead>
<tr>
<th>Detector</th>
<th>Useful detection range (nm)</th>
<th>Region</th>
<th>Remark</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lead sulphide (PbS)</td>
<td>1100-2500</td>
<td>NIR</td>
<td>PbS ‘sandwiched’ with silicon photodiodes, are often used for Vis-NIR</td>
</tr>
<tr>
<td></td>
<td>400-2600</td>
<td>Vis-NIR</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1100-4500</td>
<td>NIR-MIR</td>
<td></td>
</tr>
<tr>
<td>Lead selenide detectors (PbSe)</td>
<td>1100-5000</td>
<td>NIR-MIR</td>
<td>High detection capability and fast response speed</td>
</tr>
<tr>
<td>Silicon</td>
<td>300-1100</td>
<td>Vis-NIR</td>
<td>High speed, Sensitivity and low cost, optimised for the Vis-NIR region with typical response spectral regions between 853-950 nm</td>
</tr>
<tr>
<td>Indium arsenide (InAs)</td>
<td>1700-5700</td>
<td>NIR-MIR</td>
<td>Very high detectivity</td>
</tr>
<tr>
<td>Indium gallium arsenide (InGaAs)</td>
<td>700-1700</td>
<td>Vis-NIR</td>
<td>Very high detectivity</td>
</tr>
<tr>
<td>Indium antimonide (InSb)</td>
<td>1800-6800</td>
<td>NIR-MIR</td>
<td>Very high detectivity</td>
</tr>
<tr>
<td>Lead telluride (PbTe)</td>
<td>1500-4500</td>
<td>NIR-MIR</td>
<td></td>
</tr>
</tbody>
</table>

#### 2.3.2 Remote sensing instruments (Landsat 8 & SPOT 5)

The eighth satellite in the Landsat programme — Landsat 8 — was launched on 11 February 2013. Landsat 8 is an American Earth observation satellite that carries the Operational Land Imager and the Thermal Infrared Sensor, two sensors that take images of the entire planet every 16 days. Eleven spectral and thermal bands with different spatial resolutions are collected from these two sensors (Table 2). In this thesis, only bands 2–7 were included for the DSM (Paper III). Generally, all the spectral information is useful for different parameters of earth surface mapping. Specifically, the blue band (band 2) is used for distinguishing soil from vegetation and deciduous from coniferous vegetation; in order to assess plant vigour, the green band (band 3) is used to identify peak vegetation; band 4 (red band) focuses on discrimination of vegetation slopes; band 5 (NIR) mainly discerns biomass content and shorelines and short-wave infrared (bands 6 and 7) are used for detecting moisture content of soil and vegetation. The details of vegetation indices derived from image spectra can be found in Paper III.
Table. 2 General information of Landsat 8 Operational Land Imager (OLI) and Thermal Infrared Sensor (TIRS)

<table>
<thead>
<tr>
<th>Bands</th>
<th>Wavelength (micrometres)</th>
<th>Resolution (metres)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Band 1 – Coastal aerosol</td>
<td>0.43 - 0.45</td>
<td>30</td>
</tr>
<tr>
<td>Band 2 – Blue</td>
<td>0.45 - 0.51</td>
<td>30</td>
</tr>
<tr>
<td>Band 3 – Green</td>
<td>0.51 - 0.59</td>
<td>30</td>
</tr>
<tr>
<td>Band 4 – Red</td>
<td>0.64 - 0.67</td>
<td>30</td>
</tr>
<tr>
<td>Band 5 – Near Infrared</td>
<td>0.85 - 0.88</td>
<td>30</td>
</tr>
<tr>
<td>Band 6 – SWIR 1</td>
<td>1.57 - 1.65</td>
<td>30</td>
</tr>
<tr>
<td>Band 7 – SWIR 2</td>
<td>2.11 - 2.29</td>
<td>30</td>
</tr>
<tr>
<td>Band 8 – Panchromatic</td>
<td>0.50 - 0.68</td>
<td>15</td>
</tr>
<tr>
<td>Band 9 – Cirrus</td>
<td>1.36 - 1.38</td>
<td>30</td>
</tr>
<tr>
<td>Band 10 – Thermal Infrared 1</td>
<td>10.60 - 11.19</td>
<td>100</td>
</tr>
<tr>
<td>Band 11 – Thermal Infrared 2</td>
<td>11.50 - 12.51</td>
<td>100</td>
</tr>
</tbody>
</table>

SPOT 5 (Satellite Pour l’Observation de la Terre) is a high-resolution satellite with five bands and three different resolutions; the satellite was launched in May 2002. Only four bands were used in this thesis: band 1 (green) was defined as 0.50-0.59 μm with 10-m resolution; band 2 (red) was defined as 0.61-0.68 μm with 10-m resolution; band 3 (NIR) was defined as 0.79-0.89 μm with 10-m resolution and band 4 (SWIR) was defined as 1.58-1.75 μm with 20-m resolution. The SPOT programme was designed to explore the Earth’s resources, detect and forecast phenomena involving climatology and oceanography, and monitor human activities and natural phenomena. Similar to Landsat 8 images, SPOT 5 images have different functions for each spectral band. The green band is mainly for oceanographic applications and atmospheric corrections. The red band and the NIR band essentially detect vegetation and photosynthetic activity. Band 4 is useful for measuring ground and vegetation humidity.

2.4 Chemometrics

2.4.1 Spectral pre-processing

The foundation for using spectroscopy for soil analysis is Beer’s Law. However, this theory is strictly applicable only to transmission measurements on transparent and homogeneous
materials. The measurement will encounter scattering along with absorption if there is a non-homogeneous distribution of the particles in the sample, as in the case of soils (Gobrecht et al., 2014). Scattering could lead to a non-linear relationship between the spectrum and the concentration due to a multiplicative effect, and the useful part of spectral information is relatively small mainly because of an additive effect (Fig. 10). Thus, in order to remove or at least minimise these physical phenomena in spectra and enhance the useful information, complex mathematical spectral pre-processing is required (Rinnan et al., 2009).

![Figure 10. Representation of additive and multiplicative effects in diffuse material](image)

In general, before chemometric modelling, spectral pre-processing is the most important procedure following proper data collection. The pre-processing methods applied in this thesis are: the Savitzky–Golay 1st derivative and the Savitzky–Golay 2nd derivative (Savitzky and Golay, 1964), multiplicative signal correction (MSC) (Geladi et al., 1985) and standard normal variate (SNV) (Barnes et al., 1989). The Savitzky–Golay derivative is a method for numerical derivation of a vector that includes smoothing steps. This is a pre-processing method has the capability of removing both additive and multiplicative effects in the spectra (Rinnan et al., 2009). It uses information from a localised segment of the spectrum to calculate the derivative at a particular wavelength (Williams and Norris, 2001). The MSC is a method to remove the effects of scattering by linearising each spectrum to an average of all the samples (Geladi et al., 1985). SNV is very similar to MSC correction. It is a correction for effects of the multiplicative interferences from light scattering (Barnes et al., 1989).

On the other hand, applying wrong type or pre-processing methods that too severe could result in valuable information being missed. The choice of a suitable pre-processing method should always be considered. Therefore, the proper choice of pre-processing method can normally be assessed by comparing validated models.
2.4.2 Exploratory data analysis

One of the optimal algorithms for exploring high-dimensional data sets such as spectral data and for calculating all possible variation in the spectra is Principle Component Analysis (PCA) (Hotelling, 1933). The PCA is a technique that is useful for compression and classification of data. The purpose of PCA is to discover or reduce the dimensionality of the data set and identify new meaningful underlying variables called principle components (Esbensen et al., 2002). The number of principle components (PC) has to be less than or equal to the number of the variables.

The basic PCA structure can be given by a model with a given number of PC such as the follow equation:

\[ X = TP^T + E \]  \hspace{1cm} \text{(Eq. 4)}

where \( T \) is the scores matrix, \( P \) is the loadings matrix and \( E \) is the error matrix. The most informative part of the data structure is the combination of scores and loadings. The entire structure of the original data set will be reconstructed by multiplying these two parameters. The remaining data set is called error or residual, which cannot be modelled well. Understanding scores and loadings is very important for the interpretation of the PCA results.

Scores show the differences and similarities among all the samples, and describe the data structure according to sample patterns. Each sample has a score based on a PC. Scores reflect the coordinate of the sample along that PC and describe the major features of the sample. Thus, similar samples are supposed to have close scores along the same PC. Conversely, samples for which the scores are quite different are located far away from each other with respect to those variables.

Loadings are the correlation between the original variables and the factors, and are the key to understanding the basic properties of each factor. Similar to scores, every variable analysed has a loading on each PC, which gives information about how many variables contribute to that PC. In geometrical terms, each loading value is calculated by the cosine of the angle between the variable and each specific PC, which means the smaller the angle, the larger the loading value. Therefore, the range of the loading value is between -1 and +1 if two variables have high loadings along the same PC with the same sign, which means their angle is small, and in turn means that the two variables are positively correlated.
2.4.3 Calibration and validation

Partial Least Squares Regression (PLSR) is one of the most commonly used chemometric tools for calibrating spectral data (Wold et al., 2001). It is used to find the fundamental relations between a variable data matrix and a response variable data matrix (X and Y). The PLSR is particularly suitable when the matrix of predictors has more variables than observations and when there is multicollinearity among X values. The PLSR model aims to find the latent variables in the X space that explain the direction of the maximum multidimensional variance in the Y space (Esposito Vinzi et al., 2010). The PLSR factors are similar to PC in PCA. In general, scores and loadings are also used to interpret the PLSR model, and the way to explain the model is basically the same as for PCA, but in PLSR, two different sets of components from the X- and Y-space can be considered.

In order to choose the optimal number of factors for the PLSR model and make sure that the model will work in the future for new samples, cross-validation and test-set validation were applied in this thesis. The choice of number of factors was determined by minimising the prediction error of cross-validation. Cross-validation means the same samples are used both for model estimation and for testing in different ways, such as full cross-validation and segmented cross-validation. The test validation is based on testing the model on a subset of target samples that will not be involved in the model computation. The sample variations of the calibration set commonly cover the variations of the test set. Test set validation was only performed in Paper I and Paper III.

Variable selection is an important step in improving model performance in multivariate analysis, because the removal of many non-significant variables will simplify the model and generally lead to better predictions and better interpretation of the models. We aim for a model that is adequate while using as few variables as possible. It is particularly important to reduce the number of variables for the spectroscopic data obtained from high-resolution instruments, due to the number of variables. Therefore, two variable selection methods called jack-knifing (JF) and Cubist were applied in papers II and III. The details for these two methods can be found in the supporting papers.

2.4.4 Model evaluation

In order to evaluate how good a model is, several statistical parameters were evaluated to test prediction reliability, such as coefficient regression ($R^2$), root mean error of calibration, cross validation and prediction (RMSEC, RMSECV, RMSEP), standard error of performance (SEP) and bias (Esbensen et al., 2002). The relationship between RMSEP, SEP and bias is:
\[ \text{RMSEP}^2 \approx \text{SEP}^2 + \text{Bias}^2 \]  \hspace{1cm} (Eq. 4)

Another parameter called the ratio of performance to deviation (RPD) was also applied in Paper I. The RPD is calculated as \( RPD = \frac{SD}{RMSEP} \), where SD is the standard deviation of the validation set. The threshold of this parameter for assessing the ability of NIR spectra to predict soil properties was defined by Chang et al. (2001) where Category A: \( RPD > 2.0 \), Category B: \( RPD \ 1.4 \text{–} 2.0 \) and Category C: \( RPD < 1.4 \). This then becomes a ‘standard’ classification: (1) excellent models, with \( RPD > 2 \); (2) fair models, with \( 1.4 < RPD < 2 \); and, (3) non-reliable models, with \( RPD < 1.4 \).

As discussed in the introduction, many researchers have proved the feasibility of using Vis-NIR and MIR spectroscopy approaches to quantify soil properties. However, for this technique to replace routine laboratory analysis, accuracy of the prediction model is crucial and should be comparable to the accuracy of traditional laboratory chemistry soil analysis. Since the RMSEP value is estimated using reference values that can be inaccurate, it is necessary to include laboratory uncertainty and compare with the prediction error of soil properties. The soil spectral technique is potentially ready to replace traditional wet chemistry analysis for SOC and clay if we can constantly produce RMSEP values that are lower, equal or even slightly higher than laboratory error. In this thesis, the standard deviation of difference (SDD) and root mean square deviation (RMSD) were calculated to indicate part of the laboratory uncertainty for soil analysis. The details of the calculations for SDD and RMSD can be found in Paper I and II.
3. Results and discussion

3.1 The application of a soil spectral library for field/local-scale prediction

In Paper I, the prediction of SOC for a Danish agriculture field using a national soil spectral library was investigated. This study applied a different re-sampling strategy to calibrate soil samples that were selected from the Danish national spectral library. The method of re-sampling was based on different information on the target samples. According to the results of this study, the lowest RMSEP value was obtained by the geographically closest model and from spiking the entire spectral library with samples from the study field. Furthermore, the RMSEP value was relatively acceptable when compared to the SDD value. We found that using the geographically closest method to create a subset for calibration was the best strategy, because this method did not include any extra samples from the target site. In the last few years, several studies have used different strategies to apply national or regional soil spectral libraries for field predictions of different soil properties (Cambule et al., 2012, Goge et al., 2012, Guerrero et al., 2010, Brown, 2007, Wetterlind and Stenberg, 2010). However, the accuracy of their prediction results were not generally equal to laboratory chemical analyses. Based on all the previous research results, using the geographically closest method to create a subset and a spiking strategy can be considered as the two optimal approaches for SOC prediction at field scale. The major principle of the re-sampling strategy is to select samples from library and to explain the maximum variation of all target samples. Because chemometric modelling is an empirical method that highly depends on a similarity between training and testing samples, the re-sampling strategy is supposed to transfer the variation of target samples to the library subset. Hence, in order to use the national spectral library for further application of soil analysis at field/local scale and facilitate precision agriculture, a number of conditions have to be met:

- A spectral library which covers variations of soil type at the target site is essential;
- To apply the geographically closest method, geographical location of the target site and sampling points are required;
- As shown in our study, the model for soil samples within a 30 km radius from the target site outperformed the model with a 40 km radius, due to a change in soil parent material after 30 km. Thus, additional information such as geology and landscape maps can support the geographically closest method for the selection of samples for a subset.
In Paper I, the target site was located at the centre of a morainic landscape. Therefore all the geographically closest samples (30 km radius) represented a very similar parent material. We assumed that the geographically closest model would probably not have been the best one if the area within the 30 km radius of the target site was very heterogeneous in terms of parent material. On the other hand, it is impossible to create a subset for calibration based on the geographical information if we do not have geographical coordinate information for target soil samples. In this situation, the spiking strategy would be an alternative.

A statistical method was also applied to select the most alike spectra from the library according to target samples (Paper I). This approach was supposed to be the most efficient way of using the spectral library for the prediction of field soil properties. However, of all the models generated, this method gave the worst predictions. This was explained by the issue of spectral similarity. The selection of the most alike spectra in this study considered the general similarity between two spectra rather than matching the specific SOC bands. It means that the most alike spectra selected in this study were probably not similar in terms of SOC, at least not all SOC-specific bands. Sometimes different soils could produce similar spectra, and small spectral differences can potentially result in big differences in the predicted soil properties.

In summary, constructing a soil spectral library is a cost-effective and convenient method of facilitating DSM and soil analysis. However, for the future use for precision agriculture soil maps, the application of a soil spectral library for field prediction needs further work.

3.2 Soil sampling and laboratory uncertainty
The precision and accuracy of soil wet chemistry analysis are clearly directly related to the soil sampling, preparation, and analytical procedure of the system. We commonly consider this result as the proximity to the true value. However, the accuracy of the analytical result highly depends on the choice of field sampling strategy, reference method and valid calibration of the measurement instrument (Houba et al., 1996). Due to different determination methods of SOC and clay contents, different laboratory and even different operators can lead to variable results. In this thesis, we used dry combustion methods for SOC analysis, and hydrometer methods for measuring clay content. Table 3 shows triplicate SOC measurements for 10 samples from Paper III; two of the triplicate measurements were conducted in our own laboratory on different dates and by different operators and the other at a commercial laboratory in Germany. These results show the reproducibility of SOC laboratory analysis from two different laboratories. We can clearly find large differences
between the three replicate measurements for most of the samples. Especially sample number ten shows up to 36% uncertainty in the three measurements reported.

