Mid-Infrared spectroscopy for soil and terrain analysis

Thesis · March 2015

DOI: 10.13140/RG.2.1.2231.7688

1 author:

Alexandre Wadoux
The University of Sydney
46 PUBLICATIONS **479** CITATIONS

SEE PROFILE

Mid-Infrared spectroscopy for soil and terrain analysis

Master thesis

Alexandre Wadoux

Tuebingen, 2015

Mid-Infrared spectroscopy for soil and terrain analysis

By

Alexandre Wadoux
(Identification number: 3793657)

Submitted in Partial Fulfilment of the Requirements
for the Degree of Master of Science (M.Sc.)
in Physical Geography.
Department of Geosciences
Physical Geography and Soil Science

Supervised by
Dr. Leonardo Ramirez-Lopez
Swiss Federal Institute of Technology, Zuerich

Assessed by
Dr. Karsten Schmidt
Prof. Dr. Thomas Scholten
Eberhard Karls University, Tuebingen

Eberhard Karls University
Tuebingen, Germany
March 25, 2015

Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

This Master thesis has been conducted within the framework of the BMBF funded project YANGTZE GEO (www.yangtze-geo.de) in the sub-project "soil erosion" at the University of Tuebingen and in collaboration with the Swiss Federal Institute of Technology in Zurich for the spectroscopy.

Declaration Of Authorship

I do solemnly declare that I have written the presented research thesis by myself without undue help from a second person others and without using such tools other than that specified.

Where I have used thoughts from external sources, directly or indirectly, published or unpublished, this is always clearly attributed. In the selection and evaluation of research materials, I have received support services from following institutions: Swiss Federal Institute of Technology (ETH Zuerich), University of Tuebingen.

The presented intellectual work of this research thesis is my own. In particular, I have not taken any help of any qualified consultant.

I have not directly nor indirectly received any monetary benefit from third parties in connection to this research thesis.

Furthermore, I certify that this research thesis or any part of it has not been previously submitted for a degree or any other qualification at the University of Tuebingen or any other institution in Germany or abroad.

Date: 05.02.2015
Name: Alexandre Wadoux
Matriculation No.: 3793657
E-Mail: alexandre.wadoux@uni-tuebingen.de
Submission Date: 30.03.2015

Abstract

Diffuse Reflectance Spectroscopy is a fast, cost-efficient and non-destructive method well suited to derive a large number of soil properties from a single scan. The ability of infrared spectroscopy for predicting soil components has already been widely described in numerous studies. Especially for the Mid-Infrared (MIR) regions (400-4000 cm^{-1}), common calibration methods allow the prediction of various soil properties with high accuracy. In this respect, the bending or stretching vibration at a precise wavelength allows qualitative diagnostic on the soil components without any coupled chemical analysis. Recent studies show the importance of the soil information summarized into a few wavelengths of the spectrum. However, the use for soil monitoring remains unexplored. In this thesis, we propose a method to identify quickly which soil attributes are influenced by various and easily obtainable environmental secondary information. We demonstrate first that a few bands in our spectrum can represent most of the variability of a target soil property. In consequences, through the study of spectra-terrain relationships, we highlight the link existing between terrain derivative and the information content of the spectra. We implemented three calibration methods: Partial Least Square Regression (PLSR), Cubist and Support Vector Machine (SVM) and then used a robust linear model to define the precision and significance of the modelled terrain attributes (as independent variable) to the bands of the spectra. The 140 samples were collected from a heterogeneous 4,2 km^2 catchment area in Hubei province in central China, and scanned in the mid-infrared range using an Alpha FT.IR Spectrometer. In this work, the spectra is first linked to laboratory measured soil properties to calibrate our models and then linked to 34 terrain attributes derived from a digital elevation model with a resolution of 25m. The multivariate relationship is qualitatively interpreted based on terrain spectrograms derived from the fitted models. The results show that (i) the three calibration methods tested are efficient for predicting soil texture and organic matter; therefore our spectral library contains information about soil properties (ii) soil mineralogy and particularly clay minerals are strongly linked to the bedrock properties as well as to elevation. In contrast, soil organic matter is difficult to interpret, showing reasonable correlation to vegetation coverage and slope only for aromatic and alkyl groups. The method appears to be suitable to investigate soil-landscape relationships through Mid-Infrared spectroscopy and without any prior laboratory analysis.

Acknowledgement

I would like to thank Prof. Thomas Scholten for giving me the opportunity to participate to the Yangtze-Geo project at the University of Tuebingen. It was a great experience to attend conferences, meetings, to go to the field and meet so many interesting people.

I am deeply grateful to Dr. Leonardo Ramirez-Lopez for the supervision of this thesis. I appreciate all his contribution of time, ideas and expertise in soil spectroscopy to make my master thesis experience stimulating and productive.

All this work wouldn't have been possible without the company of Felix Stumpf during the endless field campaigns in Badong. Thank you for the nice discussions about sampling design and pedometric.

I am grateful to Dr.Karsten Schmidt and Dr.Thorsten Behrens for introducing me to the field of Digital Soil Mapping and for their advices on this thesis.

This thesis was founded by the German Ministry of Education and Research (BMBF, Grant No. 03 G 0827A) for the Germano-Sino research collaboration Yangtze-Geo.

I would like to thank the Institute of Agricultural Sciences at the ETH Zuerich for letting me use their Spectrometer and specially Dr. Lee for his time in calibrating the device.

I would like to express my gratitude to my parents for support and help during my study as well as to my friends David and Jordi without whom I wouldn't have spent a such great time in Tuebingen.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Problem statment

Predicting and managing soil efficiently became of major importance during the last decades (Hartemink and McBratney, 2008; Baveye, 2006). Soil is a complex system whose processes and structure are still fully discussed (Rossel and Chen, 2011). Historically, we base our understanding of soil system on expert field knowledge, as well as time- and cost-expensive soil chemical and physical laboratory analysis (Viscarra Rossel et al., 2006; Merry and Janik, 2001). The use of soil maps began to appear in the 1940s. They are based on deterministic components from qualitative analysis of morphological processes (Dalrymple and Conacher, 1968). Jenny et al. (1941) describes the soil forming factors with the state equation $S = f(cl, o, r, p, t)$ where cl = climate, o = organisms, r = relief, p = parent material and t = time. These factors have been largely used to delineate soil boundaries without spatial correlation. Soil is classified into defined classes where this is assumed that they are similar and with the hypothesis that their characteristics rely on the environmental covariates that affect their formation (Beckett, 1978). However, the high soil spatial variability encourages soon in using stochastic techniques with continuous soil attributes (Webster, 1977; Burgess and Webster, 1980). Such maps describe the variation of soil properties from reliable sampling and laboratory analysis. The goal being to extrapolate values to an unknown location given the environmental factors. These techniques are now commonly used but are highly variable given the sampling design and the number of samples that can be gathered (Hengl et al., 2004; Brus et al.,

2011; Minasny and McBratney, 2006). However, these methods are limited by the computing possibilities and by the expensive soil survey and chemical analysis (McBratney et al., 2003). Over the last decades, the improvement of technologies has partially filled the gap of cost and time-consuming soil data retrieval (Jarmer et al., 2009). Behind the higher computing and technological improvement, research focused on the estimation of soil properties from "non-invasive" techniques such as remote (Yurui et al., 2008) or proximal soil sensing (Viscarra Rossel et al., 2011). In this respect, the quantitative estimation of soil properties from the Vis (400-700 nm), near-IR (700-2500 nm) and mid-IR (2500-25,000 nm) regions developed quickly (Ben-Dor and Banin, 1995). They are rapid, accurate and cheap to estimate various soil properties (related to organic or non-organic components) from a single scan (Viscarra Rossel et al., 2006; Dunn et al., 2002). Determining soil characteristics with the combined VIS-NIR-MIR usually provides better results in comparison to the use of one particular region. In contrast to MIR, the use of vis-NIR spectroscopy in soil science has been largely documented. From a merely pragmatic perspective, the main differences between both techniques rely on the accuracy of the soil models. While soil mid-IR models are usually more accurate, the sample preparation required by this technique can be more complex than when vis-NIR is used. The mid-IR is however more accurate and robust to estimate soil properties (Stenberg et al., 2010; Rossel and Behrens, 2010). The reason behind the good performance of mid-IR in soil modelling, relies on the fact that several fundamental molecular vibrations occur in the mid-IR when only their overtones and combination are identified in the vis-NIR (Rossel and Behrens, 2010). Recent studies show the importance of the coupled qualitative and quantitative analysis of the spectral variability (Demattê et al., 2004; Vågen et al., 2006; McBratney et al., 2006; Shibusawa et al., 2003; Rossel and Chen, 2011; Viscarra Rossel et al., 2011; Demattê et al., 2004). Particularly for MIR, the molecular bending or stretching vibration at a precise wavelength allows diagnostics on the soil components without any coupled chemical analysis. The information contained in the MIR spectra has been studied by several authors for example for the study of the specific carbon content in soil (Calderón et al., 2013) or soil characteristics (Terhoeven-Urselmans et al., 2010; Reeves, 2012). Despite MIR has proven (to be) efficient in soil research, this technique is still not operational specially in key fields of soil science where large amounts of soil information is necessary (e.g. soil survey programs). For example, a systematic integration of MIR spectroscopy in Digital Soil Mapping (DSM)

would result in better sampling designs, higher spatial and temporal resolutions of the soil information, better mapping accuracies, and therefore a better understanding of the soil processes in the landscape.

1.2 Objective and research questions

The major aim of this research is to investigate the information content of a MIR soil spectroscopy by the prediction of soil properties and by studying the relationships with terrain attributes. We conduct both quantitative (statistical modelling) and qualitative (wavelength analysis) to answer a number of specific research questions as listed below:

- Is the information contained of our MIR spectral library enough to predict soil properties with accuracy ? The general concept that diffuse reflectance spectroscopy is suitable to derive soil properties is a long established fact. However, the particular relationship existing between a given soil property and a spectral library is complex because it varies with space, time, sampling design and accuracy of the spectrometer, especially as in our case with a very heterogeneous study area. The number of research articles on this question allows comparison, although MIR spectroscopy is not the main spectral region used for predicting soil components.