Table 3. Triplicate SOC measurements from different laboratories and operators on different dates

<table>
<thead>
<tr>
<th>Sample No</th>
<th>Measurement 1</th>
<th>Measurement 2</th>
<th>Measurement 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.96</td>
<td>1.06</td>
<td>1.24</td>
</tr>
<tr>
<td>2</td>
<td>0.72</td>
<td>0.90</td>
<td>0.69</td>
</tr>
<tr>
<td>3</td>
<td>1.27</td>
<td>1.42</td>
<td>1.40</td>
</tr>
<tr>
<td>4</td>
<td>0.59</td>
<td>0.81</td>
<td>0.64</td>
</tr>
<tr>
<td>5</td>
<td>1.20</td>
<td>1.01</td>
<td>1.34</td>
</tr>
<tr>
<td>6</td>
<td>0.83</td>
<td>0.82</td>
<td>0.93</td>
</tr>
<tr>
<td>7</td>
<td>1.00</td>
<td>1.00</td>
<td>0.99</td>
</tr>
<tr>
<td>8</td>
<td>2.05</td>
<td>1.91</td>
<td>2.03</td>
</tr>
<tr>
<td>9</td>
<td>1.28</td>
<td>1.28</td>
<td>1.34</td>
</tr>
<tr>
<td>10</td>
<td>0.99</td>
<td>1.35</td>
<td>1.05</td>
</tr>
</tbody>
</table>

Table 4 shows another triplicate SOC measurement from two different Danish standard soils. The daily measurements were conducted by the same instrument, operator and in a short time interval. These two standard samples were used for daily instrument calibration, and these results were used in the SDD value calculation (Paper I). Here we only show the results for three different days. These results show the repeatability of SOC laboratory analysis in our own laboratory. Based on the results from Table 3 and Table 4, we can see the random error for SOC measurements based on traditional lab wet chemistry analysis. Meanwhile, sample 1 (Table 4) was also used as a standard sample for daily instrument calibration of clay analysis. We randomly selected results from ten different days, and the variation in clay contents in these ten replicates was 10.7-13.2%. This gives a random error of approximately 20-30% for the clay contents analysis of this sample in our lab, which is substantial.
Table 4. Triplicate SOC measurements from the same laboratory and operator on different dates

<table>
<thead>
<tr>
<th>Date</th>
<th>Sample 1</th>
<th>Sample 2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>SOC %</td>
<td>SOC %</td>
</tr>
<tr>
<td>M 1</td>
<td>M 2</td>
<td>M 3</td>
</tr>
<tr>
<td>M 1</td>
<td>M 2</td>
<td>M 3</td>
</tr>
<tr>
<td>M 1</td>
<td>M 2</td>
<td>M 3</td>
</tr>
<tr>
<td>1</td>
<td>1.51</td>
<td>1.38</td>
</tr>
<tr>
<td>2</td>
<td>1.66</td>
<td>1.49</td>
</tr>
<tr>
<td>3</td>
<td>1.40</td>
<td>1.6</td>
</tr>
</tbody>
</table>

M - measurement number

All the errors showed in Table 3 and Table 4 could be caused by sampling variability, improper calibration of instrument, or unknown/unpredictable changes in the instruments or the environment, and so on. Technically, random errors are things that never can be avoided, but we can correct (mostly) by repeating the experiment or averaging results from replicates. However, due to very costly soil analyses in Europe (each soil sample for SOC and texture analysis costs around €60), we usually only measure the soil sample once, and commonly accept this result as a reliable value. Consequently, based on the results above, the uncertainty of SOC and clay contents from wet chemistry analysis are not negligible.

If we go back to the soil sampling step, the variability of the field could also lead to random error in soil analysis. A global positioning system (GPS) is commonly used to locate and guide us to sampling locations. Normally, we randomly choose a sampling point to collect soil after reaching the sampling location, but the soils we collect may not be representative enough for the soil properties of that area. To overcome this issue, as described in Paper III, we can combine many subsamples from the area surrounding the sampling location according to GPS coordinates. However, in laboratory analysis, of SOC especially, less than one gram of a sample is taken. This analytical result may not fully reflect field variation if we use it for DSM. Thus, the inaccuracy of traditional soil wet chemistry analysis can be defined as soil sampling uncertainty plus laboratory uncertainty. For spectroscopy measurement, the total amount of a soil sample is much higher than for traditional soil wet chemistry analysis. For scanning with the spectrometer used in Paper II and III we used around 100-g soil samples. Furthermore, each sample was scanned by seven scanning points, and each scanning point was an average of four repeated scans. The final spectrum was averaged by 27 replicates. Hence, we consider that spectroscopy measurement yields much higher repeatability than any traditional soil wet chemistry analysis method. Meanwhile, a significant uncertainty may be overlooked in not using the same test samples for spectroscopy and wet chemistry analyses. Therefore, we suggest that spectroscopy measurement should be done before wet
chemistry analysis and on the same sample. Once traditional wet chemistry soil analysis has been performed, the samples are not reusable.

3.3 Estimation of laboratory uncertainty
For quality assurance of analysis, it is very important to evaluate the performance of measurement. There is an international standard, ISO 5725, for accuracy (trueness and precision) of measurement methods and results. The second part of this standard, ISO 5725-2 (1994), refers to the basic method for determination of repeatability and reproducibility of standard measurement method. We commonly consider this standard as the basic method for estimating the precision of measurement methods by means of a collaborative inter-laboratory experiment for repeatability and reproducibility. To the author's knowledge, only Sørensen and Dalsgaard (2005) applied this standard in soil spectroscopy to estimate repeatability and reproducibility for clay contents analysis. However, in practice the standard is very expensive to use for all the soil analysis.

On the other hand, this standard does not report the "true" error for any given sample. There are many reasons for this, one of which is that there are often systematic errors in reference readings, and they will never show up in the duplicate measurements. Measuring duplicates (or triplicates, or ...) is, however, the best we can do to estimate the error, as well as to estimate the sample value. Over the years, chemists and statisticians have devised methods for overcoming these problems, but they are long, cumbersome and expensive. In this thesis, two approaches were employed to estimate laboratory uncertainty.

In Paper I, the SDD value was calculated from replicates of different types of Danish soils which were not the same samples as used for the calibration and validation sets. All replicate measurements were analysed at the same laboratory, which means the SDD value can only be used to describe part of the laboratory uncertainty and only indicates reproducibility under a single laboratory rather than several. Therefore, the SDD value in this study was possibly underestimated. On the other hand, we could use this result to compare the RMSEP value and approximately evaluate prediction results.

In Paper II, we applied the RMSD value to estimate laboratory uncertainty, the idea being to use only reference data to approximately estimate laboratory errors. We first applied the K-means algorithm to split samples into three groups. Since the values of all samples in each group are very similar to each other, the variations of the samples in each group are minimised. If the group variances are sufficiently small, we can consider the values in each group as replicates from the same sample and their mean to represent the true value. In comparison, our result for clay contents is very similar to the inter-laboratory reproducibility standard deviation of clay contents in Danish soil (clay content: 2-26%) found by Sørensen and Dalsgaard (2005).
Based on the discussion above, to evaluate laboratory performance for each soil analysis is not practical. For practical purposes, we highly recommend for soil analysis the establishment of national or regional laboratory uncertainty references where different soil properties are represented by different ranges for all types of soils within specific regions. Then the best of choice is to use the “rule of thumb” to estimate what we should expect.

3.4 True accuracy of Vis-NIR and MIR

The purpose of using spectroscopy and chemometrics for soil analysis is to establish the link between a soil property and spectral variables, and to use a model to predict unknown samples. The obtained prediction error such as the RMSEP value for unknown samples is estimated from the calculated differences between the prediction values and “known” reference values of the validation set. However, this RMSEP value cannot represent the true prediction error, because not all of the reference values of validation samples are accurate enough. It is important to note that the uncertainty of a laboratory does not contribute to the true reference values, but with the model and spectral variables we are interested in predicting the true values rather than measured one. Thus, the RMSEP value is an apparent prediction error, which is systematically larger than the true prediction error. For instance, in Paper I, the best prediction results for SOC gave an RMSEP value of 0.19%. This value indicates only the apparent prediction error, but the true prediction error is probably much smaller than 0.19%, possibly even smaller than the SDD value. This confounding effect of the inaccuracy in the reference values of the validation set has not been pointed out in soil spectroscopy. Consequently, most researchers explain the apparent prediction error as an inadequacy in the estimated calibration model and not as the contribution from laboratory uncertainty, which is incorrect.

On the other hand, there is also a recurrent misconception that spectroscopy can never be more accurate than traditional wet chemistry soil analysis because there is an error component accumulating from each stage, as discussed above. In common sense, these errors all contribute to the model error and to the prediction error. Interestingly, it is very difficult to prove this assumption in practice, since laboratory errors can only be approximately calculated from replicate values under repeatability and reproducibility conditions, which it is very costly. However, if we go back to the modelling stage, there are measurement errors associated with both spectral variables (X) and reference values (Y). But this does not mean that model error will be the sum of the errors in X and Y, because most of the random errors associated with measurement will be cancelled out during the regression process for generating a model, if a sufficient number of samples is used (Difoggio, 1995). This means that the prediction error may be lower than or equal to the laboratory error in general. So although Esbensen et al. (2002) maintained that we can never expect the prediction error to be lower than laboratory error, this might not be fully correct.
3.5 Practical implications of prediction accuracy

Fertiliser dosing depends on the type and growth of the vegetation in the field. In order to apply a fertiliser in the right place and at the correct and most economic rate, knowledge of the precise spatial distribution of different soil properties is required. No previous research has generally discussed the acceptability of Vis-NIR and MIR prediction results for practical fertilisation. In this thesis, we discuss the possibility of using the spectroscopy technique to estimate SOC and clay contents for liming and nitrogen (N) fertilisation purposes. The discussion is based on the Danish agricultural soils only.

3.5.1 Determination of lime requirement based on spectroscopy prediction

In general, soils have a pH value in the range of 5-8, according to the Danish soil library (Table 3), and around 25% of the Danish farmland is affected by soil acidity (pH below 5.5). In agriculture, depending on the crop or vegetation type, the ideal soil pH value would be 6.5-7. To achieve this, liming to neutralise soil pH is required. Thus, determination of soil liming requirement is very important. The soil lime buffer capacity (LBC) and cation exchange capacity (CEC) are commonly used to reflect soil lime requirement (Sparks, 2003). There is a general impression among farmers that the determination of lime requirement depends only on soil pH. However, soil organic matter also has a large effect on the variation in LBC for different soil types (Haynes and Naidu, 1998, Keeney and Corey, 1963, Ross et al., 1964, Aitken et al., 1990, Husni et al., 1995, Pagani and Mallarino, 2012).

Keeney and Corey (1963) proposed a calculation for determining liming requirement which accounts for the importance of soil organic matter (SOM). With a desired pH value of 6.5, the equation is as follows:

\[
\text{Lime requirement (t ha}^{-1}\text{)} = 1.6(6.5-p\text{Ha}) \times (\text{SOM}\%)
\]  
(Eq. 5)

where pHa is the actual pH value.

According to the Danish soil library (Table 5) the SOM content ranges from 0-94.8%. However, out of 2851 samples, 1323 samples had SOM values below 0.5%, for 1719 samples SOM was less than 1%, for 2114 samples less than 2%, for 2449 samples less than 3%, and for 2689 samples less than 5%. If we assume that all the soils represent the same acidity condition (pH=5.5), and the target pH value is 6.5, based on Eq. 5, the results for the liming requirement can be very different. Soils with a high organic matter content will require larger
amounts of limes to neutralise the acidity than soils lower in organic matter, given that each soil has the same pH to start with.

Table 5. General statistics on soil characteristics of the Danish soil library

<table>
<thead>
<tr>
<th>Soil properties</th>
<th>Min</th>
<th>Max</th>
<th>Mean</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>SOC %</td>
<td>0</td>
<td>55.2</td>
<td>1.1</td>
<td>3.1</td>
</tr>
<tr>
<td>SOM %</td>
<td>0</td>
<td>94.8</td>
<td>1.9</td>
<td>5.3</td>
</tr>
<tr>
<td>pH(H2O)</td>
<td>2.4</td>
<td>9.1</td>
<td>6.4</td>
<td>1.1</td>
</tr>
<tr>
<td>Clay %</td>
<td>0</td>
<td>72.3</td>
<td>10.2</td>
<td>8.6</td>
</tr>
<tr>
<td>Silt %</td>
<td>0</td>
<td>88</td>
<td>19.3</td>
<td>13.5</td>
</tr>
<tr>
<td>Sand %</td>
<td>0</td>
<td>97.6</td>
<td>65.2</td>
<td>23.7</td>
</tr>
</tbody>
</table>

SD, standard deviation;  n, number of samples

Since we can never avoid prediction error when using the spectroscopy method for soil analysis, and that includes the SOM prediction results, it is important to set up a maximum range in error that is acceptable for liming purposes based on the different SOM conditions. In practice, tonne per hectare (t ha\(^{-1}\)) (integer value) is commonly used as a unit for lime fertilisation which indicates the total amount of lime input to soil. We also assume that the soils are under the same acidity condition (pH=5.5) and for the present discussion an SOM range of between 0 and 5%, since it is not common to find high variations of SOM within fields on Danish farmland, and most of the highly organic soils are located on wetland areas. Based on Eq. 5, we calculated the lime requirement for soil with different SOM contents (Table 6).
Table 6. Lime requirement for soil with different SOM contents based on the same actual pH value and soil type

<table>
<thead>
<tr>
<th>SOM (%)</th>
<th>Lime requirement (t ha⁻¹)</th>
<th>Actual pH</th>
<th>Target pH</th>
</tr>
</thead>
<tbody>
<tr>
<td>0-0.6</td>
<td>1</td>
<td>5.5</td>
<td>6.5</td>
</tr>
<tr>
<td>0.7-1.8</td>
<td>1-2</td>
<td>5.5</td>
<td>6.5</td>
</tr>
<tr>
<td>1.9-2.4</td>
<td>2-3</td>
<td>5.5</td>
<td>6.5</td>
</tr>
<tr>
<td>2.5-3.1</td>
<td>3-4</td>
<td>5.5</td>
<td>6.5</td>
</tr>
<tr>
<td>3.2-3.7</td>
<td>4-5</td>
<td>5.5</td>
<td>6.5</td>
</tr>
<tr>
<td>3.8-4.3</td>
<td>5-6</td>
<td>5.5</td>
<td>6.5</td>
</tr>
<tr>
<td>4.4-5.0</td>
<td>6-7</td>
<td>5.5</td>
<td>6.5</td>
</tr>
</tbody>
</table>

The difference in SOM contents between each liming level was around 0.6%. This means that at this level the prediction error of SOM contents with the spectroscopy method is acceptable. If we convert 0.6% of SOM to SOC, the result will be approximately 0.35% of SOC. According to prediction results from Paper I and Paper II, the RMSEP and RMSECV values gave 0.19% and 0.35% of SOC, respectively. This means that the prediction results from these two studies were acceptable for liming purposes.

Some studies also indicate that clay contents could affect determination of liming requirement (Aitken et al., 1990, Husni et al., 1995, Pagani and Mallarino, 2012, Ross et al., 1964). However, Kummar and Devi (2012) did not observe such influence in their study. This was because the impact of clay was masked by organic matter when variability in clay content of the soils was not as pronounced as that in organic matter. According to the general statistics of the Danish soil library (Table 3), there is generally little variation in clay content in the Danish agricultural topsoil. Therefore, we did not include clay as a parameter for determination of liming requirement, since low variability in clay contents is not expected to affect CEC and LBC.

### 3.5.2 Practical nitrogen fertilisation based on spectroscopy prediction

Nitrogen is the largest nutrient element needed for crop growth and directly affects productions. Fertiliser recommendations for N are also based on yield potential and clay
content for the soil, not only soil nitrogen contents (Glass, 2003). Clay soils are generally more “fertile” and leach less than sandy soils, because clay soils usually have a higher CEC. The main soil types in Denmark are coarse sandy to loamy soils with soils higher in clay contents in the eastern part of the country (Madsen and Jensen, 1992). The total amounts of N input and frequency of N fertilisation can therefore be highly influenced by the clay percentage (especially when irrigated). Moreover, Denmark also has legal limits on the permitted levels of N fertilisers, depending on the fined texture classes and the grown crop. Thus, the accuracy of prediction results for clay contents from the spectroscopy method is very important in practical N fertilisation.