- Does the MIR spectrum have specific bands or regions "good predictors" for a given soil property ? The NIR is generally difficult to interpret because of overtones and band combinations when fundamental molecular vibration occur in the MIR range. Thus, the estimation of a target property, even in small quantity, should be the summary of important wavelength that will represent most of its variability. The identification of such bands would allow conducting future qualitative interpretation of MIR spectroscopy.

- Are specific wavelengths free of overlap for a given soil property ? We assume that a few band can represent most of the variability of a target soil property. The aim is to know whether these specific bands are only related to our soil property or if different soil properties overlap in the same wavelengths. By a literature review on the target soil properties,

we will try to define the bands where the absorption is due to a unique molecular vibration.

- Is it possible to use the spectra to highlight the relationships between soil properties and terrain ? In agreement with the fact that our spectral library contain soil information, and without any laboratory analysis, can we extract the influence of the relief and parent material factors on the soil formation ? Here, the goal is to derive information on the soil-landscape relationships from the MIR spectra. What are the particularly terrain attributes that vary according to the soil property ?.
- Is it conceivable to use soil MIR spectroscopy for soil survey ? Considering no prior information, in which extent is the spectroscopy useful to describe the soil-landscape relationships ?

1.3 Scope and layout of the thesis

The aim of this research is to link MIR spectroscopy and terrain modelling for assessing the spatial variation of soils without prior knowledge of specific soil properties which are of interest for soil surveys. This is known that topography plays a key role on soil formation, characteristics and evolution. But the link between infrared spectra and terrain remains widely unexplored. We base our methodology on the assumption that spectra contain soil information and thus the relationships between terrain and spectra can reveal information about the soil-terrain interdependences. The idea behind would be to go to an unknown study area, with no prior information about soil. By taking soil samples, scanning them in the MIR range and linking them to terrain derivatives, we obtain an overview of the influence of the terrain for some given soil attributes. For DSM purpose, the gain of information would be considerable, given that we could know if the terrain covariates are able to correctly map a given soil property. This study tries to answer the specific research questions in two steps. First, the study focuses in demonstrating the use of MIR spectroscopy in soil description by modelling a set of soil properties. The second part of the study explores the relationships between soil MIR data and terrain. In this respect, detailed terrain models of the spectral responses of each wave number are calibrated. In this sense, we indirectly investigate the terrain influence on the soil attributes. This has been seen as

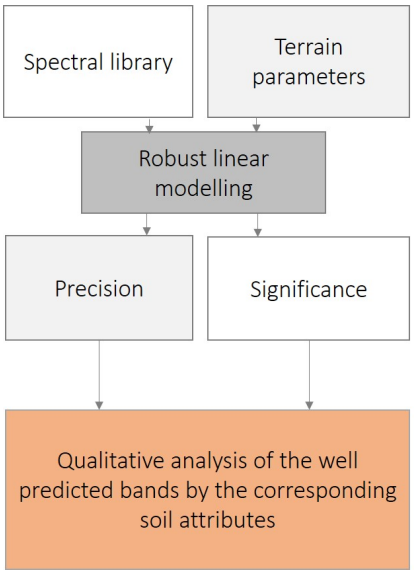


Figure 1.1: Methodology for the study of the soil-terrain relationships

a new method to describe the soil-terrain relationships quickly and without any prior information on the soil. This part is described in Figure 1.1.

Chapter 2

Research area

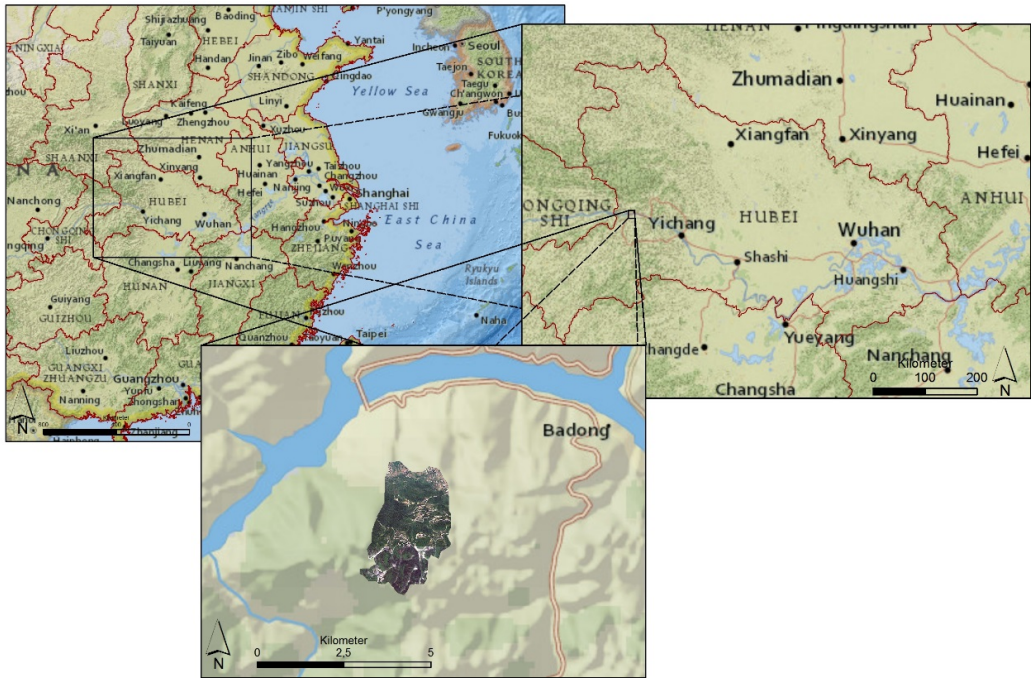


Figure 2.1: Study area of Upper Badong, background maps are provided by Olson et al. (2001)

The research was conducted in the western Hubei province, central China, in a catchment area which is located in the upper part of the city of Badong, 74 kilometres upstream the Three Gorges Dam on the Yangtze River (see

Figure 2.1).The study area is 4.2km^2 (3.2 km in the north-south and 1.8 km west-east) and centred on the geographical coordinates 31^{o}1'24"N and 110^{o}20'35"E.. The area was selected for the capability of correctly represent the variability of the soil properties as well as its unique outlet for a relatively small catchment that allows a better representativeness of the soil variability with only a few samples.

Geologically, the area is rather homogeneous. Sedimentary rocks characterize most of the area, mainly dolomite with silt and limestone formed in the middle and lower Jurassic. The lower part contains fuchsia clayed siltstone and clayey microcrystalline limestone formed in the middle Jurassic. The southern part exhibits grey microcrystalline dolomite and limestone from the lower Jurassic (Jiang J., 2012). The upper Badong is topographically heterogeneous according to the elevations range between 469m to 1 483m above sea level and with an average of 1 053m. The majority of the area is North oriented (72%) with a slope range from 0^{o}C to 53^{o}C with an average of 26^{o}C. The study area is covered by a subtropical monsoon climate with hot and humid summers and cool and dry winters. According to the Köppen-Geiger climate schemes, Badong is classified as Cwa (Rubel and Kottek, 2010). As reported by the China Meteorological Administration (Hong-Yu, 2005), the mean annual temperature is 12.9^{o}C and the annual precipitation mean is 1067mm.

The area is mainly covered by woodland (81%), cropland (15%) and infrastructures (4%). The northern part contains most of the fields when the southern part is almost entirely covered by forest. Woods are composed of coniferous and leafy trees like oaks. Cropland contain soybean, corn and cabbage with a seasonal crop rotation. Constructed area are mainly small farm located in the northern part, with a few small pork industrial buildings in the middle and upper area.

Chapter 3

Materials and methods

3.1 Multivariate statistics for the calibration of MIR spectroscopy

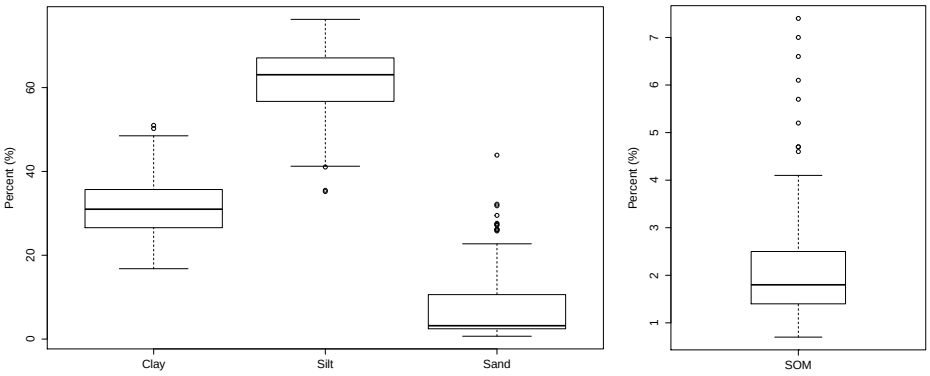
3.1.1 Soil sampling and pre-treatment

3.1.1.1 Soil sampling

The samples were collected within the framework of the Germano-Sino Yangtze-geo project, founded by the German Ministry of Education and Research. Three field campaigns were organized. The first one in June 2013, the second in May 2014 and the last one in November 2014. In total of 140 topsoil samples (0-0.2m) were collected (0.5kg). Given the accessibility problems in the area, most of the samples are collected from the lower part where the landuse is characterized by fields. However, we produced a map containing 30 spatial strata in order to cover spatially the whole area. The idea, developed in Walvoort et al. (2010a), is to distribute sample points evenly over the whole area. The stratification has been performed with the spcosa package in R (Walvoort et al., 2010b).

3.1.1.2 Chemical analysis

All the 140 samples were submitted to conventional soil analysis for Sand, Silt, Clay and SOM contents:
The samples were air dried in oven during 24 hours at 45^{o}C and sieved with a 2 millimetres filter. These first two steps were conducted at the Faculty of



(a) (b)

Figure 3.1: Boxplot of Clay, Silt, Sand (a) and SOM (b)

Engineering of the University of Wuhan, China. All the other analysis were then made at the laboratory of Soil Science and Geoecology at the University of Tübingen, among which the milling to a size of $0.63\mu m$. The Grain size was known by using a combination of the sieve- and the pipette methods of the sieved samples by DIN 19683-1 and DIN 19683-2 (Durner and Nieder, 2006). The sieve method determines the distribution of the grain size by filtering the samples through filters of 0.625mm, 0.200mm, 0.125mm, 65μm and 20μm, from a coarser to a finer filter. Finer particles, which are cohesive, need a moist sieving, we therefore employed the pipette method. This technique is based on the relationship between particle grain size and the velocity in which the particles sink into a fluid. With a sedimentation analysis, we can estimate the size of the particle lower than 0.063 mm.
The Soil Organic Matter has been calculated by a function of the carbonate ($CaCO_3$) and the total carbon estimation. For $CaCO_3$, the method used is according to Scheibler. The Scheibler is based on the reaction of carbonates into carbon dioxide using hydrochloric acid. The estimation is measured as a function of the CO_2 volume that is emitted from the reaction. In contrast, we used a CNS Vario EL III device for the total carbon content estimation in our soil sample.