In this thesis, we discuss only N fertilisation based on the arable land where barley and wheat are planted, since they are the major crops in Denmark. Danish soils have been classified into twelve classes (JB 1-12) according to soil texture (Table 7) (Aslyng, 1968, Aslyng, 1976). Based on the Danish soil library and a high-resolution (30 m) soil texture map in Denmark generated by Adhikari et al. (2013), over 90% of Danish topsoils (0-30 cm) belong to JB 1-8, according to the JB system (Table 7). This means that the range of clay contents in most Danish soils is from 0 to 45%. Based on the JB system, soil type can be roughly defined by clay contents in 5% steps. For example, sandy clayey soil is defined as JB 5 or 6 which have 10-15% clay. Clayey soil is defined as JB 7 which has 15-25% clay. These two soil types also contain very similar percentages of silt and total sand. Since N fertilisation in Denmark is partly based on the soil type (JB system), precise estimation of clay contents is very important in practical N fertilisation.
Table 7. Definition of soil types from the Danish soil classification

<table>
<thead>
<tr>
<th>Soil type</th>
<th>JB-no</th>
<th>Clay&lt;2μm</th>
<th>Silt 2-20μm</th>
<th>Fine sand 20-200μm</th>
<th>Total sand 20-2000μm</th>
<th>Organic matter</th>
<th>Lime CaCO3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Coarse sandy soil</td>
<td>1</td>
<td>0-5</td>
<td>0-20</td>
<td>0-50</td>
<td>75-100</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fine sandy soil</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td>50-100</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Clayey sandy soil</td>
<td>3</td>
<td>5-10</td>
<td>0-25</td>
<td>0-40</td>
<td>65-95</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td>40-5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sandy clayey soil</td>
<td>5</td>
<td>10-15</td>
<td>0-30</td>
<td>0-40</td>
<td>55-90</td>
<td>&lt;=10</td>
<td>&lt;=10</td>
</tr>
<tr>
<td></td>
<td>6</td>
<td></td>
<td></td>
<td></td>
<td>40-95</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Clayey soil</td>
<td>7</td>
<td>15-25</td>
<td>0-35</td>
<td></td>
<td>40-85</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Heavy clayey soil or silty soil</td>
<td>8</td>
<td>25-45</td>
<td>0-45</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>9</td>
<td>45-100</td>
<td>0-50</td>
<td></td>
<td>0-55</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>0-50</td>
<td>20-100</td>
<td></td>
<td>0-80</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Organic soil</td>
<td>11</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>&gt;10</td>
<td>0-90</td>
</tr>
<tr>
<td>Atypic soil</td>
<td>12</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>&lt;=10</td>
<td>&gt;10</td>
</tr>
</tbody>
</table>

Since ground water is the major water resource in Denmark, and N leaching is one of the biggest potential risks to ground water quality, the Danish Ministry of Agriculture proposed fertilisation standards according to soil type and crop in order to limit the amount of N fertiliser that could be input to soil on an annual basis. Table 8 shows N fertilisation standards for barley and wheat based on different soil types. For instance, for areas where the soils are defined as JB 7, indicating a clay soil, farmers are allowed to apply a maximum of 222 kg N/ha every year for bread wheat. This is in effect the optimum amount of N fertiliser for the production of bread wheat. The extra input will not be taken up by crops and remains in the soil, resulting in N leaching losses (Webster et al., 1986, Gregorich et al., 2011).
Table 8. Nitrogen fertilisation standards for Danish agricultural crops

<table>
<thead>
<tr>
<th>Crop</th>
<th>Unirrigated coarse sand (JB 1&amp;3)</th>
<th>Unirrigated fine sandy soil (JB 2&amp;4)</th>
<th>Irrigated sandy soil (JB 1-4)</th>
<th>Sandy clayey soil (JB 5&amp;6)</th>
<th>Clay soil (JB 7-9)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spring barley</td>
<td>112</td>
<td>107</td>
<td>126</td>
<td>114</td>
<td>128</td>
</tr>
<tr>
<td>Spring wheat</td>
<td>109</td>
<td>106</td>
<td>125</td>
<td>112</td>
<td>127</td>
</tr>
<tr>
<td>Winter barley</td>
<td>150</td>
<td>141</td>
<td>158</td>
<td>152</td>
<td>166</td>
</tr>
<tr>
<td>Winter wheat</td>
<td>146</td>
<td>149</td>
<td>166</td>
<td>161</td>
<td>174</td>
</tr>
<tr>
<td>Bread wheat</td>
<td>182</td>
<td>188</td>
<td>207</td>
<td>206</td>
<td>222</td>
</tr>
</tbody>
</table>

According to Table 7 and Table 8, the difference in clay contents between JB 3 and JB 7 is 5%, but the difference in the total amount of N fertilisation input for these two soils where bread wheat is planted is 40 kg N/ha. If we had wrongly estimated the clay content for a coarse sandy soil to be more than 5% and wrongly classified it as a clay soil, the farmer would apply an additional 40 kg N fertiliser every year based on the Danish N fertilisation standards. This extra N can highly affect groundwater quality in the long term, due to N leaching losses. In Paper II, a large variation in clay content within one field was reported where the range in clay content was between 5-24%. According to the JB system, the soil types in this study field ranged from JB1 to JB7. In order to optimise N fertiliser application and avoid under/over-prediction of clay content using the spectroscopy method, we recommend calibrating separate models according to the JB system and different clay ranges. Since the clay contents for unknown or target samples are unknown, it is very difficult to choose a proper model to predict target samples. In this situation, PCA can be an alternative way to distinguish variations of target samples. For example, if the variation of target samples is completely covered by the variation in the calibration samples in the PCA score plot, the calibration
model is able to represent all variations of the new samples based on spectral information, and can be successfully applied to predict target samples.
4. Conclusions
This PhD work applied laboratory Vis-NIR, MIR and RS spectroscopy combined with different chemometric tools to SOC and clay analysis as well as mapping. The work demonstrated the suitability of spectroscopy method for quantitative soil analysis, and evaluated the techniques of Vis-NIR and MIR spectroscopy to replace or partly replace standard wet chemistry soil analysis. The spectroscopy techniques can be used for SOC and clay contents modelling and mapping, and support decision systems for precision agriculture. The conclusions and implications of the results presented in the thesis are presented below:

- Creating a subset of a soil spectral library for field-scale SOC prediction is an efficient way to reduce a large data set and improve prediction accuracy.
- In terms of subset of a soil spectral library creating, geographical location, soil parent material and landscape highly improve the accuracy of the reduced library model.
- The strategy of spiking the library with local field samples is an alternative for transferring field-specific variation to the calibration model.
- Variable selection can be used to select important spectral features for SOC and clay contents.
- The optimal number of MIR measurements of SOC and clay content is 3-4 replicates, depending on spectrometer type.
- In order to estimate the uncertainty of wet chemistry soil analyses, the root mean square deviation of reference data in a validation set can be considered as a low-cost method for laboratory uncertainty estimation.
- The combination of laboratory spectra and RS spectra is able to capture the spatial pattern of SOC at regional scale.
- The uncertainty associated with traditional wet chemistry soil analysis highly affects the evaluation of accuracy using Vis-NIR and MIR spectroscopy methods for predicting soil properties.
- The prediction error of the spectroscopy method for SOC and clay contents possibly lower than or equal to the laboratory error in general.
- According to the liming and N fertilisation guidelines for Danish agriculture, we could accept prediction errors of the spectroscopic models for SOC and clay content of less than 0.41% and 5%, respectively. Separate calibration of dataset for clay content range in the JB system for N fertilisation is recommended.
5. Perspectives

This work contributed several approaches to the quantitative analysis and mapping of SOC and clay contents using Vis-NIR, MIR and RS spectroscopy techniques. A number of possible directions for future research are identified below:

- Exploration of the spectral feature selection approach to select the most alike spectra according to specific soil property and to create subset spectra from the spectral library for calibration and prediction at field scale.
- Investigation of the effect of sampling density of the soil library in terms of spectral library development and further applications.
- Introduction of Bayesian methods in chemometrics for soil spectral data to enhance the ability to use all kinds of available information to obtain more accurate and useful models.
- The use of different variable selection methods to reduce as far as possible non-significant variables to simplify the model and facilitate spectral interpretation.
- The use of the Vis-NIR and MIR combination to obtain more specific/detailed soil spectra for soil quality research.
- The use of high-resolution RS images (e.g. 2 m) from different satellites or sensors for DSM to capture more details of soil surface spatial pattern at field or larger scale.
- The use of an unmanned aerial vehicle (UAV) to carry a hyper-spectral camera and collect field information on bare soils for soil properties spatial modelling.
- Investigation of different Danish soils for SOC and clay uncertainty analysis including both repeatability and reproducibility. The result can be used for evaluation of prediction accuracy of the spectroscopy method.
6. References


REEVES, J. B. 2010. Near- versus mid-infrared diffuse reflectance spectroscopy for soil analysis emphasizing carbon and laboratory versus on-site analysis: Where are we and what needs to be done? Geoderma, 158, 3-14.


7. Supporting papers
Paper 1

*Predicting Soil Organic Carbon at Field Scale Using a National Soil Spectral Library*

Peng Yi, Maria Knadel, René Gislum, Fan Deng, Trine Nørgaard, Lis Wollesen de Jonge, Per Moldrup, and Mogens Humlekrog Greve


doi: [http://dx.doi.org/10.1255/jnirs.1053](http://dx.doi.org/10.1255/jnirs.1053)

Reprinted with permission from Journal of Near Infrared Spectroscopy
Soil organic carbon has received much attention recently, due to the increasing concentration of CO$_2$ in the atmosphere, intensifying global warming. Over recent decades, many studies in soil science have made use of visible (vis) and near infrared (NIR) reflectance spectroscopy to predict soil properties and texture. This technique is based on molecular overtones and combination vibrations, because specific molecules absorb NIR light energy at specific wavelengths. The absorptions are correlated with the concentration of the molecules according to the Beer–Lambert Law. The type of radiation referred to as vis-NIR spectroscopy is commonly defined as the wavelengths between 400 nm and 2500 nm. This region contains useful information on organic and inorganic materials in the soil, specifically the C–H, C=O, O–H and N–H vibrations.

Introduction

Soil organic carbon has received much attention recently, due to the increasing concentration of CO$_2$ in the atmosphere, intensifying global warming. Over recent decades, many studies in soil science have made use of visible (vis) and near infrared (NIR) reflectance spectroscopy to predict soil properties and texture. This technique is based on molecular overtones and combination vibrations, because specific molecules absorb NIR light energy at specific wavelengths. The absorptions are correlated with the concentration of the molecules according to the Beer–Lambert Law. The type of radiation referred to as vis-NIR spectroscopy is commonly defined as the wavelengths between 400 nm and 2500 nm. This region contains useful information on organic and inorganic materials in the soil, specifically the C–H, C=O, O–H and N–H vibrations.
bonds, which are characteristic of soil organic matter (SOM) or soil mineral fractions. In the vis region (400–780 nm), SOM can also have broad absorption bands that are dominated by darkness of organic matter and chromophores.7,8

In order to quantitatively analyse soil characteristics using vis-NIR, many multivariate calibration methods are needed to extract the desired information from the soil spectrum. The Beer–Lambert Law has been used to prove a linear relationship between absorption values and component concentrations in a sample.6 Partial least squares regression (PLSR)9 is the most common tool used to calibrate spectra and soil chemical components. The advantage of the PLSR algorithm is that it maximises the covariance between variable X and reference Y and thus the resulting spectra are directly related to the soil characteristics.5,9

To take full advantage of the vis-NIR spectroscopy technique for estimating SOC, many researchers recommend the establishment of a soil spectral library as one of the best approaches to measure soil properties.10,11 For example, Knaedel et al.11 created a Danish soil spectral library which comprised of 2851 samples, covering all of Denmark. Shepherd and Walsh12 developed a soil spectral library with more than 1000 topsoil samples from eastern and southern Africa and described a conceptual framework for using the spectral library to predict soil properties.

Another soil spectral library ("global library") with 4184 soil samples, primarily from the United States, was constructed by Brown et al.10 Using this global library, they found strong relationships between vis-NIR reflectance and SOM, clay content, inorganic carbon, dithionate citrate-extractable Fe and cation exchange capacity. They also suggested that the soil spectral library should be augmented (spiked) with local samples to support the assessment of soil variability, which means adding some local samples to the library of soil spectra and recalibrating (the difference between augmenting and spiking concerns whether or not the spectral library covers the local sample area). To test this hypothesis, Brown12 augmented the initial global library of 4184 samples with 418 local samples from Uganda to recalibrate for SOC and clay content and found that this yielded better predictions than using either the global or local calibration alone. Sankey et al.13 quantified the value of augmenting a large global library with local samples to predict clay, SOC and inorganic carbon, and showed that the use of the soil spectral library alone gave less accurate predictions than a library combined with local samples. In Sweden, Wetterlind and Stenberg14 selected 50 samples from a Swedish soil spectral library, based on mean and standard deviation in multivariate analysis, which gave the lowest value (0.37%) for the root mean square error of validation (RMSEV). Wetterlind and Stenberg14 selected 50 samples from a Swedish soil spectral library, based on mean and standard deviation in multivariate analysis, which gave the lowest value (0.37%) for the root mean square error of validation (RMSEV). Wetterlind and Stenberg14 selected 50 samples from a Swedish soil spectral library, based on mean and standard deviation in multivariate analysis, which gave the lowest value (0.37%) for the root mean square error of validation (RMSEV).

To generate a robust model for future use, a library should cover as much of the soil variation as possible. However, a high number of samples covering a large variability in soil type can lead to lower accuracy of prediction. In addition, the relationship between spectral data and reference data is not always linear and the source of non-linearity may vary widely and may be difficult to identify.16–18 The risk of non-linearity during analysis can be reduced by taking a specific set of calibration samples to predict unknown samples. For example, the locally weighted regression (LWR) method can be used to select a sub-set of samples as a calibration set similar to the unknown samples to avoid non-linearity.18,19 In France, Goge et al.17 created groups of samples from a French soil spectral library (2126 samples) for specific local calibration and selected up to 500 neighbouring samples based on eight indices combining distinct distance [Mahalanobis distance or correlation coefficient] and two methods of data reduction [principal component analysis (PCA) or fast Fourier transform]. They found that the best model for predicting SOC was created by selecting a sub-set with 400 samples, with selection based on the correlation coefficient and fast Fourier transform, which gave the lowest value (0.37%) for the root mean square error of validation (RMSEV). Wetterlind and Stenberg14 selected 50 samples from a Swedish soil spectral library, based on mean and standard deviation in multivariate analysis, which gave the lowest value (0.37%) for the root mean square error of validation (RMSEV). Wetterlind and Stenberg14 selected 50 samples from a Swedish soil spectral library, based on mean and standard deviation in multivariate analysis, which gave the lowest value (0.37%) for the root mean square error of validation (RMSEV). Wetterlind and Stenberg14 selected 50 samples from a Swedish soil spectral library, based on mean and standard deviation in multivariate analysis, which gave the lowest value (0.37%) for the root mean square error of validation (RMSEV). Wetterlind and Stenberg14 selected 50 samples from a Swedish soil spectral library, based on mean and standard deviation in multivariate analysis, which gave the lowest value (0.37%) for the root mean square error of validation (RMSEV).

Despite the advantage of taking sub-set samples from a spectral library, very few studies have focused on improving the accuracy of SOC prediction at local field scale by selecting a particularly relevant sub-set from the national spectral library. The aim of this study was, therefore, to compare models based on samples from the Danish National Spectral Library, a spiked national library and three reduced national spectral libraries. The hypothesis was that the accuracy of SOC predictions at field scale is improved by reducing the soil spectral library according to the soil type of local samples.

Materials and methods
Field site and sampling
A total of 65 topsoil (0–20 cm) samples were collected from a 1.69 ha agricultural field in Silstrup, north-west Jutland, Denmark (Figure 1), in October 2010 as part of a soil leaching study.20,21 The study site is situated on glacial till in a moraine landscape, which is the dominant geological formation in Denmark, covering around 43% of land area.22,23 The geological setting consists of a homogeneous loamy basal till rich in chalk and chert, with thin silt and sand beds.24 The site and sampling procedure are described in detail by Norgaard et al.25 Soil samples were air-dried, sieved to 2 mm and analysed...
for SOC using a LECO analyser coupled with an infrared CO$_2$ detector (Thermo Fisher Scientific Inc., CITY, STATE, USA). Soil texture (clay, silt, sand) was determined using a combined sieve/hydrometer method. Clay was defined as particles <2 µm, silt 2–50 µm, fine sand 50–200 µm and coarse sand 200–2000 µm. Bulk density was determined from the weight of the soil columns after oven-drying at 105°C.

**Vis-NIR measurements**

Reflectance spectra were obtained with a LabSpec 5100 spectrophotometer (ASD Inc., Boulder, Colorado, USA) using a laboratory protocol described in Knadel et al. The spectral range covers the visible and near infrared bands between 350 nm and 2500 nm. The spectra were collected using a Muglight sample holder (ASD Inc.). A white reference was taken every five measurements and every sample was measured twice and the average spectrum used. To eliminate noise at the edges of each spectrum, spectra at 350–433 nm were removed.

**National library and spiked library**

The soil spectra from the national library originated from the Danish soil profile database, which comprises 700 agricultural soil profiles gathered from a nationwide 7 km soil monitoring grid (Figure 2). The Danish Soil Spectral Library consists of 2851 spectra covering the range 354–2500 nm. In order to reduce the data size, the...
wavelength was decreased by averaging for every 10 nm consecutive wavelengths. In the soil library, SOC ranges from 0% to 55.2%. There are 2008 samples where the SOC content is less than 1% and only 163 samples with SOC content higher than 3%. In this study, 2688 samples were used, with SOC ranging between 0% and 3%, which covered the SOC range of the Silstrup field (1.7–2.2%, Table 1). The library data set was divided into two groups, one for calibration and one for validation. The data set was sorted in ascending order of SOC value and separated into two groups (1344 samples per group). All spectra data sets used the same wavelength range as the Silstrup spectra (averaged for every 10 nm from 434 nm to 2500 nm). For the spiked library, we randomly selected 30 samples from the Silstrup field and combined them with the library to recalibrate and predict the remaining 35 samples.

Reduced library
In order to improve the model prediction for the Silstrup field based on the Danish spectral library, we reduced the library using three methods to create calibration sub-sets.

Geographically closest method
We assumed that the study field represented a homogeneous soil type in a similar physiographical environment. Hence, in the library, the geographically closest samples around the study site should be very similar to the Silstrup field. Soil samples within a 20 km, 30 km and 40 km radius from the Silstrup site were selected from the spectral library (Figure 2), producing 51, 85 and 138 samples, respectively. Three separate data sets were used to build the models. The best geographically closest model selected was that with the lowest root mean square error of cross-validation (RMSECV) and the highest correlation coefficient ($R^2$).

Same landscape and parent material
The Silstrup study site is situated in a moraine landscape similar to that found in eastern Denmark. As soils from the same landscape and parent material are presumably similar, we selected samples from eastern Denmark with similar soil types to the Silstrup field (Figure 2). The moraine landscape data set included 777 samples where the SOC content ranged from 0% to 25.6%. In order to reduce the data set, we used the same method used for library separation to select 96 samples with SOC content below 3%.

The most alike spectra
The Mahalanobis distance between the Silstrup samples and the samples from the soil spectral library was calculated. All data were transformed by the 1st Savitzky-Golay derivative. Data from the soil library were reduced using PCA and the distance from each of the Silstrup samples to each of the samples in the soil library was calculated based on the first two components. Using these results, it was possible to identify the 100 samples from the library with spectra most alike the Silstrup spectra.

Chemometric methods
The calculations were performed with the Unscrambler X 10.1 software (Camo ASA, Oslo, Norway). Before analysis, all spectra were transformed from reflectance ($R$) to absorbance ($A$) using the function $A = \log_{10}(1/R)$. In order to enhance the relevant peaks of the spectra and reduce the influence of particle size and thus improve the model, four different spectra pre-treatment methods were used: 1st Savitzky-Golay derivative (1d), 2nd Savitzky-Golay derivative (2d), multiplicative signal correction (MSC) and standard normal variate (SNV). The PCA was applied to explore all samples from different data sets and to identify outliers. Samples with Mahalanobis distance $H > 3$ were considered spectral outliers or extreme values and removed from further analysis. Partial least squares regression calibration models were developed with leave-one out cross-validation and applied on the reduced library sub-sets. The optimal numbers of factors were determined by minimising the prediction error of validation. The models were subsequently applied and tested on 35 Silstrup samples, due to 30 samples being combined into the library for spiking purposes. The models were evaluated by $R^2$, root mean square error of calibration (RMSEC), RMSECV, bias, RMSEP and the ratio of performance to deviation (RPD), which was calculated as SD divided by RMSEP.

Standard deviation of differences (SDD) was used to compare prediction error and measurement error. RMSEP $\leq 2 \times$ SDD is usually accepted for many applications.
of differences was calculated on 120 soil samples covering different types of Danish soils as:

\[
SDD = \sqrt{\frac{\sum |d_i - d_{m}|^2}{n-1}}
\]

where \(d_i\) is the difference in \(Y\) between two replicates of sample \(i\), \(d_m\) is the mean value of all replicate differences \(\left| d_{m} = \frac{\sum d_i}{n}\right|\); and \(n\) is number of samples.

Results

Exploratory data analysis

The content of SOC in the Silstrup samples ranged from 1.7% to 2.2%, the mean value of clay content is 15.9% (Table 1), which represents a homogeneous data set with small variations in SOC. Descriptive statistics for SOC analysed in the laboratory for the different calibration sets are shown in Table 2. The SOC content of all calibration samples was lower than 3%. For clay content, the greatest variations were found in library calibration samples. Here we separated the library into two groups, one for calibration and one for test set validation. The SOC standard deviation and the mean of the validation data set (data not shown) were the same as for the calibration data set, with only the maximum value differing and then only by 0.06%. There was no clear difference in SOC content range and SD in the calibration data sets and the SOC content in all data sets ranged from almost 0 to 3. The only differences were in the mean of the most alike spectra data sets, which was twice as high as for other data sets, and the number of samples from the library and spiked library, which was much higher than that from the reduced library (Table 2).

One and three extreme samples were detected for the 30 km and 40 km distance data sets (geographically closest), respectively, representing samples with high SOC contents (data not shown). These samples were removed before data analysis.

The first two principal components (PCs) for absorbance spectra explained 96% of the total variance. The score plot of
Table 3. Calibration and internal validation statistics for SOC using PLSR models and VIS-NIR spectra (434-2500 nm) applying 1d.

<table>
<thead>
<tr>
<th>Data set</th>
<th>Transformation</th>
<th>Factor</th>
<th>( R^2 )</th>
<th>( RMSEC^b )</th>
<th>( R^2 )</th>
<th>( RMSECV^c )</th>
<th>RPD^d</th>
</tr>
</thead>
<tbody>
<tr>
<td>Geographically closest (30 km)</td>
<td>1d^e</td>
<td>8</td>
<td>0.91</td>
<td>0.21</td>
<td>0.86</td>
<td>0.27</td>
<td>2.6</td>
</tr>
<tr>
<td>Same parent material</td>
<td>1d</td>
<td>9</td>
<td>0.91</td>
<td>0.18</td>
<td>0.87</td>
<td>0.24</td>
<td>2.5</td>
</tr>
<tr>
<td>Most alike spectra</td>
<td>1d</td>
<td>9</td>
<td>0.79</td>
<td>0.30</td>
<td>0.64</td>
<td>0.40</td>
<td>1.8</td>
</tr>
<tr>
<td>Library</td>
<td>1d</td>
<td>8</td>
<td>0.75</td>
<td>0.34</td>
<td>0.74</td>
<td>0.34</td>
<td>2.0</td>
</tr>
<tr>
<td>Spiked library</td>
<td>1d</td>
<td>8</td>
<td>0.76</td>
<td>0.34</td>
<td>0.76</td>
<td>0.34</td>
<td>2.0</td>
</tr>
</tbody>
</table>

*aCorrelation coefficient of determination*
*\(^b\)Root mean square error of calibration*
*\(^c\)Root mean square error of cross-validation*
*\(^d\)Ratio of performance to deviation*
*\(^e\)1st derivative (nine points)*
*\(^f\)Root mean square error of prediction (library and library+30 model used test data set for validation instead of cross-validation)*

These first two components (Figure 3) were grouped into five classes representing Silstrup, spectral library, geographically closest, same landscape and parent material and the most alike spectra samples. There was no distinct grouping according to the Silstrup samples in the different calibration samples, but some of the samples from the reduced library data sets were close to the Silstrup group and a few samples from the most alike spectra data set overlapped Silstrup samples.

**Calibrations and predictions**

The internal validation results for the best calibration models are shown in Table 3. The calibrations were obtained using 1d with a 2nd order polynomial and smoothing over nine points and up to nine PLSR PCs. The PLS regression coefficients are shown in Figure 4. All regression coefficients had very similar absorbance peaks, for example, 570–700 nm, 1720–1760 nm, 1950 nm, 2030 nm, 2044 nm, 2134 nm, 2200 nm, 2309 nm and 2344 nm (Figure 4), which could be assigned to OM. Other similar peaks at 1384 nm, 1454 nm and 1874 nm are related to water signatures.

Of the three models based on the geographically closest method, that for 30 km distance from the study site (84 samples) yielded the best results in terms of \( R^2 \) and \( RMSECV \) values (data from 20 km and 40 km distance not shown).

The prediction results are shown in Figure 5. In general, all models except those including the most alike spectra and the same parent material showed reasonable predictive abilities. Comparing models from the reduced library, the geographically closest model gave the best results, with the highest \( R^2 \) (Table 3) and lowest \( RMSECV \) values for prediction (Figure 5). The calibration model for the same parent material gave a similar \( R^2 \) value but a lower \( RMSEP \) value than the geographically closest model (Table 3). However, the same parent material model used one more PC and the \( RMSEP \) was also higher than for the geographically closest model. Furthermore, the geographically closest model had the highest RPD (3.7) for the prediction test (Figure 5). The models for the most alike spectra and the same parent material had similar \( RMSEP \) values, but the RPD for the most alike spectra model was higher than that for the same parent material. This is because they had different standard deviations, but the prediction using the most alike spectra model had a low \( R^2 \) value of 0.1.

Calibration of the library and spiking library gave almost the same results in terms of \( R^2 \) and \( RMSEP \) (Table 3). The
calibration results for the spiked library model were not as good as for the geographically closest model. The SOC predictions from the spiked library model gave the same low RMSEP value as the geographically closest model (0.19%; lowest value observed) and the highest \( R^2 \) value of all predictions, but it was only slightly higher than for the geographically closest model (Figure 5).

The SDD in four different soils calculated using replicate measurements 30 times for each soil was in the range 0.04–0.17%.

### Discussions

This study demonstrated that using both a soil spectral library and selecting sub-sets from the library in vis-NIR predictions can provide a rapid and inexpensive method for analysing SOC.

The geographically closest model, which used the fewest number of samples (84), gave the lowest RMSEP and highest RPD value. The data used in the geographically closest data set did not include any samples from the target site, but still yielded the same RMSEP and RPD values as the model based on the spiked library. However, the geographically closet model
is preferable because it can be selected directly from the spectral library, making it cheaper and more efficient, whereas the spiked library data set still needs to combine several samples from the target site and the reference data.

The standard error of prediction (SEP) and RMSEP are the most commonly used parameters in NIR spectroscopy to describe the predictive ability of a model. Some studies present the relationship between RMSEP, SEP and bias as $\text{RMSEP}^2 = \text{SEP}^2 + \text{bias}^2$, i.e. if the bias is zero, $\text{RMSEP} = \text{SEP}$. In the present study, the bias with cross-validation was close to zero and, therefore, we only compared the models based on RMSEP and $\text{RMSECV}$. For all models of vis-NIR spectra in this study, RMSEP gave SOC values of 0.19–0.24%, which is acceptable for some, but not all, soils. A decision on whether a particular model is acceptable should then be made separately for each case. It is important to note that $\text{SDD}$ can only be used to describe the reproducibility of the measured SOC. Another error is the stability over consecutive measurements, described as repeatability of SOC, which was in the range 0.04–0.16 using the same data as for calculation of SDD.

**Library and reduced library**

Our spiked library model gave more accurate predictions than the unspiked library model, a finding consistent with previous SOC studies. Of the three different distances (20 km, 30 km and 40 km) used to construct the geographically closest data sets, the model for 30 km gave the highest $R^2$ values for calibration and validation and the lowest $\text{RMSECV}$. The calibration results for the 30 km data set were slightly better than those of the 20 km data set and much better than those of the 40 km data set. This was because, at around 40 km from the Silstrup site, the soil parent material changes from moraine with carbonate to moraine without carbonate. It is important to note that the present study site was located at the centre of a moraine landscape. However, if it had been located at the border of two different landscape sites, prediction using the geographically closest soil samples would probably not have been the best model.

The morainic soil of eastern Denmark is similar to that at the Silstrup site and the geographically closest samples were also dominated by moraine material. The calibration results from the eastern Denmark and Silstrup data sets were almost the same, but the predictions were different, which might be explained by geological history. The Silstrup field site is located in northern Jutland, where the soil parent material is covered by a glacial advance from Norway, while in eastern Denmark the soil is covered by a glacial advance from the eastern Baltic Sea. No previous study has, in fact, used the geographically closest and the same parent material method to reduce the soil spectral library. This study found that soil parent material and landscape are relevant to the accuracy of a reduced library model.

Selecting the most alike spectra from the library did not improve the results compared with spiking the library. Wetterlind and Stenberg found that spiking a reduced library was better than spiking the entire library for SOC prediction, because the smaller data set could more easily integrate the characteristics of the target site after spiking. Our study confirms that spiking the library can improve the predictive abilities of the model.

**Conclusions**

Comparison of model predictions based on a national spectral library, a spiked library and three sub-sets of the library showed potential for the national vis-NIR spectroscopy library to predict SOC at field scale. Of all the models tested, using the geographically closest method to reduce the library resulted in the best prediction of SOC at field scale. The accuracy of the reduced library model was found to be highly dependent on geographical location, soil parent material and landscape. While the calibration based on the library was not the best model, it was still robust because the library data set covered the range of soil types in Denmark. Spiking the library with 30 samples improved the final predictive ability of the model.

**Acknowledgements**

This study was funded by China Scholarship Council (CSC) and the international project Soil Infrastructure, Interfaces and Translocation Processes in Inner Space (Soil-it-is) funded by the Danish Research Council for Technology and Production Sciences [http://www.agrsci.dk/soil-it-is/], and by the Danish Pesticide Leaching Assessment Programme [www.pesticid-varsling.dk].

**References**


23. NERI, Areal Informations Systemet [AIS]. [In Danish.], National Environmental Research Institute, Denmark. (2000).


Quantification of SOC and Clay Content Using Vis-NIR-MIR Spectroscopy with Jack-knifing Partial Least Squares Regression

Peng Yi, Maria Knadel, René Gislum, Kirsten Schelde, Anton Thomsen, Mogens Humlekrog Greve

Accepted for publication in Soil Science (2014)

Reprinted permission of Lippincott Williams & Wilkins
Quantification of SOC and Clay content Using Vis-NIR-MIR Spectroscopy with Jack-knifing Partial Least Squares Regression

Accepted for publication in Soil Science

Yi Peng, Maria Knadel, René Gislum, Kirsten Schelde, Anton Thomsen, Mogens Humlekrog Greve

1Department of Agroecology, Aarhus University, Blichers Allé 20, 8830 Tjele, Denmark
2Department of Agroecology, Aarhus University, Forsøgsvej 1, DK - 4200 Slagelse, Denmark
3Department of Chemistry, Faculty of Chemistry and Chemical Engineering, Zunyi Normal College, Shanghai Road 830, 563002, Zunyi, Guizhou, China

Corresponding author: Yi Peng

Phone: +45 8715 4789; E-mail: Yi.Peng@agrsci.dk
ABSTRACT

A total of 125 soil samples were collected from a Danish field varying in soil texture from sandy to loamy. Visible near-infrared reflectance (Vis-NIR) and mid-infrared reflectance (MIR) spectroscopy combined with chemometric methods were used to predict soil organic carbon (SOC) and clay contents. The main objective of this study was to find the best model for predicting SOC and clay content in the sampled field using Vis-NIR, MIR, and the combination of Vis-NIR and MIR, and using different model development techniques. The secondary objectives were: (i) to use iterations of calculation to find the optimal number of replicates for MIR measurements based on the root mean square error of cross validation (RMSECV) and (ii) to apply partial least squares regression (PLSR) in combination with jack-knifing (JK) to identify the most important part of spectral variables and the best model for predicting SOC and clay content. The study showed that with repeated MIR measurements it was possible to improve RMSECV by 20%. The optimal number of repeated MIR measurements was between 3 and 4 for SOC and clay content. Comparing all the prediction results, the combination of MIR and Vis-NIR with the PLSR-JK technique resulted in the lowest prediction errors, (RMSECV_{soc} of 0.35% and RMSECV_{clay} of 1.05%). The average uncertainties of laboratory measurements were 0.39% and 1.86% for SOC and clay contents, respectively. All models had acceptable and – to a large extent – comparable margins of error. PLSR with jack-knifing simplified and enhanced the interpretation of the developed models due to a reduction in the number of variables used in the models.

Key words: SOC, Clay, Spectroscopy, Vis-NIR-MIR, PLSR, Jack-knifing
Introduction

The determination and quantification of soil properties is one of the most important aspects in digital soil mapping, precision agriculture and landscape management. However, traditional laboratory analysis of soil physical and chemical properties is relatively expensive and time-consuming (Stenberg et al., 2010). Moreover, the precision of conventional wet chemistry analysis may not always be constant due to uncertainties in the analyses which cannot be eliminated (Mark et al., 1989; Aastveit and Marum, 1991; Faber and Kowalski, 1997; Sørensen, 2002). In order to overcome the limitations of conventional wet chemistry analysis, spectroscopy can provide a cheaper and more precise analytical technique for soil analysis (Rossel et al., 2006).

Visible near-infrared reflectance (Vis-NIR) spectroscopy is one the spectroscopic methods, it is characterized as the region from (350-2500 nm). This technique is based on molecular overtones and combination vibrations. Over the last 20 years, Vis-NIR spectroscopy and chemometrical tools have provided a rapid, low-cost and accurate analytical technique for soil analysis (Stenberg et al., 2010; Knadel et al., 2011). In soil spectroscopy research area, more than 200 publications have been generated in the last 15 years. Sørensen and Dalsgaard (2005) used 174 samples from different regions of Denmark to determine the uncertainty of wet chemistry analysis of clay content (range in clay content: 2-26%). The results showed that the estimated true standard error of prediction using Vis-NIR was 1.5-1.7% and the reproducibility standard deviation of the wet chemistry analysis was 1.3%. This means that the uncertainty of the Vis-NIR method was very similar to the reproducibility standard deviation of the lab measurements. Moreover, the intra-laboratory (different laboratory) reproducibility standard deviations of the Vis-NIR measurement and reference method were 0.7% and 1.1%, respectively. The results showed that the Vis-NIR method was more reproducible than the reference method for clay measurement. Peng et al.
(2013) improved the prediction accuracy using a different approach including subdividing the Vis-NIR spectral library of Danish soils for building the model for prediction of soil organic carbon (SOC) of validation samples from a Danish agricultural field. They created a subset of the library by selecting the samples that were geographically closest to the study area. The results showed that the root mean square error of prediction (RMSEP) for SOC was 0.19%. The RMSEP value was 0.01% greater than the standard deviation for the uncertainty of wet chemistry analysis. The authors concluded that a model based on a regionalized Vis-NIR spectral library is preferable compared to a model based on a global library.