The data extracted are given in Table 3.1 and drawn in Figure 3.1. The distribution of silt seems to be skewed (Asymmetry = -0.86) which means

that the probability of finding sand values is higher when looking to the higher values. It can be compared to the three other distribution where SOM and sand have positive values, the cumulative probability is more constant. The kurtosis coefficient gives indication about the shape of the probability distribution. High kurtosis value is often linked to low standard deviation. For Clay, we get a negative value which means that most of the values are near the centre of the probability distribution.

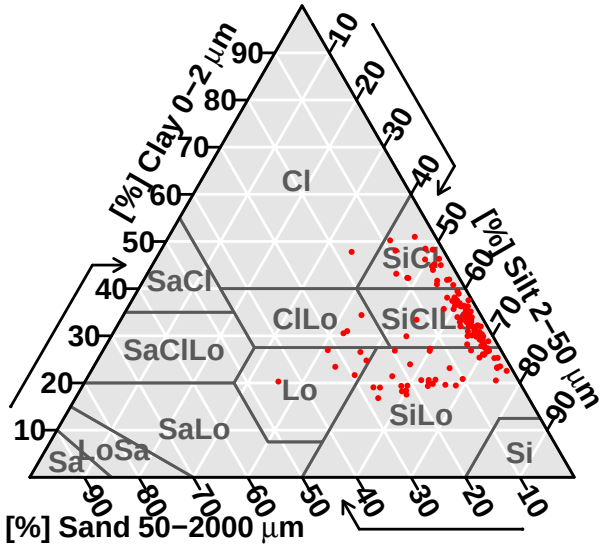


Figure 3.2: Soil texture with USDA classification in background

The soil textural triangle exhibits a soils mainly composed by silt soil classes (USDA Soil classification). Most of the samples are classified as Silt-clay-loam. The two others main classes are Silt-loam and Silty-clay. Silt soils are generally the most fertile ones. They are composed of minerals like quartz and fine organic particles. They retain large amount of moisture and improve the drainage as the proportion of clay decreases. In the USDA classification, the Silty-clay-loam soil is composed of 27 to 40% of clay and less than 20% of sand.

	Soil Organic Matter	Clay	Silt	Sand
Min	0.7	16.79	35.2	0.68
1^{st}Qu.	1.4	26.58	56.71	2.46
Median	1.8	30.99	63.07	3.19
Mean	2.17	31.27	61.06	7.66
3^{rd}Qu.	2.5	35.68	67.08	10.61
Max	7.4	51	76.28	43.88
SD*	1.28	7.86	8.51	8.65
Asymmetry	1.78	0.36	-0.86	1.69
Kurtosis	3.49	-0.34	0.33	2.16

Table 3.1: Descriptive statistics for soil texture and Soil Organic Matter

3.1.1.3 Optical measurement (spectral scanning)

Each samples were air dried, 2mm sieved and ball milled as for the laboratory experiments. Plants and stone are therefore excluded. The background measurement is first measured with an empty sample compartment through 32 co-added repetitions. The soil sample is then placed into the sample cup which is 0.5 cm in diameter and 0.5 cm in depth. A spatula was used to smooth the surface of the cup and to provide a maximum light reflection and a minimal signal to noise ratio (Mouazen et al., 2005). The scans were done using an Fourier Transform Infrared (FTIR) Vertex 70 Spectrometer (Bruker, Germany). This device is equipped with an air cooled IR source and ZnSe optics for controlling air humidity. We used a MIR-KBr beamsplitter with a spectral range of 7500-370cm^{-1}. We scanned all the samples three times in the MIR-range (570-5500cm^{-1}) for absorbance with 64 co-added scans. The background measurement were done every 3 samples or every ten minutes. Three references samples were scanned every 4 hours to have a look to the influence of the environmental factors like air humidity. After all the scans, we calculated the Standard Deviation (SD) of the spectrum. The samples in which the spectrum has SD higher than 1.5 were re-scanned. The result is given in absorbance per unit with a spectral resolution of 1.4cm^{-1} in Figure 3.3.

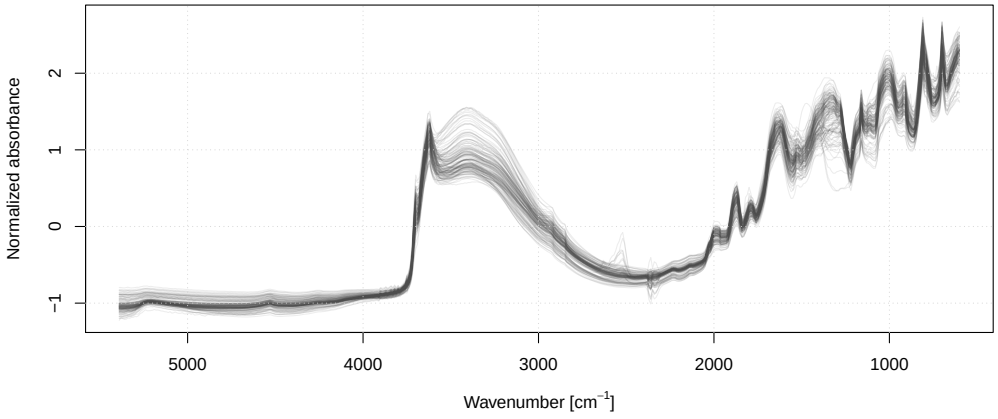


Figure 3.3: Scanned samples in the MIR range

3.1.2 Calibration methods

3.1.2.1 Partial least square regression (PLSR)

The dataset was split into calibration and validation sets by the k-means sampling algorithm (80% and 20%)(Stevens and Ramirez-Lopez, 2013; Næs, 1987). The PLSR was used to calibrate MIR models that allow the prediction of soil attributes from the MIR data. This multivariate regression method aims to deal with many and correlated predictor variables and with only a few observation (Mevik and Cederkvist, 2004). PLSR has been widely used in applied sciences (Höskuldsson, 1988) and in chemistry (Martens and Martens, 2001). The model aims to find a latent structure by projecting the predicted variables and the observed variables to a new space (Tenenhaus, 1998). PLSR model first find new variables called latent variables $t_a (a = 1, 2, \ldots, A)$, from a linear combination between the original variable x_k with the weight coefficients $w^*_{ka} (a = 1, 2, \ldots A)$,(Wold et al., 2001):

$$t_{ia} = \sum_k W^*_{ka} X_{ik} \tag{3.1}$$

The predictor variables X are summarized by the loadings p_{ak} multiplied by the scores t_{ia} with e_{ik} residuals:

$$X_{ik} = \sum_a t_{ia} p_{ak} + e_{ik} \tag{3.2}$$

In the case of several Y, the matrix of the observed variables is estimated like the last equation:

$$y_{im} = \sum_a u_{ia} c_{am} + g_{im} \tag{3.3}$$

where u_{ia} are the Y-scores, c_{am} are the weight and g_{im} the residuals. The X-scores are predictors for the observed values so we obtain:

$$y_{im} = \sum_a c_{ma} t_{ia} + f_{im} \tag{3.4}$$

Thus, the PLSR model is expressed as a multiple linear equation:

$$y_{im} = \sum_a c_{am} \sum_k w^*_{ka} x_{ik} + f_{im} = \sum_k b_{mk} x_{ik} + f_{im} \tag{3.5}$$

The coefficient of regression for each predictor variable is denoted as b_{mk} and calculated as follow:

$$b_{mk} = \sum_a c_{ma} w^*_{ka} \tag{3.6}$$

The PLSR model is validated using a Leave One Out (LOO) cross validation technique. This method allows to understand how accurate the predictive model will perform in practice. It can be run internally with the training dataset which constructed the model or externally with new and independent variables. For the training set validation, the LOOCV excludes one observation and the model is recalculated with the other observations. The cross validation fitted values are calculated as the difference between the predicted and the original observation and run as many time as the number of observation in the dataset. The model is validated with an external dataset; the square root of the prediction (RMSEP) and the coefficient of determination (R^2) are calculated from the following equations (Mevik and Cederkvist, 2004)

$$R^2 = \frac{\sum_{i=1}^{n} (\hat{y}_i - \bar{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y}_i)^2} \tag{3.7}$$

$$RMSEP = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (\hat{y}_i - y_i)^2} \tag{3.8}$$