Nevertheless, The absorption bands for soil organic matter in the Vis-NIR region are usually weak because vibration bands in the Vis-NIR spectral region (350-2500nm) are broad and overlapping, resulting in a weak signal for specific chemical compounds. In addition, water has strong absorption bands around 1400 and 1900 nm. Several other organic and inorganic molecules also absorb light energy in this region, which can make the spectra difficult to interpret. Furthermore, some inorganic chemicals (e.g. silica) are not significant contributors to the NIR spectra (Reeves, 2010; Stenberg et al., 2010; Bellon-Maurel and McBratney, 2011).

More recently, mid-infrared reflectance (MIR) spectroscopy has also been adopted for soil analysis. Mid-infrared radiation is found between 2500 nm and 25000 nm (Reeves, 2010; Bellon-Maurel and McBratney, 2011). Many types of organic and inorganic compounds can be more specifically identified by the MIR spectrum, due to intense vibration fundamentals. The MIR spectral features can be characterised by specific organic and inorganic spectral bands, whereas the NIR region gives much weaker and broader signals from vibration overtones and combination bands (McCarty et al., 2002; Stuart, 2004; Grinand et al., 2012). Thus, many studies have combined Vis-NIR and MIR regions for
quantitative and qualitative analysis of soil properties. Rossel et al. (2006) calibrated separately spectra from visible (400-700nm), NIR (700-2500nm), MIR (2500-25000nm) and the combination of all regions to predict soil properties (including, among others, SOC and soil texture). The results showed that there was only a minor difference in the predictions of SOC and soil texture using NIR, MIR or the combination of all spectral regions. However, these results also demonstrated a potential of using Vis-NIR and MIR for efficiently interpretation of soil spectra. Many studies have shown that the most important vibration bands for soil organic matter are in the MIR region, from around 2750 nm to 9523 nm (Janik and Skjemstad, 1995; McDowell et al., 2012). In the MIR spectral region, specific spectral features for aromatic (around 6250 nm and 6600 nm), carboxylic (around 3425nm, 5780nm) and carbohydrate groups (around 2857 nm and 4650 nm) can be found due to the stretching of the C-H, C=O and C=C double bonds and carbohydrate C-OH deformation vibration, which are essential for identifying SOC (Ibrahim et al., 2008; Calderon et al., 2011; Grinand et al., 2012; Song et al., 2012). For the main constituents of clay, kaolinite could be detected around 5236, 5747, 5525, 10940 nm by Si-O stretching, montmorillonite around 2915 nm by O-H stretching of water molecules and illite around 6112 nm by H-O-H stretching vibrations (Nguyen et al., 1991; Madejova and Komadel, 2001; Oumabady Alias Cannane et al., 2013).

However, not all spectral data are useful or important for the modelling of a given property. If we can remove useless and unreliable variables, the model may not suffer and it may even improve the extraction of important spectral features for specific soil properties (e.g. SOC). It would also make the model easier to interpret than a model based on the full spectral range (400-25000 nm). Thus, it would be an advantage if variable selection was applied to the interpretation of soil spectra and quantification of soil properties.
The term variable or feature selection is also called “frequency” or “wavelength” selection when applied to spectroscopic data (hereafter referred to as ‘variable selection’). Variable selection is an important step in improving model performance in multivariate analysis, because the removal of many non-significant variables will simplify the model and generally lead to better predictions. We aim for a model that is adequate while using as few variables as possible. It is particularly important to reduce the number of variables for the spectroscopic data obtained from high-resolution instruments, due to the amount of data (Andersen and Bro, 2010).

Jack-knifing (JK) (Efron and Efron, 1982; Martens and Martens, 2000) is a powerful variable selection tool for assessing the significance of variables commonly used in chemometrics. It is a statistical significance test seeking to eliminate variables with non-significant regression coefficients. Variables with relatively low regression coefficients are usually considered non-significant and supernumerary to the model (Andersen and Bro, 2010). JK can conveniently be performed by the models generated during cross-validation.

In this study, partial least squares regression (PLSR) (Martens, 1991) in combination with JK was applied to identify the important variables for SOC and clay content estimation from Vis-NIR, MIR and Vis-NIR-MIR spectral regions, and compared to the results of PLSR alone.

The precision of any soil analysis depends on the repeatability of the measurement. The way to represent the variation in soil samples is to increase the number of replicate measurements. However, to the authors’ knowledge no studies have demonstrated the repeatability of NIR and MIR measurements and their effect on the final prediction results.

The aims of this study were: (i) to test the repeatability of Vis-NIR and MIR measurement using the iterations of calculation to find the optimal number of replicates for NIR and MIR measurements and (ii) to apply the PLSR-JK technique on different spectral regions
in order to find the best model and relevant spectral information for SOC and clay content estimation. Additionally, we hypothesised that using the important variables identified by PLSR-JK could produce better or equivalent prediction results compared to the model using the entire spectrum and PLSR alone.
Materials and methods

Field site and sampling
A total of 125 topsoil (0-20 cm) samples were collected from a 14.6 ha agricultural field in Sørvad, west Jutland, Denmark (Figure 1), in May 2012. The study site is situated on an old moraine; with the major soil type classified as a podzol (Madsen et al., 1992). The site and sampling procedure have been described in detail by Knadel et al. (2013a). Soil samples were air-dried, sieved to 2 mm, and analysed for SOC using a LECO analyzer coupled with an infrared CO$_2$ detector (Thermo Fisher Scientific Inc., USA). Soil texture (clay, silt, sand) was determined by a combined sieve/hydrometer method (J.H. Dane and C. Topp, 2002). Clay was defined as particles <2 µm, silt 2-63 µm, fine sand 63-200 µm, and coarse sand 200-2000 µm.

Vis-NIR measurements
Vis-NIR spectra were obtained with a NIRS$^\text{TM}$ 2500 instrument (FOSS, Hillerød, Denmark). The spectral range was between 400 and 2500 nm with a resolution of 0.5 nm, giving 4200 data points. The soil Vis-NIR spectra were obtained using the following method. A rotating sample cup (diameter: 7cm) containing around 50 g of soil was scanned four times at seven different points and averaged. Ten measurements were carried out for each sample. The Kennard-Stone algorithm (Kennard and Stone, 1969) was applied to select 50 representative soil samples for SOC and soil texture lab analysis. The details of selection procedure being described in (Knadel et al. submitted). The present study only included these 50 samples (Figures 1 and 2).

MIR measurements
MIR spectral measurements were performed on air-dry samples finely ground to powder (approximately < 80 µm) (Rossel et al., 2006). Samples were carefully placed on the
surface (2mm diameter) made of diamond and illuminated with light from below. To test the effect of replication, 10 measurements were carried out for each of the 50 samples. Measurements were made using a Fourier transform (FT) IR MB3000 spectrometer from ABB (Q-interline, Roskilde, Denmark) with an attenuated total reflectance (ATR) accessory at a resolution of 16 cm$^{-1}$. Sixty-four scans were taken per sample before Fourier transformation of the spectra to an average spectrum. Spectra were recorded in the absorbance mode from 2500 to 25000 nm. To standardize the units for both instruments, nanometer (nm) was hereafter used. After noise removal at the edges of the MIR spectra, only spectra between approximately 2500-14089 nm were used for further data analysis. Thus, each spectrum consisted of 428 spectral data points in total.

**Multivariate data analysis**

The calculations were performed with the Unscrambler X 10.2 software (Camo ASA, Oslo, Norway). Principle component analysis (PCA) (Wold et al., 1987) was used for exploratory data analysis of Vis-NIR and MIR spectra to detect outliers or extreme values defined with a Mahalanobis distance $H > 3$. Different spectra pre-processing methods ($1^{st}$ derivative, $2^{nd}$ derivative (Savitzky and Golay, 1964), multiplicative signal correction (MSC) (Geladi et al., 1985), and standard normal variate (SNV) (Barnes et al., 1989)) were applied to raw spectra to correct for sample variation and measurement noise. Calibration models were developed with PLSR with segmented cross-validation. The optimal number of segments, the optimal pre-processing method and number of principal components were determined by minimizing the RMSECV. The models were evaluated by correlation coefficients ($R^2$), root mean square error of calibration (RMSEC), RMSECV, bias and the ratio of performance to interquartile distance (RPIQ) (Bellon-Maurel et al., 2010). The RPIQ was defined as $RPIQ=IQ/SEP$, where $IQ=Q3-Q1$; IQ being the interquartile distance of the validation set, Q1 the median of the first half of the validation set and Q3 the median.
for the second half of the validation set. In this study, RMSECV was used for calculations instead of SEP. We consider the best model to be the one with the lowest RMSECV and the highest $R^2$ and RPIQ values.

Initially we applied PLSR to correlate Vis-NIR, MIR and Vis-NIR-MIR spectra with SOC and clay contents individually. Thereafter PLSR with JK was applied to the same data sets to identify important variables and recalibrate to improve model performance. The best model for each soil property was identified from the lowest RMSECV.

*Repeated Vis-NIR and MIR measurements*

In order to identify outliers of replicate measurements for each sample, PCA was applied to each sample (10 replicates) individually. If outliers were found the measurement was performed again to ensure there were 10 acceptable measurements for each soil sample.

Next we averaged replicates from one up to ten for each sample. For each average calculation the measurements were selected randomly with replacement. To eliminate the effect of random selection, 10, 100 and 1000 iterations of the calculation were individually applied for each average calculation.

Partial least squares regression models were developed on all 50 samples, for each modelling progressively with one to ten replicates. The average spectra were an average of the spectra from the same samples. The optimal number of replicates was determined by RMSECV and these optimal average spectra were used for further analysis.

*PLSR with Jack-knifing*

Jack-knifing is a resampling method based on the assessment of the relative importance of the PLS regression coefficients (Efron and Efron, 1982; Karaman *et al*., 2013). It has been modified for PLSR by Martens and Martens (2000). Normally, the analysis of variability of the PLS regression coefficients gives information about the significance of variables. Here
JK was used to select variables by calculating the uncertainty estimates associated with the regression coefficients and the loadings (Martens and Martens, 2000).

Jack-knifing is based on PLSR with cross-validation in which samples are split into $M$ subsets. In the present study 50 samples were randomly split into five subsets. First, regression coefficients $b$ were calculated for each variable when PLSR model was completed. Afterwards, regression coefficients for each sub-model of the cross-validation steps $b_m$ were calculated. The uncertainty variance of the regression coefficients $s^2(b)$ was estimated by following formula:

$$s^2(b) = \left( \sum_{m=1}^{M} (b - b_m)^2 \right) \left( \frac{M - 1}{M} \right)$$

Where $M$ is the number of the sample sets for cross segment validation.

The $t$-test of significance was performed individually for each element in $b$ relative to the square root of its estimated uncertainty variance $s^2(b)$, giving the significance level ($p$-value) for each parameter (Westad and Martens, 2000). Variable selection was done by comparing the $p$-value associated with the regression coefficients with a predefined threshold value ($p<0.05$). We removed variables with $p$-value greater than 0.05 and a new model was generated on the remaining variables. In order to improve validation results and remove irrelevant variables as much as possible, we repeated JK 10 times for the new calibration model.

**Prediction and lab error**

Root mean square deviation (RMSD) of reference data was used to estimate the uncertainty in lab wet chemistry analysis. To evaluate the accuracy of prediction results, we considered a value of RMSECV equal to or lower than RMSD as acceptable. Since soil samples in the present study are highly variable, we first applied the K-means algorithm to split samples into two subsets, and separately calculated and compared RMSD and RMSECV for each
subset. Next we followed the same criteria for creating three subsets for each property. We compared the results from the two-subset and three-subset groups. The RMSD was calculated by the following formula:

$$RMSD = \sqrt{\frac{\sum (y_i - y_{mean})^2}{n}}$$

where $y_i$ is the reference value of sample $i$, $y_{mean}$ is the mean value for each group, and $n$ is the number of samples.
Results and discussion

Exploratory data analysis

The study site represents a highly heterogeneous agricultural field. The SOC, clay, silt and sand contents ranged from 1.7 to 2.2%, 5 to 24%, 0.95 to 14% and 58 to 90%, respectively (Table 1). The 50 representative samples selected for lab analysis covered all variation of the total 125 samples (Figure 2). It is important to note that the Kennard-Stone algorithm in this study was applied only to soil Vis-NIR spectra. Thus, the 50 representative samples covered only the variation of the 125 soil Vis-NIR spectra and may not cover all soil characteristics.

Repeated Vis-NIR and MIR measurements

Of the 500 replicate Vis-NIR and MIR spectra, five outliers were found in the MIR spectra (data not shown). We remeasured these samples and as the new spectra did not show any extreme values they were used in the further data analysis.

The difference in the RMSECV value between the single and the 10-replicate Vis-NIR measurement was 0.04% and 0.07% for SOC and clay, respectively. This is considerably lower than the RMSECV values; thus, the repeatability of Vis-NIR measurement could not be significantly improved by repeating measurements rather than using a single measurement. The result showed that the Vis-NIR instrument used in this study produced very repeatable results. This is because the Vis-NIR instrument was equipped with a 7cm diameter sample cup, each measurement therefore being an average of seven different measuring points.

Figure 3 shows that the RMSECV value from the MIR measurement decreased significantly from one to three replicates, by about 20%. The RMSECV value for all soil properties fell markedly with increasing number of replicates up to 3-4 replicates,
whereafter the improvement was minor. Using 1000 iterations also gave more stable results. We stopped iterations at 1000 due to time constraints. Thus, we used four replicates with 1000 iterations for further analysis.

In theory, random errors exist in all measurements, and the poor repeatability of measurement could be overcome by using repeated measurements. Our results showed that 3-4 replicates was the optimal number for soil MIR measurement with the instrument used in this study. For this setup soil samples were measured from the 2mm diameter measuring surface. If we assume that random error in this measurement was mainly due to the small area scanned, we could improve the repeatability of MIR measurements by enlarging the area.

**Comparison of cross-validation using Vis-NIR, MIR, Vis-NIR-MIR and JK**

The best performance of the calibration models using different spectral pre-processing methods (1st and 2nd derivative, SNV, MSC, SNV-Detrend) is shown in Table 2 and Figure 4. Different calibration models based on the different spectral region generally produced very similar validation results for different parameters (Figure 4 and Table 2). The results from Vis-NIR and the Vis-NIR-MIR combination were almost the same for clay and SOC. Comparing the calibration models of Vis-NIR, MIR and Vis-NIR-MIR for each parameter, there was little difference in the validation results, though the highest RMSECV value was obtained with MIR. In previous studies, the MIR calibration model gave more accurate results than the Vis-NIR model (Rossel *et al*., 2006). However, in our study, the RMSECV value using Vis-NIR was better than when using MIR. This could be explained by the resolution of the instrument. As previously mentioned, Vis-NIR spectra were collected with a spectral resolution of 0.5nm giving a much higher resolution than MIR spectra that had a spectral resolution of 16 cm$^{-1}$. Additionally noise was removed from the spectra, leaving a total of 428 spectral data points for further analysis. Thus, the MIR spectra in this
study probably lost some information on soil chemistry, mineralogy and physical properties.

For all the models, the results from using important variables selected with the JK-PLSR technique always gave lower RMSECV and higher $R^2$ values, which were comparable to using PLSR alone (Table 2). This is because JK-PLSR selected important wavelengths based on the regression coefficients and removed unreliable variables at the same time. Additionally, one of the hypotheses in this study was proved in that the performance of the model using important variables was slightly better than the model based on the entire spectrum.

Of the six calibration models based on the different spectral data for SOC and clay, the PLSR-JK calibration model used fewer variables and factors, resulting in lower RMSECV and higher $R^2$ values (Table 2). The models based on the Vis-NIR (JK) and MIR (JK) yielded more or less the same prediction results as models based on the full spectrum, but the JK technique simplified the models. Comparing the use of Vis-NIR-MIR for the full spectrum with the use of important variables, the RMSECV value was reduced by approximately 17% and 10% for SOC and clay, respectively, when using important spectra, which was a substantial improvement for both. The Vis-NIR-MIR model used 4628 spectral variables for calibrating while the JK-PLSR technique only around 20% of the Vis-NIR-MIR spectra for SOC and clay calibration. Our results are in agreement with those of Martens and Martens (2000) who found that JK-PLSR is an efficient technique for removing unimportant variables when there are many spectroscopic variables. In order to remove all unreliable variables, we repeated the JK-PLSR procedure on the previously selected important variables. However, no further improvement was found after 10 iterations (data not shown).
Important variables and description of spectra

The PLS regression coefficients of important variables are shown in Figure 5. All the important variables were selected by the JK technique. For both SOC and clay content, the regression coefficients in the MIR spectra were generally higher than the values in the Vis-NIR spectra. Therefore, MIR spectra yielded more important information on SOC and clay content than the Vis-NIR spectra (Figure 5). The important bands from the JK procedure mainly represented the functional groups C-H, C-O, C=O, C-N, C=C, C≡C. The functional groups O-H, H-O-H, Si-O, and Al-O were related to clay content.

For SOC, information related to organic matter was found in the Vis-NIR region around 474, 556, 638, 832, 1265, 1720, and 2165-2443nm. The water signature was found around 1900nm (Figure 5). In the MIR region, C-H stretching bands from aliphatic hydrocarbon compounds were found in the range 2986-3522 nm. Around 3420nm and 3503nm two strong absorptions could be ascribed to methylene asymmetric C-H stretching and methylene symmetric C-H stretching, respectively (Smith, 1999; Stuart, 2004; Janik et al., 2007). The C≡C triple-bond could be detected in the 4379-4900nm region. The most intense stretching bands in the spectra, C=O and C≡C, were found in the 4970-5761nm region. These features are formed by esters and carboxylic acids (Stuart, 2004; Janik et al., 2007; Calderon et al., 2011; Song et al., 2012). Only three variables were selected between 6000-7000nm. Absorption at 6323nm and 6354nm could be ascribed to pyridine C=N stretching and C≡C stretching, respectively (Stuart, 2004). Another absorption band representing CH₂ bending was found at 6931nm. This alkyl deformation band is very important for SOC determination (Janik et al., 2007; Calderon et al., 2011). Absorption bands between 7323-8878nm have been ascribed to C-O stretching and O-H bending in humic acid and aromatic amines (Stuart, 2004; Janik et al., 2007; Grinand et al., 2012).
Clay minerals are hydrous aluminium phyllosilicates, with variable amounts of iron and magnesium. For clay minerals, in the Vis-NIR region the strongest absorption bands were found in the visible part (410-572nm), these features being mostly due to Fe (Rossel and Behrens, 2010). The first overtone of OH-stretching of kaolin and water bands was found near 1400 and 1900 nm. The aluminum and magnesium could be found at 2200nm and 2300nm, respectively (Hunt, 1977; Rossel and Behrens, 2010; Stenberg et al., 2010).

In the MIR region, the kaolin group with O-H stretching vibration was generally found in the 2500-3600nm region. We found strong absorption bands in the ranges 5469-6085nm, 9258-9970nm and 10624-12708nm, which could also indicate kaolinite (Farmer and Russell, 1966; Farmer, 1968; Chester and Elderfield, 1973; Marel and Beutelspacher, 1976; Hunt, 1977; Vaculikova and Plevova, 2005).

**Accuracy of Vis-NIR-MIR spectroscopy**

In order to apply a spectroscopy technique to soil property analysis, the accuracy of the technique has to be evaluated. In the present study the comparison of the uncertainty of wet chemistry analysis and RMSECV is shown in Table 3. Generally the RMSECV values were lower than the RMSD values, indicating that the prediction results were acceptable. Data sets with wide-ranging data always give a high value for the uncertainty: consequently, the high RMSD value in this study was from a data set with a broad range. The average RMSD values from three-subset group were lower than the RMSD values from the two-subset group.

When comparing the two-subset group we found that all RMSECV values were lower than the RMSD for both SOC and clay contents. The average values of RMSD were 0.56% and 2.73% for SOC and clay content, respectively (Table 3). Sørensen and Dalsgaard (2005) used a similar range of data (clay content: 2-26%) for Danish soils as in the present study (5-24%) to calculate the uncertainty for the wet chemistry analysis of the clay content.
They obtained a reproducibility standard deviation of 1.3%. Hence, the two-subset group of RMSD value for clay and SOC in this study could be much higher than the real uncertainty of the wet chemistry analysis, at least for the clay content.

Because the RMSD values from the two subsets did not indicate reliable results, we focused on the RMSD values for the three subsets (Table 3). For SOC, the average RMSD value gave the same average RMSECV value of 0.39%. However, when comparing the results between the three subsets, we found that the result from one of the subsets was not acceptable (extreme subset). In this extreme subset, the RMSD value of SOC was 0.19% and the RMSECV of SOC was 0.36%, indicating that the prediction error was approximately twice as high as the uncertainty of wet chemistry analysis. This result can be explained by the variation of the calibration data set. In this comparison all the RMSECV values of SOC were predicted by one model rather than sub-models; this model was developed on the SOC range of 1.51%-6.61%, but the extreme group data set held the lowest SOC value and smallest range of 1.51%-2.09%. This means that the model we used here could not represent well the samples from the extreme group. Therefore, the RMSECV value was relatively higher than the RMSD value.

Comparing the clay content for the three subsets, we found a similar situation as with SOC. The average value of RMSD for the three subsets was 1.86% for clay content, which was close to 1.3% found by (Sørensen and Dalsgaard, 2005). Thus, we concluded that the RMSD value using three subsets would be more practical than if based on two subsets in this study. It is important to note that the choice of data division is highly dependent on the data range and sample number.

In order to apply the spectroscopy technique to soil analysis on a routine basis in agriculture and ecology, the evaluation of the accuracy of the technique is a crucial step in the analysis. Essentially, to calculate repeatability or reproducibility of lab measurements,
you need to also perform repeated wet chemistry analyses. However, we were only able to use reference data to estimate lab uncertainty, due to the high cost of wet chemistry analysis.
Conclusions

The present work demonstrated that the JK-PLSR technique yielded an efficient variable selection method for soil Vis-NIR-MIR spectra. We applied PLSR and PLSR-JK chemometrics tool to calibrate Vis-NIR, MIR and Vis-NIR-MIR spectra. Of the models tested, JK-PLSR only used around 20% of the entire Vis-NIR-MIR spectral data set for calibration and gave the lowest RMSECV (RMSECV$_{soc}$ of 0.35% and RMSECV$_{clay}$ of 1.05%). The predictive ability of JK-PLSR was better than using PLSR alone. Since the JK-PLSR technique simplified the model and selected important variables, most of the spectral features of SOC and clay were described. When comparing RMSECV and RMSD, the accuracy of the prediction results from soil spectra in this study was acceptable. According to these results, we consider that the technique of Vis-NIR and MIR spectroscopy combining with JK-PLSR could replace laboratory traditional wet chemistry SOC and clay analysis in this study. These two models can be applied to more unknown soil samples from target site for SOC and clay contents analysis, and facilitate soil digital mapping or other agriculture applications on target field. We conclude that the spectroscopy technique combining with proper chemometric tool could efficiently at least partly replace expensive and time consuming laboratory traditional wet chemistry SOC and clay analysis. The study also showed that 3-4 replicates were the optimal number for MIR measurement of SOC and clay content. Thus, we suggest that the accuracy of spectroscopy is highly dependent on the resolution of the instrument.

Acknowledgements

This study was funded by China Scholarship Council (CSC) and the HOBE project. We also thank FOSS Company for providing the NIR instrument.
References


Knadel, M., Y. Peng, K. Schelde, A. Thomsen, F. Deng and M.H. Greve, 2014, Sensor fusion for the identification of hydrological response units at field scale, European Journal of Soil Science, (Submitted)


Oumabady Alias Cannane, N., M. Rajendran and R. Selvaraju, 2013. Ft-ir spectral studies on polluted soils from industrial area at karaikal, puducherry state, south india. Spectrochimica acta. Part A, Molecular and biomolecular spectroscopy, 110: 46-54

Reeves, J.B., 2010. Near-versus mid-infrared diffuse reflectance spectroscopy for soil analysis emphasizing carbon and laboratory versus on-site analysis: Where are we and what needs to be done? Geoderma, 158(1-2): 3-14


Figure Legends

**Figure 1.** Location of study site and all 125 sampling points (●), and 50 representative samples for calibration (∗).

**Figure 2.** PCA score plot for PC1 versus PC2 for all 125 soil spectra (●), and 50 representative samples for calibration (O).

**Figure 3.** RMSECV value of SOC and clay content vs. number of replicate MIR measurements and using 10, 100 or 1000 iterations of the calculations. MIR spectral range from 2500 to 14089nm.

**Figure 4.** RMSECV value of SOC and clay content from different calibration data sets (Vis-NIR, MIR, Vis-NIR-MIR, Important spectra).

**Figure 5.** Regression coefficient plots from JK-PLSR models for SOC and clay content. JK selected 1017 and 1156 important variables for SOC and clay content, respectively.
Figure 1.
Figure 2.
Figure 3.

**SOC**

- RMSECV(%) vs Number of replicates
- 10 iterations
- 100 iterations
- 1000 iterations

**CLAY**

- RMSECV(%) vs Number of replicates
Figure 4.
Figure 5.
Table 1. General statistics of soil properties of the 50 calibration samples at the study site.

<table>
<thead>
<tr>
<th>Soil Properties</th>
<th>Min</th>
<th>Max</th>
<th>Mean</th>
<th>SD&lt;sup&gt;a&lt;/sup&gt;</th>
<th>Q1&lt;sup&gt;b&lt;/sup&gt;</th>
<th>Q3&lt;sup&gt;c&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>SOC</td>
<td>1.51</td>
<td>6.61</td>
<td>3.30</td>
<td>1.35</td>
<td>2.12</td>
<td>4.39</td>
</tr>
<tr>
<td>Clay</td>
<td>5.00</td>
<td>24.00</td>
<td>9.49</td>
<td>3.92</td>
<td>6.53</td>
<td>10.00</td>
</tr>
<tr>
<td>Silt</td>
<td>0.95</td>
<td>14.00</td>
<td>5.17</td>
<td>2.50</td>
<td>3.40</td>
<td>6.05</td>
</tr>
<tr>
<td>Sand</td>
<td>58.00</td>
<td>90.00</td>
<td>79.68</td>
<td>7.49</td>
<td>76.00</td>
<td>86.00</td>
</tr>
</tbody>
</table>

<sup>a</sup> standard deviation

<sup>b</sup> the median of the first half of the data set

<sup>c</sup> the median of the second half of the data set
Table 2. Calibration and segmented cross-validation (5 segments) statistics for SOC and clay using PLSR and reduced spectra with jack-knifing on Vis-NIR (400-2500nm), MIR (2500-14000nm) and the combination of Vis-NIR-MIR (400-14000nm).

<table>
<thead>
<tr>
<th></th>
<th>Calibration</th>
<th>Validation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Number of Variables</td>
<td>Transform</td>
</tr>
<tr>
<td><strong>SOC</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Vis-NIR</td>
<td>All wavelengths(4200)</td>
<td>log(1/R)</td>
</tr>
<tr>
<td>Vis-NIR(JK)</td>
<td>Important variables (925)</td>
<td>log(1/R)</td>
</tr>
<tr>
<td>MIR</td>
<td>All wavelengths(428)</td>
<td>2&lt;sup&gt;nd&lt;/sup&gt;e 9point</td>
</tr>
<tr>
<td>MIR</td>
<td>Important variables (122)</td>
<td>2&lt;sup&gt;nd&lt;/sup&gt;e 9point</td>
</tr>
<tr>
<td>Vis-NIR-MIR</td>
<td>All wavelengths(4628)</td>
<td>SNV</td>
</tr>
<tr>
<td>Vis-NIR-MIR(JK)</td>
<td>Important variables (1017)</td>
<td>SNV</td>
</tr>
<tr>
<td><strong>CLAY</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NIR</td>
<td>All wavelengths(4200)</td>
<td>MSC</td>
</tr>
<tr>
<td>Vis-NIR(JK)</td>
<td>Important variables (1038)</td>
<td>MSC</td>
</tr>
<tr>
<td>MIR</td>
<td>All wavelengths(428)</td>
<td>snv-detrend</td>
</tr>
<tr>
<td>MIR</td>
<td>Important variables (62)</td>
<td>snv-detrend</td>
</tr>
<tr>
<td>Vis-NIR-MIR</td>
<td>All wavelengths(4628)</td>
<td>1&lt;sup&gt;st&lt;/sup&gt;f 21point</td>
</tr>
<tr>
<td>Vis-NIR-MIR(JK)</td>
<td>Important variables(1156)</td>
<td>1&lt;sup&gt;st&lt;/sup&gt;f 21point</td>
</tr>
</tbody>
</table>

<sup>a</sup> Root mean square error of calibration,  <sup>b</sup> Correlation coefficients,  <sup>c</sup> Root mean square error of cross validation,  <sup>d</sup> ratio of performance to interquartile distance,  <sup>e</sup> 2<sup>nd</sup> derivative,  <sup>f</sup> 1<sup>st</sup> derivative,  <sup>g</sup> Jack-knifing with partial least squares regression.
Table 3. The comparison between the RMSD of reference data and RMSECV in different groups of soil samples for SOC and clay content.

<table>
<thead>
<tr>
<th></th>
<th>N\textsuperscript{a}</th>
<th>RMSD\textsuperscript{b}</th>
<th>RMSECV\textsuperscript{c}</th>
<th>Min</th>
<th>Max</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SOC</td>
<td>2 subsets</td>
<td>25</td>
<td>0.43</td>
<td>0.38</td>
<td>1.51</td>
<td>2.73</td>
</tr>
<tr>
<td></td>
<td></td>
<td>25</td>
<td>0.68</td>
<td>0.41</td>
<td>3.48</td>
<td>6.61</td>
</tr>
<tr>
<td></td>
<td>average</td>
<td></td>
<td>0.56</td>
<td>0.40</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2 subsets</td>
<td>13</td>
<td>13</td>
<td>0.19</td>
<td>0.36</td>
<td>1.51</td>
<td>2.09</td>
</tr>
<tr>
<td>3 subsets</td>
<td>13</td>
<td>13</td>
<td>0.32</td>
<td>0.39</td>
<td>2.20</td>
<td>3.48</td>
</tr>
<tr>
<td></td>
<td>average</td>
<td>24</td>
<td>0.66</td>
<td>0.41</td>
<td>3.65</td>
<td>6.61</td>
</tr>
<tr>
<td>Clay</td>
<td>2 subsets</td>
<td>10</td>
<td>3.76</td>
<td>1.60</td>
<td>12.00</td>
<td>24.00</td>
</tr>
<tr>
<td></td>
<td></td>
<td>40</td>
<td>1.69</td>
<td>1.15</td>
<td>5.00</td>
<td>11.00</td>
</tr>
<tr>
<td></td>
<td>average</td>
<td></td>
<td>2.73</td>
<td>1.38</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3 subsets</td>
<td>6</td>
<td>6</td>
<td>3.37</td>
<td>1.64</td>
<td>15.00</td>
<td>24.00</td>
</tr>
<tr>
<td></td>
<td></td>
<td>23</td>
<td>1.37</td>
<td>1.25</td>
<td>8.30</td>
<td>13.00</td>
</tr>
<tr>
<td></td>
<td>average</td>
<td>21</td>
<td>0.85</td>
<td>1.12</td>
<td>5.00</td>
<td>8.20</td>
</tr>
<tr>
<td></td>
<td>average</td>
<td></td>
<td>1.86</td>
<td>1.34</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

\textsuperscript{a} Number of samples for each group, \textsuperscript{b} Root Mean Square Deviation (RMSD) of reference data, \textsuperscript{c} Root mean square error of cross validation
Paper 3

*Modeling SOC at Regional Scale by Combining Multi-spectral Images with Laboratory Spectra*

Yi Peng a, Xiong Xiong b, Kabindra Adhikari c, Maria Knadel a, Sabine Grunwald b, Mogens Humlekrog Greve a

a Department of Agroecology, Faculty of Science and Technology, Aarhus University, Blichers Allé 20, P.O. Box 50, DK-8830 Tjele, Denmark

b Department of Soil and Water Science, University of Florida, Gainesville, FL 32611, USA

c Department of Soil Science, FD Hole Soils Lab, University of Wisconsin—Madison, Madison, WI 53706, USA

Manuscript in preparation
Abstract

There is potential to use soil proximal and remote sensing derived spectra concomitantly to develop soil organic carbon (SOC) models. Yet mixing spectral data from different sources and technologies to improve soil models is still in its infancy. The objective of this study was to incorporate soil spectral features indicative of SOC from laboratory visible near-infrared reflectance (Vis-NIR) spectra and incorporate them with remote sensing (RS) images to improve predictions of top SOC in the Skjern river catchment, Denmark. The secondary objective was to improve prediction results by separately calibrating samples from upland and wetland. We hypothesize that final prediction accuracy is significantly improved by incorporating laboratory Vis-NIR images upscaled from point-based spectra to catchment scale and RS data for topsoil SOC spatial modeling. A total of 328 topsoil samples were collected and analyzed for SOC. Satellite Pour l’Observation de la Terre (SPOT5), Landsat Data Continuity Mission (Landsat 8) images, laboratory Vis-NIR spectroscopy and other ancillary environmental data including terrain parameters and soil maps were compiled to predict top SOC using Cubist regression and Bayesian kriging. The results show that the model developed from RS data, ancillary environmental data and one specific laboratory spectral wavelength (1930nm) image yielded lower RMSE of 2.8% and higher $R^2$ of 0.59 than the model developed from only RS data and ancillary environmental data (RMSE = 3.6%, $R^2$=0.46). The prediction accuracy was significantly improved after adding laboratory Vis-NIR spectral images predictors. Vegetation indices, such as the Normalized Difference Vegetation Index (NDVI) and Enhanced Vegetation Index (EVI), were very important predictors for SOC spatial modeling. Furthermore, the upland model produced more accurate prediction results for samples from upland compared to the upland & wetland model. However, separately calibrate samples from upland and wetland did not improve prediction accuracy of wetland samples, due to vegetation coverage of wetland ecosystem. We conclude that a laboratory spectral feature adds critical information to
significantly improve the prediction accuracy of SOC compared to using RS data alone. Additionally, separate calibration of the wetland and upland datasets could improve prediction accuracy only for upland by narrowing down variation of soil samples, the complexity of wetland area results in difficulty in predicting SOC using RS data in Skjern river catchment.
1. Introduction

The reliable information of the spatial distribution of soil physical and chemical properties is required for sustainable land management and precision agriculture (Sumfleth and Duttmann, 2008, McBratney et al., 2003). Recently, remote sensing (RS) has been widely applied to digital soil mapping (DSM) due to its ability to overcome the issue of sparsely sampled soil property data at regional and national scale (Mulder et al., 2011, Ben-Dor et al., 2008). Grunwald et al. (2011) proposed STEP-AWBH (S, soil; T, topography; E, ecology; P, parent material; A, atmosphere; W, water; B, biotic; H, human) model to determine soil properties and classes. Based on this model, they emphasized that the “B” factor such as vegetation cover could be assessed through spectral indices derived from remote sensing imagery. These data can be easily acquired from authorities, for instance Satellite Pour l’Observation de la Terre and Landsat Data Continuity Mission images. In addition, some researchers derived different vegetation indices from imagery to indirectly obtain information which correlate with soil properties (Mulder et al., 2011, Rivero et al., 2009). Overall, in the last ten years RS showed to be useful as part of digital soil mapping. Mulder et al. (2011) reviewed the RS data that can be analyzed using empirical methods to predict soil properties, and this technique facilitates mapping large areas by reducing the need for costly sampling work and laboratory chemistry analysis.