3.1.2.2 Cubist

For the calibration of the MIR spectra, we secondly used a rule-based regression model (Holmes et al., 1999) called Cubist (Minasny and McBratney, 2008). Cubist is based on a decision tree having a linear model at the leaf nodes (Holmes et al., 1999). This is an enhancement of preliminary method combining instance-based and model-based learning called M5 or model tree (Quinlan, 1993b) and M5' (Wang and Witten, 1997). Model tree is a technique that constructs a piecewise function. Instead of using discrete data at the leaves of a decision tree, a linear model is used. As for the model tree, M5'chooses factor that minimize the variation at each branch of the tree rather than maximizing the information gain (Quinlan et al., 1992). The target is to maximize the error reduction:

$$\Delta error = sd(T) - \sum_i \frac{|T_i|}{|T|} \times sd(T_i) \tag{3.9}$$

where T is the training dataset in which every potential test is computed and T_i the number of T cases with i tests. After having computed every combination, the model chooses the one that will minimize this error. In the original model of decision tree, Breiman et al. (1984) uses the variance of the absolute deviation to choose between the tests. The linear model fitted at each leaf uses standard regression techniques simplified in order to minimize the error. A smoothing procedure is determined by the variance and the covariance between two sets of residuals (Kuhn and Johnson, 2013) as follow:

$$PV(S) = \frac{n_i \times PV(S_i) + k + M(S)}{n_i + k} \tag{3.10}$$

Where S_i is the followed bran of the node S, n_i is the number of training cases, $PV(S_i)$ the predicted value at S_i and $M(S)$ the value given at S. A boosting scheme called committees can also be implemented in the model tree (Kuhn et al., 2012). The first tree is calculated using the procedure described above. The outcome result is "analysed" by a sequence of test trees. If the values is over-estimated by the model, the following value is adjusted downward to compensate. The final value is an average of the smoothed value at each node. The procedure, as described in Quinlan (1993a):

$$y^*_{(m)} = y - (\hat{y}_{(m-1)} - y \tag{3.11}$$

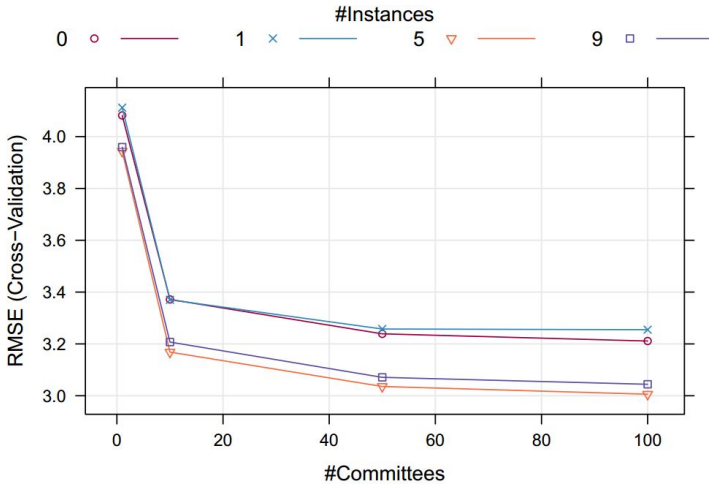


Figure 3.4: Performance of Cubist using the tuning parameters committees and nearest neighbours. After Kuhn et al. (2012)

where $\hat{y}$ is the predictive model.
At the end of the model, Cubist adjusts the prediction using nearest-neighbours in the training set. With the predicted sample, cubist finds its nearest neighbour and smooth with the surrounded values. Let $\hat{y}$ be a new predicted sample and w_l be the weight of distance between the new sample $\hat{y}$ and the neighbours in the training set. For the model prediction $\hat{t}_l$ and t_l as observed outcome for a training set neighbour we obtain:

$$\frac{1}{N}\sum_{l=1}^{K}[t_l + (\hat{y} - \hat{t}_i)] \tag{3.12}$$

Cubist method is becoming more and more employed in the calibration of Vis-NIR and MIR spectroscopy (Miklos et al., 2010; Minasny et al., 2009) because of its capability to deal with missing data or to handle non-linear relationships. As for PLSR, the accuracy of the model is calculated with RMSEP and R^2, see equations (3.7) and (3.8).

3.1.2.3 Support vector machine (SVM)

Support Vector Machine (SVM) is a supervised machine learning method used for classification and regression and first developed by Cortes and Vapnik (1995). This method uses the concept of "dimension superiority" (Li

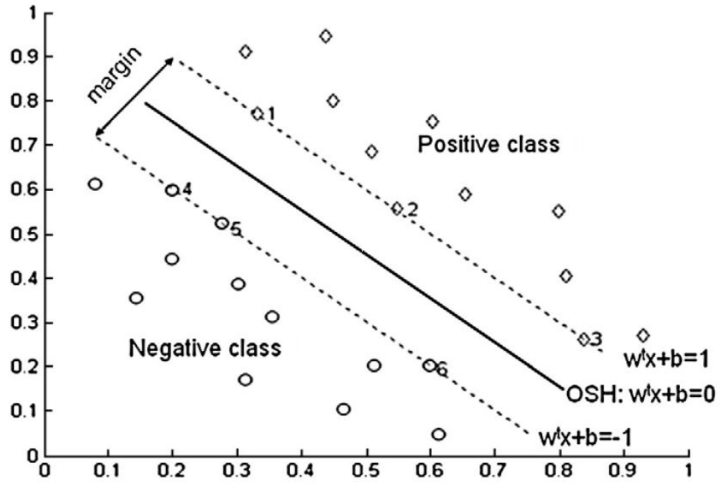


Figure 3.5: Linearly separable case for SVC. The hyperplane is expressed by optimizing the margins

et al., 2009) to increase the information content of the database and separate the data. SVM draws inseparable samples in a low dimensional space to a higher dimensional space where a linear or a kernel hyperplane can solve the separability problem. For linearly separable data, the question that arise is about the best hyperplane to efficiently separate the data. Support Vector Classification (SVC) introduces the concept of margin. The model selects the best separator line so that it maximizes the margins to the nearest samples (Figure 3.5). The model locates the best line by maximizing (Ivanciuc, 2007):

$$\frac{2}{||w||} \tag{3.13}$$

with the constrain:

$$(w^t x_i + b) y_i \geq 1 \tag{3.14}$$

where w is the normalized weight vector and b the bias for the hyperplane. For non-linearly separable cases, two techniques are introduced:

- The soft margin technique proposes a penalizing factor C for the observations inside the margins of the hyperplane. The constrain is then expressed as follow:

$$(w^t x_i + b) y_i \geq 1 - \xi, \xi \geq 0 \text{ for } i = 1, 2, ..., N \tag{3.15}$$

where ξ_i is the variable measured to be away from the margins. The construction of the hyperplane is then done with the above constrain by minimizing:

$$\frac{1}{2}||w|| + C\sum_{i}^{N}\xi_i \tag{3.16}$$

Kernel methods can help to classify non-linearly separable data into a

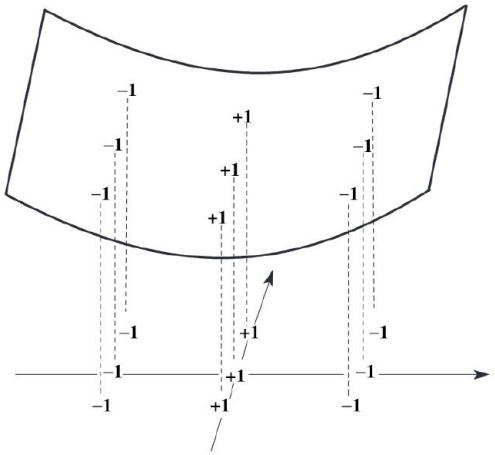


Figure 3.6: Hyperplane transformation for (x, y, x^2)

high dimensionality hyperplane. Kernel based methods map the input space in a feature space by using the function $\varphi(x)$ (see Figure 3.6).

- The kernel methods calculate the inner product in the original input space and avoid therefore distortion due to the projection in a high (or infinite) feature space. The following kernel based methods are commonly used:
  The linear kernel

$$K(x_i, x_j) = x_i \cdot x_j \tag{3.17}$$

  The polynomial kernel

$$K(x_i, x_j) = (1 + x_i \cdot x_j)^2 \tag{3.18}$$

  The Gaussian Radial Basis Function kernal (RBF Kernel)

$$K(x_i, x_j) = \exp\left(-\frac{||x - y||^2}{2\sigma^2}\right) \tag{3.19}$$

The sigmoid kernel

$$K(x_i, x_j) = \tanh ax_i \cdot x_i + b \tag{3.20}$$

The SVMC can be extended to a regression problem (Vapnik and Vapnik, 1998) by using a ε-insensitive loss function (see Figure 3.7). SVR identifies

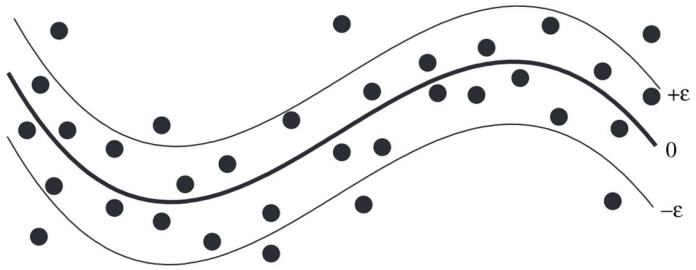


Figure 3.7: A radius ε is fitted to the data to maximize the margins

a function $f(x)$ so that all the x have a maximum deviation ε from the value y. SVR creates a model analogous to SVC by using soft-margins with the introduction of slack variables ξ and where C control the penalty linked with a deviation higher than ε.

3.2 Robust modelling of the spectra-terrain relationships

3.2.1 Data preprocessing

3.2.1.1 Terrain parameters extraction

A Digital Elevation Model (DEM) with a spatial resolution of 25m was derived from a topographic map using a 25" by 25" grid (ESRI, 2011). The covariates were obtained by digital terrain analysis (Hutchinson, 1989) based on the DEM. A total of 29 terrain attributes were derived by ArcGIS Desktop 10, SAGA GIS and the statistical software R (SAGA, 2013; R et al., 2012; ESRI, 2011). Furthermore, a layer containing landuse information classified according to the Chinese landuse classification system (Liu et al., 2005; Di Gregorio, 2005) were available, as well as a geological map. All layers were

Table 3.2: Terrain derivatives used as independant variables

Topographic indice	Description	Method	Software
Northing	Orientation of the pixel	Bivand et al. (2008)	ArcInfo, R
Easting	Orientation of the pixel	Bivand et al. (2008)	ArcInfo, R
Wetness Index	Steady state wetness index	Moore et al. (1993)	SAGA
Slope	Angle of inclination of the soil surface from the horizontal	Burrough et al. (1998)	ArcInfo
Slope, Maximum slope	Maximum angle to or from the centre cell	Travis et al. (1975)	SAGA
Slope, Maximum triangle	Steepest downwards slope on triangular facets formed in a 3 x 3 pixel window	Tarboton (1997)	SAGA
Slope length	Length dimension of the slope	Olaya and Conrad (2008)	SAGA
Catchment area	Hydrological contribution to the area	Quinn et al. (1991)	SAGA
Catchment area, parallel	Contributing area for specific basin	Quinn et al. (1991)	SAGA
Catchment area, recursive	Processing of cells for flow accumulation	Quinn et al. (1991)	SAGA
Plan curvature	The perpendicular to the direction of the maximum slope	Moore et al. (1993)	ArcInfo
Profile curvature	The parallel to the direction of the maximum slope	Moore et al. (1993)	ArcInfo
Combined curvature	Combination of profile and planform curvature	Moore et al. (1993)	ArcInfo
Protection Index	Analyse the surrounded protection of the relief	Olaya and Conrad (2008)	SAGA
Flow accumulation	Cumulative count of the number of pixels that naturally drain into outlets	Jenson and Domingue (1988)	ArcInfo
Flow direction	Determines into which neighbouring pixel any water in a central pixel will flow naturally	Jenson and Domingue (1988)	ArcInfo
Mass balance index	Identify different soil-related landforms	Möller et al. (2008)	SAGA
Distance to channel	Vertical distance to a channel network base level	Olaya and Conrad (2008)	SAGA
Overland flow distance	Calculates overland flow distance to a channel network	Olaya and Conrad (2008); Freeman (1991)	SAGA
Vertical flow distance	The vertical distance of a cell above the nearest water body	Olaya and Conrad (2008)	SAGA
Horizontal flow distance	The horizontal distance of a cell above the nearest water body	Olaya and Conrad (2008)	SAGA
Altitude above channel	Difference between the DEM and a surface interpolated from the channel network	Olaya and Conrad (2008)	SAGA
Digital Elevation Model	Height above sea level	Olaya and Conrad (2008)	SAGA
Convergence index	Terrain structure of channel and ridges	Koethe and Lehmeier (1996)	SAGA
LS-Factor	Calculation of slope length (LS) factor as used by the Universal Soil Loss Equation (USLE)	Olaya and Conrad (2008); Moore et al. (1993); Desmet and Govers (1996)	SAGA
Ruggedness index	Amount of elevation difference between adjacent cells of a digital elevation grid	Riley (1999)	SAGA
Position index	Identifies position between channel and ridges	Guisan et al. (1999)	SAGA
Terrain ruggedness	Topographic heterogeneity	Riley (1999)	SAGA

standardized to fit a spatial resolution of 25m. Description of the derived terrain and references are summarized in Table 3.2.

3.2.1.2 Normalization of the terrain attributes

The 34 terrain attributes were imported in ArcMap 10.1 and the tool Extract by points were used to obtain the terrain attributes values at each sample location. The distribution of the values were analysed for normal distribution using Quantile-Quantile (QQ) plots. The QQ plots are used to check whether the distribution of a dataset validates the distributional assumption by computing a theoretical expected value for each data point and to check if the points follow a straight line. In the case the statistical population of the terrain attribute seems to follow a non-linear pattern, the dataset is transformed to come as close as possible to a normal linear distribution $y = x$. We used logarithmic $y = log(x)$, mean $y = mean(x)$, and exponential $y = exp(x)$ functions to correct non-linear datasets. After checking the assumption of normal distribution, we checked the colinearity of the covariates. In order to do this, a correlation matrix was computed. The terrain attributes with a correlation higher than 0.8 were extracted and the colinearity was avoided by deleting the terrain parameters with the lowest value of colinearity.

3.2.1.3 Mid-Infrared data pre-treatment

The acquisition of the MIR spectra is described in Part 3.1.1.3. Before to implement the robust linear model, we tested different spatial transformations for the spectra.

- Savitzky-Golay local polynomial regression (Savitzky and Golay, 1964); item the 1^{st} and the 2^{nd} differentiation order with a polynomial order $p = 4$ implemented in the R package prospectr (Stevens and Ramirez-Lopez, 2013);
- coversion of the absorbed spectra to reflected spectra by using the function $y = \exp(-x)$;
- the Standard Normal Variate (SNV) -detrend to remove the scatter effect of each spectra individually (Barnes et al., 1989);
- the Savitzky-Golay 1st differentiation order spectra.

The best results for the robust linear modelling are found for the detrent spectra.

3.2.2 Robust linear model

3.2.2.1 Basic concepts

Linear regression models provide a tool to summarize the relationships in the data. These linear methods has been shown as extremely sensitive to minor deviation in their assumptions (Huber, 2011). These assumptions, among which linearity and additivity, statistical independence, homoscedasticity, and normality may cause havoc in the model and lead to erroneous conclusions. The normality assumption can be violated by the presence of outliers. They can be object that have a different property or they can fake values produced by the data generation process (Filzmoser et al., 2009) which is likely in terrain processing. Handling outliers in regression analysis may avoid distort estimates. Robust methods can handle regression analysis with influential cases. They are insensitive to small deviations from the assumption (Huber, 2011) and deal with a wide range of probability distributions including non-normal. Robust methods fit the bulk of the data. If the dataset contains only a few small outliers, the models gives approximately the same result as a classical regression model. In contrast, in high-dimensional multivariate situation with large outliers and when the typical regression model shows erroneous results, the robust method provides reliable information (Maronna et al., 2006). Robust methods have been first developed in the 1960s and in the early 1970s with the work of Tukey (1960, 1962), Huber et al. (1964); Huber (1967) and Hampel (1971, 1974). Since they have high computationally requirements, robust statistics have experienced only recently a growth of publication number.
Formulated by Mosteller and Tukey (1977) and reported in Andersen (2008), the robust estimator satisfies two conditions: the robustness of validity refers to the estimator that should have an optimal efficiency at the model and reflects the resistance of the estimator to outliers. In other words, the precision should not be impacted by a change in the data. Robustness of efficiency depicts the efficiency of the estimator under a wide range of circumstances.
The concepts underlying the robustness of an estimator can be defined as follow:

- The breakdown-point defines the degree of robustness of an estimate

with one or several outliers (Yohai, 1987). This concept has been first introduced by Hampel (1971) and then developed by Donoho (1982) and Donoho and Huber (1983). The breakdown-point (BDP) gives a measure of the resistance of an estimator (Andersen, 2008) by providing the largest amount of unusual observations that the data can contain without damaging the estimate. Huber (1984) gives the maximum bias that can be caused by outliers as follow:

$$bias(\varepsilon; T, Z) = \sup |T(Z') - T(Z)| \tag{3.21}$$

where $T = (t_1 \dots t_n)$ is an estimator and $T(Z)$ is its values at the sample Z. We assume that all ε-corrupted samples are contained in Z'.In the case of the number of ε samples infinitely grows in T, the estimator "breaks down" and fails to fit the bulk of the data. The breakdown point is more generally defined as:

$$BDP(T, Z) = \inf(\varepsilon | bias(\varepsilon; T, Z) = \infty) \tag{3.22}$$

The breakdown point varies strongly between the estimators. The dataset contains generally 10% of unusual observation that differ from the volume of the data (Hampel et al., 1986) and suggests therefore having a breakdown point of at least .1. However, several estimates fail to have a high breakdown point, such the M-estimates introduced by Huber (1984) with a BDP of 0 or the GM-estimates developed by Yohai and Maronna (1979) which has a BDP which tends to 0. In response to the low breakdown point of the M-estimator, the S-estimators (Rousseeuw and Yohai, 1984) and the MM-estimators (Yohai, 1987) were developed to implement a high breakdown point of 30-50% and 50% respectively. They are now the most commonly used methods and are largely inspired by the M-estimation procedure as described is the following parts.

- The measure of location is the measure of a position in the distribution with a value . Common measures of location for a distribution are the centre, the mean, the median and the interquartile mean. Linear regression estimates a conditional mean of a dependant variable with one or several predictors. The conditional mean is not robust and can therefore be strongly influenced by a few outliers and misrepresent the

real relationships between the data. The arithmetic mean is commonly used in statistic to estimate the central tendency of the distribution and denoted $\bar{x}$.

$$\bar{x} = \frac{x_1 + x_2 + x_3 + ... + x_n}{n} \tag{3.23}$$

Considering the low resistance to the mean to outliers or extreme values, the mean is less efficient than many other measures of centre (Andersen, 2008). Hampel (1974) suggests to use a robust measure of location rather than a two-step procedure consisting in removing the outliers before calculating the mean. The trimmed mean is a robust measure of the centre that discard parts of the samples, typically 10% at the high and low end. It is calculated as follow:

$$\bar{x}_{tk} = \frac{1}{n - 2k} \sum_{i=k+1}^{n-k} x_i \tag{3.24}$$

where k is the number of times that the trimmed-mean is computed, n is the number of observation and x_i is the ith arranged in increasing order:

$$x_1 \leq x_2 \leq ... \leq x_n \tag{3.25}$$

M-estimation are also commonly used to robustly estimate the location (Huber, 2011). It includes various estimators based on the idea of the maximum likelihood. Main robust regression estimates are derived from this measure of location and scale such as M-estimates, GM-estimates, S-estimates and MM-estimates. In the M- estimator, the Maximum Likelihood is generalized. The M-estimator of location μ for an assumed distribution T_n is the solution of the equation:

$$\sum_{i=1}^{n} \psi \left(\frac{x_i - \mu}{c_s} \right) = 0 \tag{3.26}$$

where S is the measure of the scale with c as tuning constant and ψ is the score function. M-estimate is differently estimated by using another function that gives less weight to the outliers. The most popular are the Huber weight and the biweight functions as shown in Figure 3.8.