Since Singh and Kathpalia (2007) found the variation of vegetation cover was linked soil physical and chemical properties, a series of vegetation indices have been developed based on the RS imagery. One of the most common indicators that reflect biomass status and indicate soil qualities of a specific site is Normalized Difference Vegetation Index (NDVI) (Rouse Jr et al., 1974). The NDVI is calculated from the Vis-NIR light reflected by vegetation. According to the theory of photosynthesis, during the growing season chlorophyll strongly absorbs visible light (from 400 to 700 nm) and reflects a large portion of the near-infrared light (from 700 to 1100 nm)
simultaneously. Thus, higher NDVI indicates higher photosynthesis activity and gross primary productivity. Sumfleth and Duttmann (2008) combined terrain variables and NDVI in a multiple linear regression model to predict total carbon content in paddy soil landscapes. The results showed that 41% of total carbon variation was explained in final prediction map, and they also found the spatial distribution of carbon was closely related to the NDVI of vegetated surfaces.

In order to extract sufficient information from RS imagery for prediction of soil properties, several indices have been proposed besides NDVI, e.g., Normalized Difference Vegetation Green Index (NDVI green), Simple Ratio (SR), Transformed Vegetation Index (TVI). Kim. et al. (2014) used 9 different spectral vegetation indices derived from 3 different satellite images incorporating with environmental ancillary data, and developed prediction models for soil phosphorus (TP) and total nitrogen (TN) for a wetland area. They found the indices of the TVI, SR, NDVI green were the top three most important predictors for TP predictions. The results also showed the NDVI index does not contribute strongly to the prediction ability in the complex wetland ecosystem. The main drawbacks of these data sources are: (i) large size of the pixels; (ii) and the wide spectral band that is difficult to capture detailed spectral information from soil (Ben - Dor et al., 2008). Thus, spectral indices derived from remote sensing imagery might not provide sufficient information for determining soil properties (McBratney et al., 2003, Mulder et al., 2011). Thus, combing RS images with high spectral resolution data source (i.e., laboratory spectroscopy) for modeling provides a promising solution.

The Vis-NIR spectroscopy has been well developed to analyze soil properties in the laboratory, with over 200 publications in past 15 years (Bellon-Maurel and McBratney, 2011). This technique has been commonly considered as a potentially efficient and low-cost technology for determination of soil properties. Vis-NIR is a spectroscopic method that uses both visible and near infrared
regions of the electromagnetic spectrum, between the range from 350-2500nm. The fundamental of Vis-NIR is based on molecular overtone and combination vibrations. Theoretically, there is a linear relationship between a concentration of a given constituent and absorbance (Williams and Norris, 1987). Many studies have shown that Vis-NIR combining with chemometric tools is a rapid and objective method to quantify several soil physical and bio-chemical properties (Sørensen, 2002, Peng et al., 2013). In comparison with RS image, laboratory Vis-NIR produces much higher resolution spectrum, better signal-to-noise ratio, no atmosphere attenuation, and more stable spectrum for the soil of a specific site. However, the key limitation of this technique is that laboratory Vis-NIR only provides spectral data at the point other than image. Hence, upscaling laboratory Vis-NIR spectra from point to image (laboratory Vis-NIR image) is necessary.

We hypothesize that final prediction accuracy of topsoil SOC is significantly improved by incorporating laboratory Vis-NIR images upcaled from point-based spectral feature to catchment scale and RS data. The objective of this study was to incorporate soil spectral features indicative of SOC from laboratory visible near-infrared reflectance (Vis-NIR) spectra and incorporate them with remote sensing (RS) images to improve predictions of top SOC in the Skjern river catchment, Denmark. The secondary objective was to improve prediction results by separately calibrating samples from upland and wetland.
2. Materials and Methods

2.1 Study area

The study area is Skjern river catchment located on the west coast of Jutland, Denmark (Figure.1) and covers an area of approximately 2500 km². This area is selected for a Hydrological Observatory (HOBE). The climate in this region is temperate maritime with a mean annual precipitation of 990 mm, and mean annual temperature of 8.2°C (Jensen and Illangasekare, 2011). The predominant soil texture is loamy sand (Greve et al., 2007). The maximum elevations of the area is around 125m above the sea level. Over 50% of land use is agriculture (i.e., grain and crops), followed by grass (30%), forests (7%), and heath (5%) (Jensen and Illangasekare, 2011).

2.2 Sampling and Laboratory analyses

A total of 328 topsoil samples were collected from the study area during middle of 70’s, 84 of sampling sites were located on the wetland area (Figure.1). All the samples were originally from Danish Soil Classification (Greve et al., 2007), which contains SOC value for each sample. Each sample was composed of 25 sub samples which were taken within a 70m × 70m square represent this site’s soil characteristics. SOC was determined by combustion in a LECO IR-12 furnace (Krogh et al., 2003). Soil texture were determined only for samples with SOC value lower than 6%, resulting in 296 samples with soil texture values (Table.1).

2.3 Laboratory spectra and RS data

Laboratory Vis-NIR spectra were obtained with a NIRSTM 2500 instrument (FOSS, Hillerød, Denmark). The spectral range was between 400 and 2500 nm with a resolution of 0.5 nm, giving 4200 data points. The soil laboratory Vis-NIR spectra were obtained using the following method. A rotating sample cup (diameter: 7cm) containing around 50 g of soil was scanned four times at seven different points and averaged. In total, 328 soil laboratory Vis-NIR spectra were collected.
The RS data were extracted from Satellite Pour l’Observation de la Terre (SPOT5) and Landsat Data Continuity Mission (Landsat 8) images. Due to less cloud cover in summer, SPOT5 images for this study were obtained from June 2012 and Landsat8 images were taken from June and July 2013. The SPOT5 images were provided by The Danish Geodata Agency, with Band1 (500-590nm), Band2 (610-682nm), Band3 (780nm-890nm) with 10m spatial resolution and Band4 (shortwave-infrared (SWIR); 1580-1750nm) with 20m spatial resolutions. The Landsat 8 images with spatial resolutions in the visible/near-infrared part of the spectrum of 30 m, which included Band1 (blue 450-510nm), Band2 (green 530-590nm), Band3 (red 640-670), Band4 (NIR 850-880nm), Band5 (SWIR1 1570-1650nm), Band6 (SWIR2 2100-2290nm) were provided by U.S. Geological Survey USGS, http://landsat.usgs.gov/band_designations_landsat_satellites.php, last accessed 25.03.2014). An atmospheric correction was performed for all the images, the details of image correction are described by Guzinski et al. (2014).

In order to enhance spectral information and improve predictive capability to infer on SOC, the spectral indices were generated (Table.3) (Rivero et al., 2009, Mulder et al., 2011). The following spectral vegetation indices were derived for both satellite images depending on their spectral bands: (i) from the Landsat8 image – Mid-Infrared Index (MidIR), Moisture Stress Index (MSI), Normalized Difference Vegetation Index (NDVI), Normalized Difference Vegetation Green Index (NDVI green), Normalized Difference Water Index (NDWI), Reduced Simple Ratio (RSR), Simple Ratio (SR), Transformed Vegetation Index (TVI), Enhanced Vegetation Index (EVI), Land Surface Temperature (LST), Emissivity (Emiss) and Leaf Area Index (Lai); (ii) from the SPOT image – MSI, NDVI, NDVI green, NDWI, RSR, SR, EVI and TVI. The formulas used to derive indices are shown in Table.3. The details of estimation for LST, Emiss and Lai were described in Guzinski et al. (2014). ArcGIS 10.2 was used to extract all the spectral indices and reflectance values to the
sampling locations from the images, and relate the extracted values to the corresponding SOC values.

2.4 Ancillary environmental data

This study used plant available water (PAW), elevation, soil maps, geology, land use and landscape types as predictors of SOC in the selected watershed in Denmark. The PAW was considered as one the most important environmental factors for SOC (Vasques et al., 2012, Xiong et al., 2014). The Danish national PAW map was derived from a Danish soil property map (Greve et al., 2007) via pedotransfer function. Elevation of the soil surface was derived from LiDAR (light detection and ranging) technology where the LiDAR points were interpolated using triangulation method and a fine-resolution digital elevation model (DEM) (grid spacing 1.6 m x 1.6 m) was generated by National Survey and Cadastre of the Danish Ministry of Environment in 2011. The DEM was further resampled to a grid size of 30m. Before resampling, DEM was processed to remove the artificial sinks and peaks of 50 cm to ensure proper delineation of drainage networks.

Two soil maps, namely soil class map and soil texture map, were used. The soil class map consists of major FAO-UNESCO soil groups in the study area, e.g., Podzols, Luvisols, etc. derived by Adhikari et al. (2014) using decision tree modeling. The geology map shows the geological origin of soil material at 1m depth, whereas the landscape map describes different landform types such as terminal moraine or glacio-fluvial plains, etc. in Denmark based on the geo-morphological features. Land use and land cover map used in this study is based on CORINE data adjusted for Denmark (Stjernholm and Kjeldgaard, 2004). Areas covered with forest, crop lands, grass meadow etc. could be extracted from this map. The last three maps were originally vector-based and were rasterized to a grid of 30 m for use in this study. More details of these ancillary environmental data can be found in Adhikari et al. (2013).
2.5 Regression modeling

All the models were developed using Cubist regression, and executed using the R statistical software with the Cubist package (Kuhn et al., 2014). The coefficient of determination ($R^2$) and Root mean square error (RMSE) were used for assessment of model performance. We considered that the best model with the highest $R^2$ and the lowest RMSE value.

Cubist is a powerful data-mining tool for generating rule-based predictive models from data, which is developed on an earlier version of C4.5 and M5 model tree (Quinlan, 1993). A tree is grown where the terminal leaves contain linear regression models. Also, there are intermediate linear models at each step of the tree. All the models are based on the variables used in previous splits. A prediction is made using linear regression model at each terminal node of the tree. At last, the tree is reduced to a set of rules, which initially are paths from the top of the tree to the bottom, and then the linear model will be adjusted and simplified to reduce absolute error (Quinlan, 1992, Minasny and McBratney, 2008). Generally, Cubist regression model produces good prediction capability, while also being easy to understand and interpret (Minasny and McBratney, 2008). This data-mining tool can also apply boosting called committees. The details of boosting procedure can be found in Quinlan (1992). Furthermore, Cubist model also gives the usage of each variable, and then we could easily detect the importance of each variable based on this statistics. In the present study, we applied “trainControl” function (R package: Caret (Kuhn, 2012)) with 10 times iteration of 10-fold cross validation to find optimal number of committees.

2.6 Upscaling point spectral feature to image spectral feature and dataset definition

A Cubist model was developed on the Vis-NIR laboratory spectral against SOC, the calibration sample set was based on all 328 samples, according to attribute usage of each variable from output of the model, and the most important laboratory spectral wavelength/feature was selected.
Afterwards, a kriging map based on this important spectral wavelength/feature was generated using Empirical Bayesian kriging (ArcGIS 10.2). This map (hereafter referred to as estimated spectral map) was converted to raster with 30m resolution, then estimated spectral value were extracted to all 328 sampling points for further data analysis. The details of upscaling procedure can be found in the flow chart in Figure.2.

In the spatial modeling step, for each dataset, we randomly selected 75% samples for calibration, withholding the rest 25% samples for independent validation (Figure.1 & Table.2). Firstly, we only calibrated RS data and ancillary environmental data (RSAE) against SOC based on upland & wetland dataset. In theory, the prediction results should be the same or even slightly better if non-significant variables were removed. Thus, the predictors without usage were removed according to output of the model, and then model was recalibrated based on the remaining predictors. Secondly, to improve the prediction results based on the previous step, the estimated spectral map as additional information was combined with the RSAE predictors, and model was recalibrated. We validated both models by the same validation set. Lastly, we calibrated samples from upland (244 samples) and wetland (84 samples) individually (Table.2), and compared prediction results with the model from previous step.

In the end, four models were generated:

Model A: developed by RS data and ancillary environmental data (RSAE), estimated spectra data, model based on total of 328 samples (calibration: 246 samples, validation: 82 samples) from upland and wetland;

Model B: developed by only RSAE data and based on total of 328 samples (calibration: 246 samples, validation: 82 samples) from upland and wetland;
Model C: developed on the RSAE and estimated spectra data based on total of 244 samples (calibration: 183 samples, validation: 61 samples) from upland

Model D: developed on the RSAE and estimated spectra data based on total of 84 samples (calibration: 63 samples, validation: 21 samples) from upland

Two maps were generated from different models. The one was predicted by upland and wetland model which based on the RSAE data and estimated spectral map, another one was predicted by upland and wetland two different models. The predicting maps were produced using R with package : raster (Hijmans and van Etten, 2012), rgdal (Bivand et al., 2013).
3. Results and discussion

3.1 Exploratory data analysis

The general statistics on soil characteristics in the study area of Skjern river catchment are shown in Table.1. The SOC content of 328 samples in entire catchment ranged from 0.7% to 31.6%, which indicates Skjern river catchment represents a heterogeneous study area with high variations in SOC. The descriptive statistics of SOC from three different datasets are shown in Table.2. In these three datasets, all the calibration sets included minimum and maximum SOC value. First, upland and wetland set with all 328 samples, which included all samples from wetland and upland. Second, 244 samples were from upland ranged from 0.7% to 5.5%, the calibration and validation set had almost the same mean value and standard deviation (Table.2). Lastly, 84 samples were obtained from wetland with generally high SOC content. In comparison with the upland data set, the wetland dataset certainly presented a high variation data set with higher mean value and standard deviation (Table.2). Generally, wetland soils showed high SOC content. But in this study, some wetland samples with SOC content only around 1%, and a few samples were located on outlet of catchment with very high organic soil (Figure.1). This is because human activities strongly affected part of this area, the lower part of the Skjern River was straightly channelized, and river valley wetland was reclaimed for agricultural cultivation from 1960s until to 2000 (Pedersen et al., 2007). Furthermore, the wetland map was originally delineated in 1920’s, SOC content could be decomposed if some part of the wetland area with thin peat soil was drained or plowed by agricultural cultivation after 1920.