- The measure of scale characterizes the spread or the variability of a dataset. The idea, that is the same as the robust measure of location,

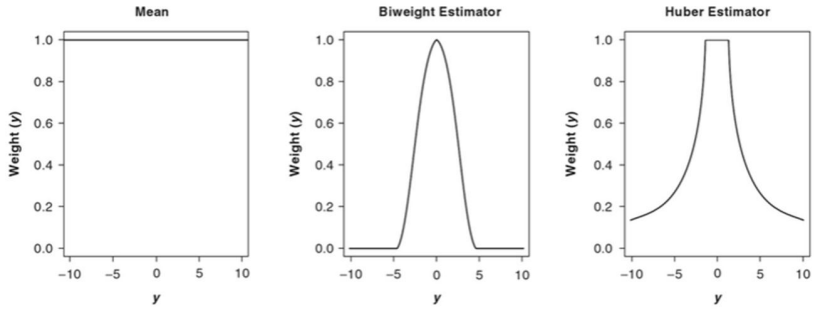


Figure 3.8: M-estimator functions compared to the mean

is to give less weight to unusual observations when traditional measure of scale are sensitive to outliers. The two common measure are first the variance:

$$s^2 = \sum_{i=1}^{N} \left(\frac{Y_i - \bar{Y}}{N-1} \right)^2 \tag{3.27}$$

where $\bar{Y}$ is the mean of the data and s^2 represent the arithmetic average of the squared distance from the mean. The second is the standard deviation that is the square root of the variance and defined as follow:

$$s = \sqrt{\sum_{i=1}^{N} \left(\frac{Y_i - \bar{Y}}{N-1} \right)^2} \tag{3.28}$$

The standard deviation is particularly affected by outliers because it calculates the squares of the deviations from the mean, which is also strongly impacted by outliers. The outlier's effects are amplified. To avoid having distortions of the estimator, the Interquartile Range (IQR or Qn) and the Median Absolute Deviation (MDA) are commonly used. The MAD is defined as the median of the absolute deviations from the median (Huber, 2011):

$$MAD_n = med\{|x_i - M_n|\} \tag{3.29}$$

with

$$M_n = med\{x_i\} \tag{3.30}$$

The MAD estimate of scale considers that most of the information is in the tail of the data, therefore MAD exclude the lower and higher

values of the median of the absolute deviations from the data median. The estimate achieves high breakdown point, around .5. The IQR is a measure of scale that have also high breakdown point: 50% of the data are discounted which means that the estimate has a breakdown point of .5. The difference between the .25 and the .75 range produce he interquartile range:

$$QR_q = y_{1-q} - y_q \tag{3.31}$$

where

$$0 < q < .5 \tag{3.32}$$

The SD, MDA and IQR are compared in Figure 3.9. The standard deviation behaves inadequately and shows the impossibility to extract information in presence of outliers. MAD shows better behave and the dispersion of IQR is even lower. As for the measure of location, the M-

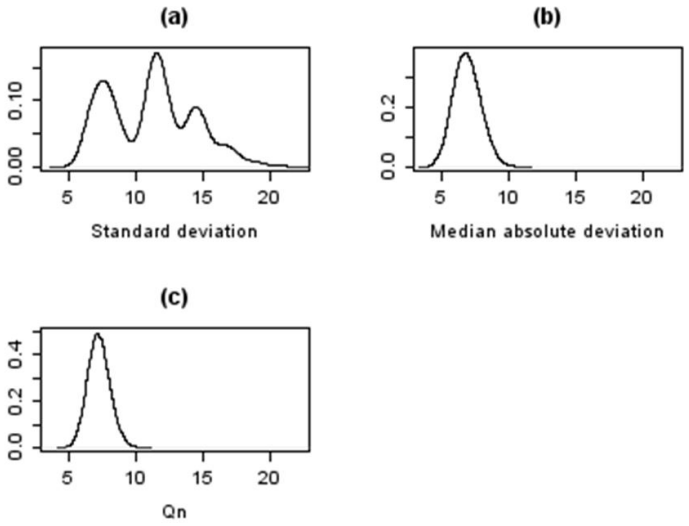


Figure 3.9: Bootstrap distribution of the Standard Deviation, the Median Absolute Variance (MDA) and the Interquartile Range (Qn). Based on Rousseeuw and Croux (1993). The image shows that SD can not be used as measure of scale in presence of outliers.

estimation provides tools to handle outliers using maximum likelihood. The M-estimator of scale is defined as the solution of the equation:

$$\frac{1}{n}\sum_{i=1}^{n}\rho\left(\frac{x_i - \hat{\mu}0}{\sigma}\right) = \kappa \tag{3.33}$$

with

$$0 < \kappa < \rho(\infty) \tag{3.34}$$

3.2.2.2 The MM-estimator

In this study, we used a MM-estimator that "is perhaps now the most commonly employed robust regression technique" (Andersen, 2008). First proposed by Yohai (1987), it combines high breakdown estimator of regression (50%) with good efficiency. The "MM" has been chosen because the estimator needs to compute multiple M-estimator procedures to carry out the computation of the estimator. To reach the final estimates, the MM-estimator makes use of the S-estimation for the high breakdown point (Rousseeuw and Yohai, 1984) and the iteratively reweighted least squares (IRLS) with the M-estimation as follow:
Stage 1: S-estimator is used to fit a highly resistant regression (BDP=50%) with the coefficient $\widetilde{\beta}$ and estimated residuals $r_i(\widetilde{\beta}) = y_i - x_i^T\widetilde{\beta}$
Stage 2: The residuals $\widetilde{\beta}$ of the first estimated regression are used to compute an M-estimate of scale with high BPD (50%) and denoted $s_n = s(r_1\widetilde{\beta}, ..., r_n\widetilde{\beta})$ [1]. The chosen objective function (mainly Huber of biweight) is expressed as ρ_0.
Stage 3: The residuals $\widetilde{\beta}$ from the first stage and the residuals scale s_n in the second stage are used to compute the MM-estimator as a solution of iteration weighted least squares of the M-estimator with a redescending score function:

$$\sum_{i=1}^{n} x_{ij}\psi_1\left(\frac{y_1 - x_i^T\beta}{s_n}\right) = 0 \tag{3.35}$$

where $\psi_1(u)$ is the score function.The first two stages refer to the MM-estimator's high breakdown point when the third one is responsible of the high efficiency (Stuart, 2011; Rousseeuw and Leroy, 2005).
The estimation of the probability for the results to be obtained when the null hypothesis H_o is true is estimated by the p-value. We commonly use the p-value to determine the evidence against H_o; the higher the evidence against H_o, the smaller the the p-value (Wasserman, 2004). It is important to determine the level at which H_o can be rejected, assuming that if the test

[1] Note that the function s_n is here not given. Information in Stuart (2011) and Andersen (2008)

rejects at α, it will reject at level $\alpha^{'} > \alpha$. Hence, the lower level at which H_o is excluded is (Wasserman, 2004):

$$p - \text{value} = \inf\{\alpha : T(X)^n \in R_\alpha\} \tag{3.36}$$

with a rejection region R_α for $\alpha \in (0, 1)$. The significance levels are classified as follow:

$$\begin{gathered} p \leq 0{,}01 \\ 0{,}01 < p \leq 0{,}05 \\ 0{,}05 < p \leq 0{,}1 \\ p > 0{,}01 \end{gathered}$$

For interpretation purposes, the p-value can be accounted as the value at which the observed results for a sample is considered as the results of the relationships between the dependant variable and the independent variables. The statistical measure of the goodness of the robust fit is typically done by using a coefficient of determination (or R^2 value), as for classical linear models. The R^2 indicates the proportion of the total sum of squared explained by the model as described in equation (3.7).

3.2.2.3 Terrain modelling of MIR variables

A model based on the set terrain attributes (L) was calibrated for each wave band in the MIR region. The following equations summarize the concept used:

$$A = a_{k=1}^{d} \tag{3.37}$$

and

$$a_k = \hat{f}(L)_k + \varepsilon_k \tag{3.38}$$

where A represent the d spectral variables present in the MIR region, a_k represent the kth spectral variable, $\hat{f}(L)_k$ is the calibrated function that relates the terrain attributes to the kth spectral variable and ε_k is the error of the kth model. This indicates that the final number of models is equal to the number of spectral variables. Robust linear modelling was used for calibrating the terrain models. These models were validated using leave-group-out cross validation and the R^2 values as well as the RMSE values were obtained to assess their accuracy. In this respect, it was possible to construct

the spectra of the obtained R2s, RMSEs and also the spectra of the p-values (significance) of the terrain attributes along the different wavebands. We selected to show in Figure 4.9 only six terrain attributes that were mainly related to the spectra.

Chapter 4

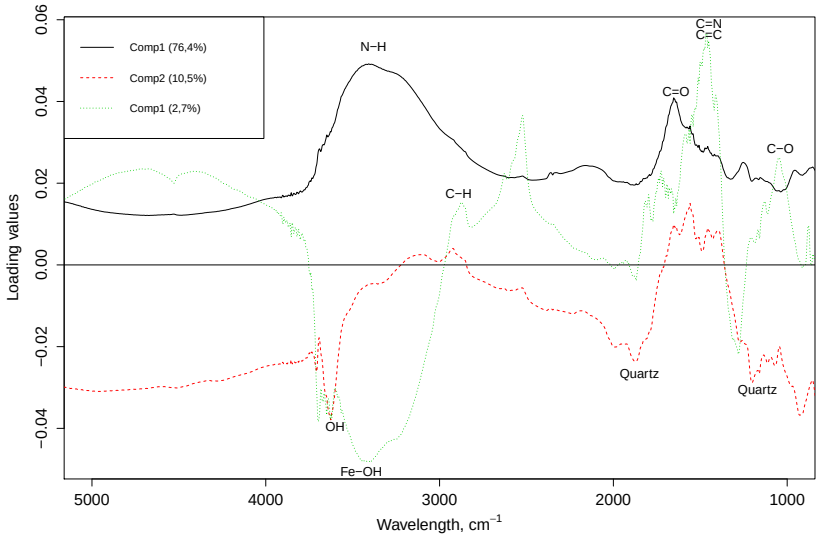
Results and discussion

4.1 Predicting abilities of the mid-infrared spectra

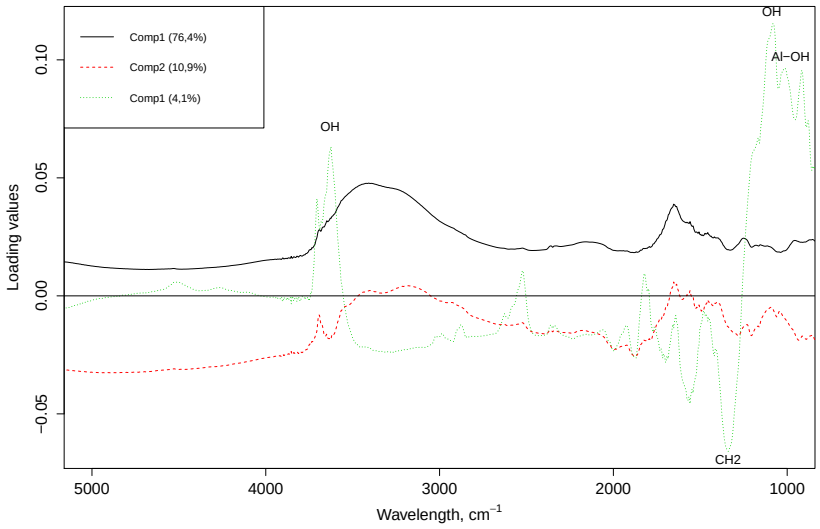
4.1.1 Performance

4.1.1.1 Partial least square regression (PLSR)

The information contained into the MIR data are described in the PLSR model by the eigenvectors (see Figure 4.2). Each spectral band has an eigenvalue which can be interpreted to know what are the spectral region that contribute to the prediction (Haaland and Thomas, 1988; Rossel and Chen, 2011). The bands with a positive eigenvalue will exhibit an absorption of energy (positive relationship) when a negative eigenvalue indicates a reflexion (negative relationship).
Soil organic matter (Figure 4.1a) is positively correlated in its first principal component with the N-H stretch of amines near 3330cm^{-1}, the C=O bond of carboxylic acids at 1730cm^{-1} and the C-O stretching vibration of polysaccharides at 1050cm^{-1}. The third component shows a peak around 1510-1530cm^{-1} which could be likely assigned to the C=N and C=C stretching vibration of amides. In contrast, we observe negative relationships with some sand minerals like quartz near 800 and 1789cm^{-1}, with the Fe-OH bond near 3530cm^{-1} corresponding to the glauconite and especially with the OH stretching vibration of kaolinite at 3620cm^{-1}.

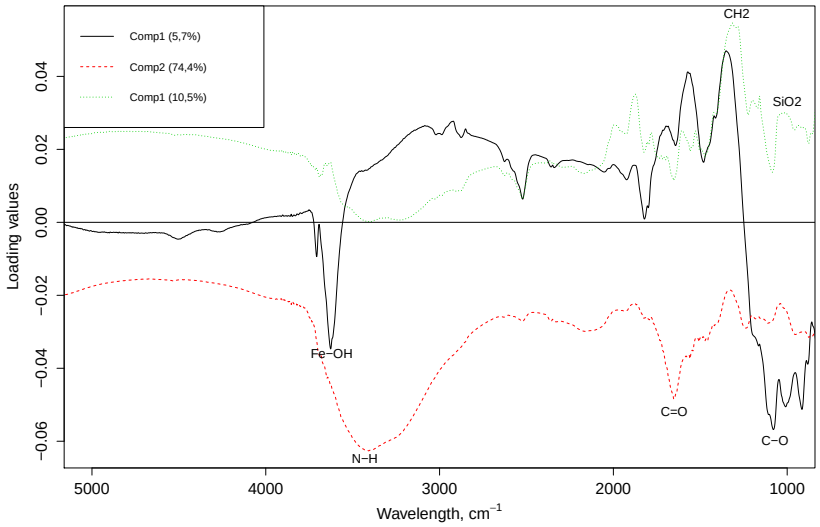


(a) SOM

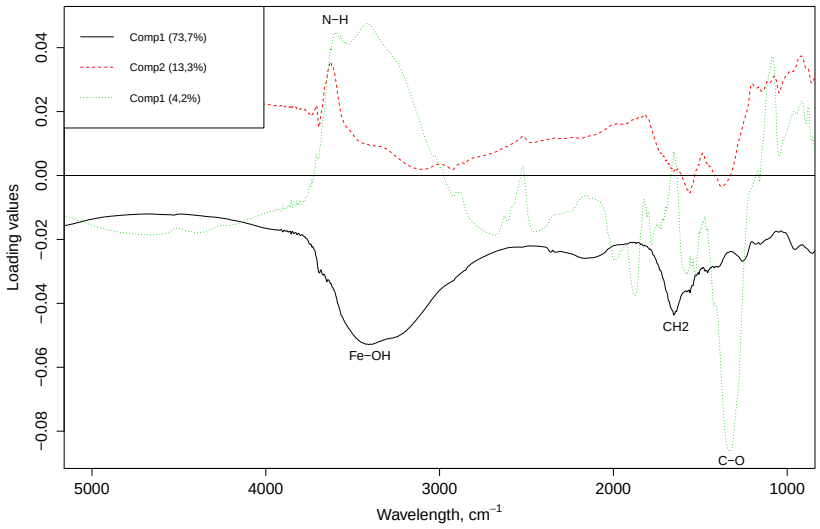


(b) Clay

Figure 4.1: Loadings for the PLSR factors that represent the main spectral variation caused by the absorption for Organic matter and clay. The positive or negative significant peaks are linked to the supposed molecular vibration.



(a) Silt



(b) Sand

Figure 4.2: Loadings for the PLSR factors that represent the main spectral variation caused by the absorption for silt and sand. The positive or negative significant peaks are linked to the supposed molecular vibration.

Clay (Figure 4.1b) has not pronounced peaks in its 1^{st} or 2^{nd} principal component, showing only pronounced positive peaks to the OH stretching vibration of kaolinite at 3620cm^{-1}, to the OH bond due to lattice minerals and to the Al-OH bending vibration of kaolinite near 915cm^{-1} in the 3^{rd} principal component. The third loading has also positive values near 1100cm^{-1} with the OH bending linked to the silicate structures. A strong negative correlation is observed at 1415cm^{-1} with the CH_2, corresponding to an organic matter like alkyl.
Silt (Figure 4.2a), as well as for Clay, shows positive correlation in its third component near bands that are primarily linked to the OH stretching vibration of kaolinite around 3620cm^{-1}. A second peak is observed at 915cm^{-1} and might be due to alkyls or to carbonates and near 1100cm^{-1} with the OH bending linked to the silicate structures. As well as clay, a negative loading is detected at the bands near 1415cm^{-1} with the CH_2. The loadings of the silt prediction is extremely close to the loadings of the clay. It indicates that the main contributors for their prediction are the same.
Sand (Figure 4.2b) shows heterogeneous eigenvalues, with strong positive peaks in its first principal component as, surprisingly, at 1415cm^{-1} with the CH_2 stretching vibration of alkyls. The quartz appears to be strongly related to sand with the SiO2 streching and bending vibration around 800cm^{-1}. Negative peaks are also strong with the fe-OH at 3367cm^{-1}, the N-H stretch at 3330cm^{-1}, the C=O bond at 1730cm^{-1} and the C-O stretching vibration at 1050cm^{-1} corresponding respectively to aromatics, carboxylic acids and to carbohydrates in SOM.

The prediction performance has been assessed by R^2 and RMSE (see 3.1.2.1). We commonly define the R^2 values for prediction of soil properties as very good (>0.81), good (0.61-0.8), fair (0.41-0.6) and poor (<0.41), (Rossel and McBratney, 2008). Table 4.1 shows the results for SOM, Clay Silt and Sand with the internal validation dataset and the external validation of the PLSR model.

The cross validation of the prediction shows generally good results for SOM (R^2=0.98) and for sand (R^2=0.84). Good predictions are obtained for Clay with a value of 0.73 and silt with 0.64. Due to the fact that organic carbon is very well defined in the MIR region, the performance of the MIR models to predict SOM is generally good. Furthermore, the results of the external validation are extremely close to those referenced in Rossel and McBratney (2008) computed from various authors between 1986 and 2006. The R^2 value for the independent validation is in our study of 0.93 for SOM,

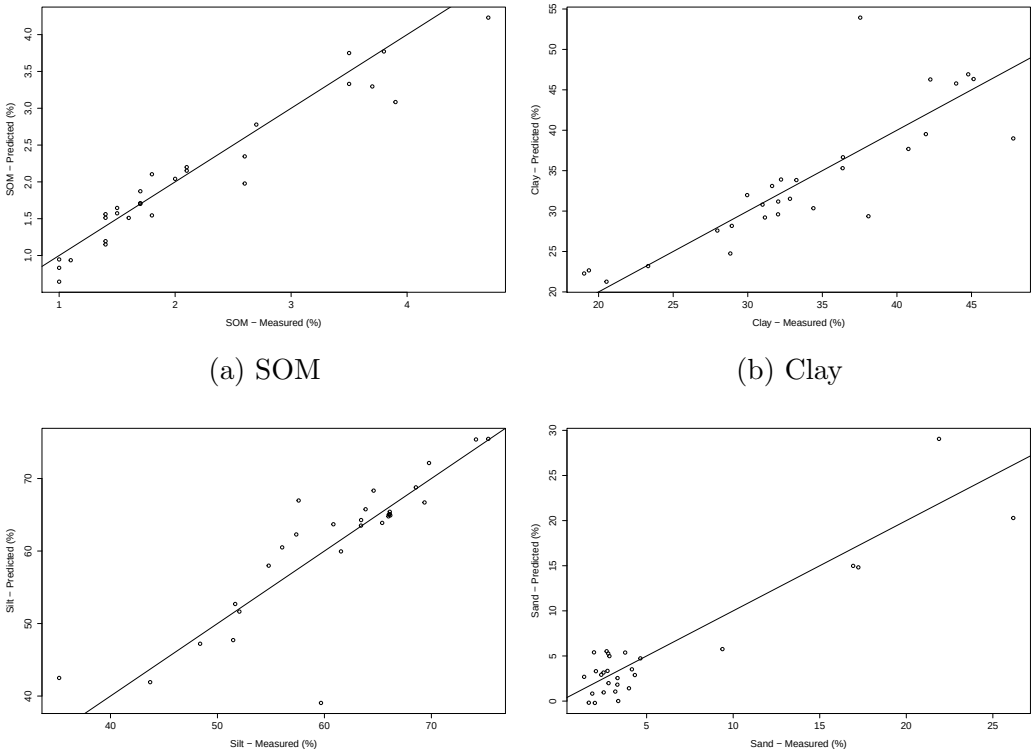
Soil property	Latent variables	Training dataset (n=111)		Independant validation (n=28)	
		R^2	RMSE	R^2	RMSE
SOM in %	14	0.98	0.21	0.93	0.26
Clay in %	8	0.73	4.00	0.73	3.94
Silt in %	7	0.64	4.98	0.82	3.83
Sand in %	10	0.84	3.56	0.80	2.86

Table 4.1: Results of the calibration with the PLSR model for organic matter, clay, silt and sand with the number of selected latent variables. The latent variables are selected by cross-validation as implemented in Mevik and Wehrens (2007).

very close to the R^2 of 0.95 for the mean average prediction between 1986 and 2006. The R^2 for clay is at 0.73 and 0.78 in the study, as close as for sand with a R^2 of 0.80 in comparison with the 0.84 in the paper. Excepted for silt where the difference is high between our R^2 and the study with an R^2 of 0.67. This is however close to the internal validation (R^2=0.64).