3.2 Laboratory spectral variable selection

The most important laboratory spectral wavelength/feature was selected by the Cubist model. Wavelength of 1930nm gave the highest attribute usage of 100%, which means all the sub-models
used this variable during the calibration process, and this spectral variable showed very strong signal for SOC and positively contributed to modeling. In general, the absorption features in the NIR region between 1000nm and 2500nm can be due to water, clay minerals and organic matter, meanwhile, the major proportion of soil organic matter is SOC. Organic molecules overtones and combination bands occur in NIR region, which is due to organic matter result from the stretching and bending of NH, CH and CO groups (Bellon-Maurel et al., 2010, Williams and Norris, 2001). The main process by which molecules absorb energy is electronic transitions in atoms from ground to higher energy states. Previous research shows that bands around 1100, 1600, 1700-1800, 1900, 2000, and 2200 to 2400nm are important for SOC calibration (Bellon-Maurel and McBratney, 2011, Brown, 2007, Chang and Laird, 2002, Chang et al., 2001, Du and Zhou, 2009, Gholizadeh et al., 2013, Goge et al., 2012, Rossel and Behrens, 2010). Rossel et al. (2006) also found the strong absorption peak around 1930 nm that is associated with organic compounds such as organic acids and alkyl. Our result confirms that 1930nm is highly related to carboxylic acids (RCOOH) (Stenberg et al., 2010, Rossel and Behrens, 2010). Carboxylic acids exist as dimers, because of strong intermolecular hydrogen bonding. Thus, O-H stretching band, commonly assign to free water, also can be found around 1930 nm.

3.3 Prediction Models

All observed SOC values versus prediction results are shown in Figure. 2. Since we developed models from different dataset based on different predictors, we separately compared results from different models. We only showed predictions results from 3 different models, Model A, B and C:

3.3.1 Model A v.s. Model B

The observed versus predicted values for Model A and Model B are shown in Figure.2 (a) and (b), respectively. These two models were based on the same calibration and validation samples from upland & wetland, but developed by different predictors. The only different predictor was one
estimated spectral map (1930nm). Model A that used estimated spectral map produced better prediction accuracy with an $R^2$ of 0.59 and RMSE of 2.8 % than Model B with an $R^2$ of 0.46 and RMSE of 3.6%. The basic hypothesis of the present study was to test whether this additional information (estimated spectrum) is able to significantly improve the prediction accuracy of spatial modeling. The result proved that our hypothesis was correct. This might be explained by fundamentals of laboratory spectroscopy, which could provide spectral information and directly link to SOC. Moreover, all soil samples applied in this study were composed of 25 sub samples which were taken within a 70m × 70m area; soil samples represented each site’s soil properties very well. It means each soil laboratory spectrum provided soil property information not only for single sampling point, but also represented neighboring area, which could suitably incorporate with RS images with 30 meters resolution.

To the best of our knowledge, no articles have been found in literature deriving laboratory spectrum from point to image and combing with RS image for spatial modeling. In fact, Cubist model selected more than 10 spectral variables with over 90% attribute usage. Interestingly, we only applied the most important laboratory spectrum (1930nm) for spatial modeling instead of all high usage spectral variables. This was because estimated spectral map was derived from point spectrum value using Empirical Bayesian kriging, which means all the values from image are estimated rather than measured values. Thus, we did not apply all the important laboratory spectra for spatial modeling. Lastly, the final predicting model was developed on 28 predictors, which included only one estimated spectral map (1930nm) and another 27 RSAE predictors.

### 3.3.2 Model A v.s. Model C

Model C was developed on the same predictors as Model A, but both models were based on different number of samples. The prediction results from Model c was shown in Figure.2 (c), Model C gave an $R^2$ of 0.66 and RMSE of 0.59 %. Since it was very difficult to directly compare
prediction accuracy for both models, we specifically calculated RMSE value which only contained predicting value of upland samples that were predicted by Model A, the RMSE value gave 0.87%. By contrast, Model C outperformed Model A. This was because the sample variation of SOC in Model A was much higher than Model C, Model A was developed on soil sample with SOC content range from 0.7 to 31.5%, but Model C was developed on soil sample with SOC content range from 0.7 to 5.5%. Additionally, the upland area of Skjern catchment was dominated by agriculture cultivation and mainly covered by crops during the June and July, the soil on upland could be more homogeneous than soil on wetland, in terms of SOC variation. Hence, we considered that individually calibrate samples from upland could avoid some effects from high SOC variation wetland samples during the modeling process.

We also individually calibrated samples only from wetland (Model D), however, the results were not acceptable, it gave very high RMSE value and $R^2$ was close to zero (results are not shown). The reason why this model could not perform well might be explained by vegetation coverage of wetland and the complexities of wetland ecosystem. The wetland area in Skjern catchment was mostly dominated by perennial rye and reed canary grasses, and vegetation coverage in this area were very uniform all the year. The fundamental of using RS indices for soil spatial modeling was to indirectly extract soil information from vegetation. Theoretically, vegetation planted on high SOC soil was assumed to produce more biomass than vegetation grew on poor SOC soil, it means that the RS vegetation indices such as NDVI, EVI value for high SOC site should be relevantly high. However, in present study, all samples from wetland showed very similar value of RS vegetation indices. For instance, in wetland validation sample set, the minimum and maximum value of SOC samples were 1.8% and 20% (Table.2), respectively, but NDVI value in June for these two sites were 0.21 and 0.19, respectively. In this situation, it was very difficult to find linear relationship between SOC content and RS vegetation indices. Consequently, most of validation samples with
high SOC content were underestimated in Model A and Model B (Figure.2 a and b). Instead, in upland validation sample set, the minimum and maximum value of SOC were 0.8% and 5%, respectively, the NDVI value in June for these two samples were 0.41 and 0.75, respectively.

3.4 Prediction maps
Two prediction maps are shown in Figure.3, upland map was predicted by Model C, and upland & wetland map was predicted by Model A. Both maps were developed on the same predictors, but based on different samples. In general, these two maps have very similar spatial distribution, north-east part and middle of the catchment with relatively higher SOC content, because north-east part of catchment were dominated by forest, and dense of wetland was located on middle of catchment. On the other hand, western part of catchment showed relatively lower SOC content, which might be explained by geological history. In this area, the soil was dominated by Aeolian deposits, mainly with coarse sand, most of organic matter and clay was removed by wind erosion during 16 and 18 century (H. B. Madsen et al., 1992). Upland & wetland map was generally showed slightly higher value of SOC content when compare to upland map, except wetland area (Figure.3). This was because model based on different sample range, upland map was predicted by Model C based on SOC range from 0.7% to 5.5%, but upland and wetland map was predicted by Model A based on SOC range from 0.7% to 31.6% (Table.2). Therefore, Model A was affected by high SOC content samples more than Model c.

3.5 Cubist regression rules and Variable importance
The final values used for the upland & wetland Cubist model and upland Cubist model were 20 and 10 committees, respectively. The lowest value of RMSE by leave group out cross validation was used to select optimal model. Soil class map and land use were selected to set the rule conditions in most of the rules, for instance, soil class map was used for rule setting in 46% of the models.
Furthermore, Cubist model only chooses a number of predictors for use in prediction, for the linear model building, estimated spectral map, PAW, the vegetation indices such as, NDVI, LST, EVI, etc. provided very high attribute usage (Table.4). Cubist regression tree produces amount of liner equation for SOC predicting based on different rules, which was very easy to understand and interpret the model. Here we showed an example of the rule used for SOC prediction in upland and wetland model.

Rule 1: [237 cases, mean 3.29, range 0.65 to 31.56, est err 1.33]

If

FAO map in (2, 4, 5, 6, 7, 8)

PAW > 14.30

then

SOC= 13.6151138 - 0.3815 LST (June) + 0.3723 LST (July) - 3.315 NDVI (June) + 3.166 NDVI (July) – 1.28 SR (July) + 6.92 NDWI (June) + 0.369 PAW + 2.54 SR (June) + 13.9 1930nm - 3.39 NDVIgreen (July) + 2.24 EVI (July) + 1.92 NDVIgreen (June)

Table 4 shows top 5 predictors selected by Cubist model and their attribute usage. Plant available water, estimated spectral map provided very high attribute usage among the all predictors for both models. Predictor of PAW showed 100% attribute usage, this predictor was derived from Danish national soil organic matter map, which was highly correlated with SOC content. The estimated spectral map ranked as the second best of predictor; this single spectral wavelength/feature strongly supported modeling process by directly giving SOC information from soil sample. In this study, different vegetation indices were applied to accentuate a particular vegetation property. Vegetation indices such as NDVI, EVI, and LST also showed high attribute usage in this study, this result was
consistent with previous soil studies (Fernandez-Buces et al., 2006, Mulder et al., 2011, Rivero et al., 2009, Sumfleth and Duttmann, 2008). In present study, the study area was dominated by non-irrigated agriculture land, and mainly covered by crops. Furthermore, all the satellite images used in this study were from June and July, which were the most productive months (season) for crop growth. The NDVI is chlorophyll sensitive and indicates the greenness cover on the land surface that reflects crop growth characteristics and indirectly gives information of soil properties at specific site. Therefore, this index revealed a good relationship with vegetation productivity or biomass status (McKenzie and Ryan, 1999, Rondeaux et al., 1996, Yang and Chen, 2004). The results also showed that the EVI gave a very high attribute usage for upland & wetland model. The EVI is more responsive to canopy structural variations, including LAI, canopy type, plant physiognomy, and canopy architecture compared to NDVI, these two indices could complement each other and extract canopy biophysical parameters (Huete et al., 2002, Gao et al., 2000). Interesting, we found LST provided high attribute usage for upland model, but did not for upland & wetland model. The LST was for estimating surface energy fluxes, being more closely related to the physiological activities of leaves in vegetated area (Weng et al., 2004). Upland ecosystem release electromagnetic energy back to atmosphere much faster than wetland, it means that LST index is more sensitive for upland than wetland ecosystem. There was no raw RS spectral band selected by both Cubist models, because the study area was mainly covered by crops, grass and forest during June and July. Thus, we suggest that raw RS spectral band should be applied to bare soil, which could detect soil property information directly from soil rather than vegetation.
4. Conclusion

The present study upscaled laboratory spectral wavelength/feature from point to image, and combined with multi-spectral images to predict SOC at regional scale. The results showed that upland and wetland model developed on RS data, ancillary environmental data and one estimated laboratory spectra image yielded lower RMSE of 2.8% and higher $R^2$ of 0.59 than the model developed from only RS data and ancillary environmental data (RMSE = 3.6%, $R^2$=0.46). Additionally, separately calibrate wetland and upland dataset could improve prediction results only for upland by narrowing down variation of soil samples, the complexity of wetland ecosystem makes prediction of SOC difficult using RS data in Skjern river catchment. Vegetation indices such as NDVI, EVI were very important predictors for SOC spatial modeling when there is vegetation cover. Empirical Bayesian kriging can be an alternative method to upscale laboratory spectra from point to image for digital soil mapping, but estimated spectral map only can be used for additional information incorporated with other RS images or other ancillary environmental data, due to the laboratory spectral image is estimated. We concluded that laboratory spectra as additional information significantly improves prediction accuracy of SOC compared to only applying RS data.

5. Acknowledgements

This study was funded by China Scholarship Council (CSC) and the HOBE project. We also thank Radoslaw Guzinski provided Landsat 8 images, Danish Geodata Agency provided SPOT5 images.


KUHN, M. 2012. The caret Package.


Figure 1 Study area of Skjern river catchment and spatial distribution of measured SOC
Figure 2 Flow chart of point spectral wavelength/feature upscale to image spectrum and calibration procedure.
Figure.3 Cubist model predicted topsoil organic carbon value versus reference values using different predictors based on the different dataset.

(a) Cubist model developed using ancillary environmental data, RS data, and one estimated spectra (1930nm), and based on upland & wetland dataset (calibration: 246 samples, validation: 82 samples),

(b) Cubist model developed only using ancillary environmental data and RS data based on the same dataset as model (a).

(c) Cubist model developed using ancillary environmental data, RS data, and one estimated spectra (1930nm), model based on only upland dataset (calibration: 183 samples, validation: 61 samples).
Figure 4: Prediction maps of SOC for upland and the combination of upland and wetland, all models were developed by ancillary environmental data, RS data and estimated spectra (1930nm).
Table 1: Descriptive statistics on soil characteristics at the Study area of Skjern river catchment

<table>
<thead>
<tr>
<th>Soil properties</th>
<th>Min</th>
<th>Max</th>
<th>Mean</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>SOC (n=328)</td>
<td>0.7</td>
<td>31.6</td>
<td>3.7</td>
<td>4.1</td>
</tr>
<tr>
<td>Clay (n=296)</td>
<td>1.1</td>
<td>14.3</td>
<td>4.2</td>
<td>1.9</td>
</tr>
<tr>
<td>Silt (n=296)</td>
<td>0.5</td>
<td>22.2</td>
<td>4.9</td>
<td>2.5</td>
</tr>
<tr>
<td>Sand (n=296)</td>
<td>63.5</td>
<td>97.3</td>
<td>91</td>
<td>3.9</td>
</tr>
</tbody>
</table>

SD, standard deviation; n, number of samples
Table 2: Descriptive statistics of SOC from 3 different datasets

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Min</th>
<th>Max</th>
<th>Mean</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>SOC (%)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Upland &amp; wetland calibration set (n=246)</td>
<td>0.7</td>
<td>31.6</td>
<td>3.5</td>
<td>3.9</td>
</tr>
<tr>
<td>Upland &amp; wetland validation set (n=82)</td>
<td>0.8</td>
<td>26</td>
<td>4.2</td>
<td>4.6</td>
</tr>
<tr>
<td>Upland calibration set (n=183)</td>
<td>0.7</td>
<td>5.5</td>
<td>2.5</td>
<td>0.94</td>
</tr>
<tr>
<td>Upland validation set (n=61)</td>
<td>0.8</td>
<td>5</td>
<td>2.3</td>
<td>0.98</td>
</tr>
<tr>
<td>Wetland calibration set (n=63)</td>
<td>1</td>
<td>31.6</td>
<td>7.5</td>
<td>7.3</td>
</tr>
<tr>
<td>Wetland validation set (n=21)</td>
<td>1.8</td>
<td>20</td>
<td>6.3</td>
<td>5.1</td>
</tr>
</tbody>
</table>

SD, standard deviation; n, number of samples
Table. 3 Vegetation Indices Description

<table>
<thead>
<tr>
<th>Indices¹</th>
<th>Equation</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>NDVI</td>
<td>( \frac{\text{NIR} - \text{Red}}{\text{NIR} + \text{Red}} )</td>
<td>(Rouse Jr et al., 1974)</td>
</tr>
<tr>
<td>MSI</td>
<td>( \frac{\text{MidIR}}{\text{NIR}} )</td>
<td>(Rock et al., 1986)</td>
</tr>
<tr>
<td>NDWI</td>
<td>( \frac{\text{NIR} - \text{SWIR}}{\text{NIR} + \text{SWIR}} )</td>
<td>(Gao et al., 2000)</td>
</tr>
<tr>
<td>RSR</td>
<td>( \frac{\text{NIR}}{\text{Red}} \left(1 - \frac{\text{SWIR} - \text{SWIR}<em>{\text{min}}}{\text{SWIR}</em>{\text{max}} - \text{SWIR}_{\text{min}}}\right) )</td>
<td>(Brown et al., 2000)</td>
</tr>
<tr>
<td>TVI</td>
<td>( \left(\frac{\text{NIR} - \text{Red}}{\text{NIR} + \text{Red}} + 0.5\right)^{1/2} \times 100 )</td>
<td>(Nellis and Briggs, 1992)</td>
</tr>
<tr>
<td>SR</td>
<td>( \frac{\text{NIR}}{\text{RED}} )</td>
<td>(Cohen, 1991)</td>
</tr>
<tr>
<td>EVIᵇ</td>
<td>( \frac{\text{NIR} - \text{Red}}{\text{NIR} + C_1 \text{Red} - C_2 \text{Blue} + L} (1 + L) )</td>
<td>(Huete et al., 1997)</td>
</tr>
</tbody>
</table>

\( \text{NDVI Green} \) \( \frac{\text{NIR} - \text{Green}}{\text{NIR} + \text{Green}} \) (Gitelson et al., 1996)

¹ NDVI, normalized differential vegetation index; EVI, Enhanced vegetation Index; MSI, Moisture Stress Index; NDWI, normalized difference water index; MSI: moisture stress index; NDWI: normalized difference water index; RSR: reduced simple ratio; SR: Simple Ratio

ᵇ Empirical parameters for EVI: \( C_1=6; \ C_2=7.5; \ G=2.5; \ L=1 \)
Table 4 Predictors selected by Cubist model and their attribute usage

<table>
<thead>
<tr>
<th>Variables for Upland &amp; Wetland</th>
<th>Attribute usage (Top5)</th>
<th>Variables for Upland</th>
<th>Attribute usage (Top5)</th>
</tr>
</thead>
<tbody>
<tr>
<td>PAW&lt;sup&gt;a&lt;/sup&gt;</td>
<td>100%</td>
<td>PAW</td>
<td>100%</td>
</tr>
<tr>
<td>1930nm</td>
<td>96%</td>
<td>1930nm</td>
<td>98%</td>
</tr>
<tr>
<td>Landsat8 NDVI_July</td>
<td>88%</td>
<td>Landsat8 NDVI_July</td>
<td>90%</td>
</tr>
<tr>
<td>Landsat8 NDVI_June</td>
<td>85%</td>
<td>Landsat8 NDVI_June</td>
<td>87%</td>
</tr>
<tr>
<td>Landsat8 EVI_July</td>
<td>80%</td>
<td>Landsat8 LST_July</td>
<td>84%</td>
</tr>
</tbody>
</table>

<sup>a</sup> PAW, plant available water