The calibration was established using PLSR with all the sample and the model was tested on the training and validation data set with the R^2 and the RMSE given in Table 4.1. In Figure 4.3d are drawn the corresponding linear regressions between the observed (x-axis) and predicted values (y-axis) for the external validation. As described before the relationship seems to be clear, mainly for SOM and Clay. For Clay, Silt and Sand, it seems that a few outliers account for most of the error, given the correct linear distribution of the dataset.

The predicted values of SOM exhibit very good relationships with the observed data that is due to the fundamental bands occurring in the MIR range (Rossel and Behrens, 2010). Clay gives surprisingly low results in prediction even if the MIR range is largely described as good predictor. For Silt, the prediction corresponds to the results described in Rossel and McBratney (2008) when sand has in our study are very high variability (See Table 3.1) and is therefore difficult to estimate properly.



(a) SOM

(b) Clay

(c) Silt

(d) Sand

Figure 4.3: External validation of the model PLSR for organic matter, clay, silt and sand.

Soil property	Committees	Neighbors	Training dataset (n=111)		Independant validation (n=28)	
			R^2	RMSE	R^2	RMSE
SOM in %	50	0	0.95	0.30	0.91	0.31
Clay in %	10	0	0.73	4.10	0.74	3.90
Silt in %	50	5	0.74	4.27	0.73	4.77
Sand in %	10	5	0.78	4.22	0.88	2.28

Table 4.2: Results for the calibration using cubist model. The values for Committees and Neighbors were selected according to the tuning function of the cubist package in R (Kuhn et al., 2014)

4.1.1.2 Cubist

Table 4.2 shows the accuracy of the cubist model on the internal and external validation sets. The model performs generally well for both validation and prediction with R^2 comprises between 0.73 and 0.91. As expected, we get the best R^2 for the internal and external validation of SOM with $R^2 = 0.95$ and $R^2 = 0.91$. This is near the results found in Rossel and McBratney (2008) for the calibration of C (total) with a MIR library ($R^2 = 0.95$) and to the results in Minasny and McBratney (2008) for the calibration of a MIR library with cubist with $R^2 = 0.96$. Surprisingly, the external validation shows also very good results, with $R^2 = 0.91$, that is higher than the two above reference articles. For clay, we obtain lower results than expected for the validation of the training set ($R^2 = 0.73$, RMSE = 4.10) but good prediction of the independent dataset ($R^2 = 0.74$, RMSE = 3.90). For silt content, cubist gives similar prediction than for clay with a $R^2 = 0.74$ and $R^2 = 0.73$ but with higher dispersion of the residuals for the external validation (RMSE = 4.27 and RMSE = 4.77). Sand fraction is well predicted with cubist with an external validation showing better R^2 and RMSE than the internal validation ($R^2 = 0.78$, RMSE = 4.22 and $R^2 = 0.88$, RMSE = 2.28).

A positive part of the cubist model is the facility of interpretation. A set of rules is produced and allows us to have a look to the bands used to build the prediction of the soil property. The model is built from rules that select only a few bands for prediction. Using a model with only one rule — that is, using conventional multiple linear regression analysis. When several rules are decided, this means that the correlation tends to the non-linearity. An example is given with the first rule for the prediction of soil organic matter:

```
Model:
  Rule 1: [105 cases, mean 1.94, range 0.7 to 4.7, est. err 0.26]

    SOM =     3.56    + 140.84 X1542.80535
                      + 121.91 X1471.45061
                      - 113.09 X1479.16463
                      - 49.81 X1558.23341
                      - 48.16 X1560.16191
                      - 48.85 X1540.87685
                      + 46.52 X1577.51847
                      - 31.16 X813.82982
                      + 28.91 X815.75833
                      - 18.42 X1546.66237
                      - 15.29 X1457.95106
                      - 8.28 X860.11399
                      - 7.14 X1446.38002
                      - 6.35 X1322.95559
                      + 4.34 X742.47508
                      + 2.78 X1035.60809
```

Where X refers to the value assigned to the waveband used for the model. The bands used for the prediction of each soil property are visualized in Figure 4.4. They are extracted from the sets of rules and allow finding the wavelengths responsible of the estimation of a particular soil property.

Figure 4.4 shows the wavenumbers taken by cubist to construct the model for SOM, clay, silt and sand. It can be compared to the PLS loadings for interpretation purposes (see Figure 4.2). The wavelengths used for the prediction of SOM are comprise between 870 and 1700cm^{-1} and correspond to the general trend observed in Figure 4.10 in Part 4.2.2.1. Most of them are comprise in the double bond region that is free of overlap for soil organics. The bonds correspond for example at the C=N or C=C stretching bond of amides around 1510-1530cm^{-1} or at the alkyls asymmetric-symmetric doublet with the CH_2 deformation near 1450cm^{-1}. The wavelengths selected for SOM are concentrated into main areas, but with a few other suitable bands at 742cm^{-1}, 860cm^{-1} and 1035cm^{-1} corresponding to the fingerprint region. They might be attributed to the C-O stretch of polysaccharides that are mainly predictable in this area (Skjemstad and Dalal, 1987). For the soil texture, clay and silt seem to have a few close bands that are good predic-

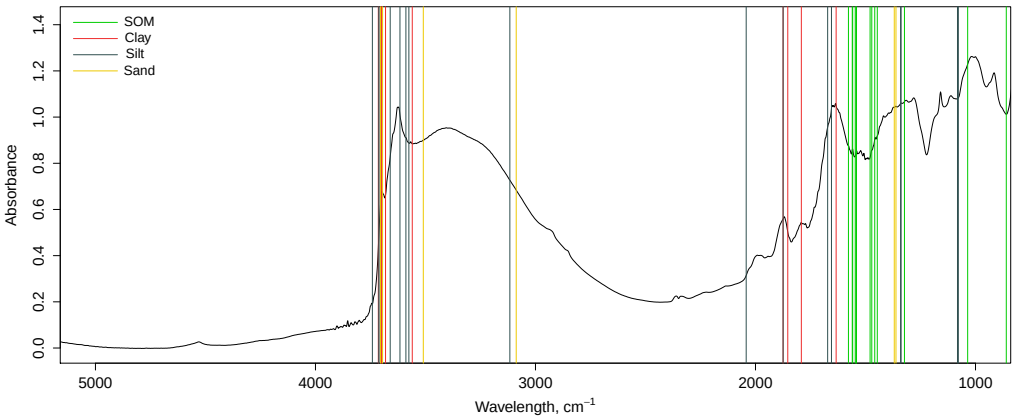


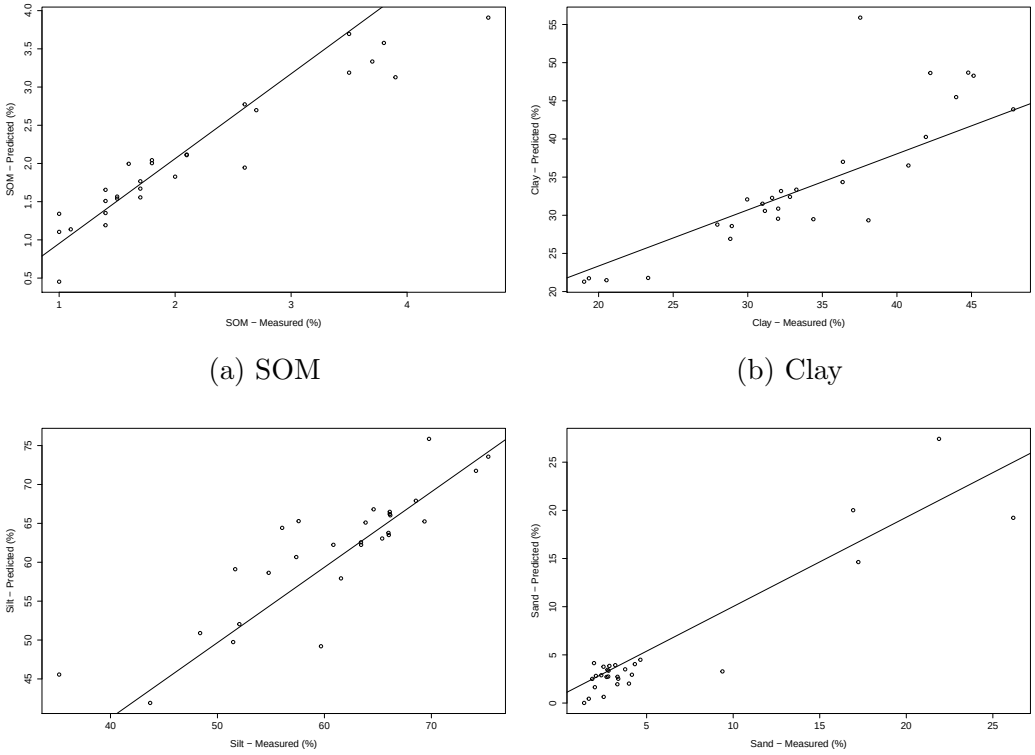
Figure 4.4: Bands "good predictor" for soil organic matter, clay, silt and sand. They are extracted from the set of rules used in cubist to build the prediction.

tors. Around 1640cm^{-1}, the peaks corresponding to the water are suitable as well as between 1700 and 2050cm^{-1}, corresponding to the quartz. Around 3620cm^{-1} most of the good predictors are found. These bands are linked mainly to clay minerals. Kaolinite has an OH vibrational bond at 3620cm^{-1} and 3695cm^{-1}. Smectite shows broad absorption bands between 3600 and 3800cm^{-1} due to the OH stretching vibration. For more description about the correspondent wavelengths to each soil property, see Part 4.2.2.1.

Figure 4.5 represents the observed values plotted against the predicted ones for the external validation. For clay and silt, the high RMSE and low R^2 are due to a few values that seem to influence the slope of the regression line. In contrast, for sand the regression line fits the pattern of the bulk of the data and is not influenced by the four extreme cases that fit approximately the line. SOM is well predicted

4.1.1.3 Support vector machine (SVM)

The results for the calibration using the support vector machine model are in Table 4.3. The model shows generally lower results than for the two other methods. This was an expected result, given that Behrens and Scholten (2006) described SVM as performing poorly comparing with other calibra-



(a) SOM

(b) Clay

(c) Silt

(d) Sand

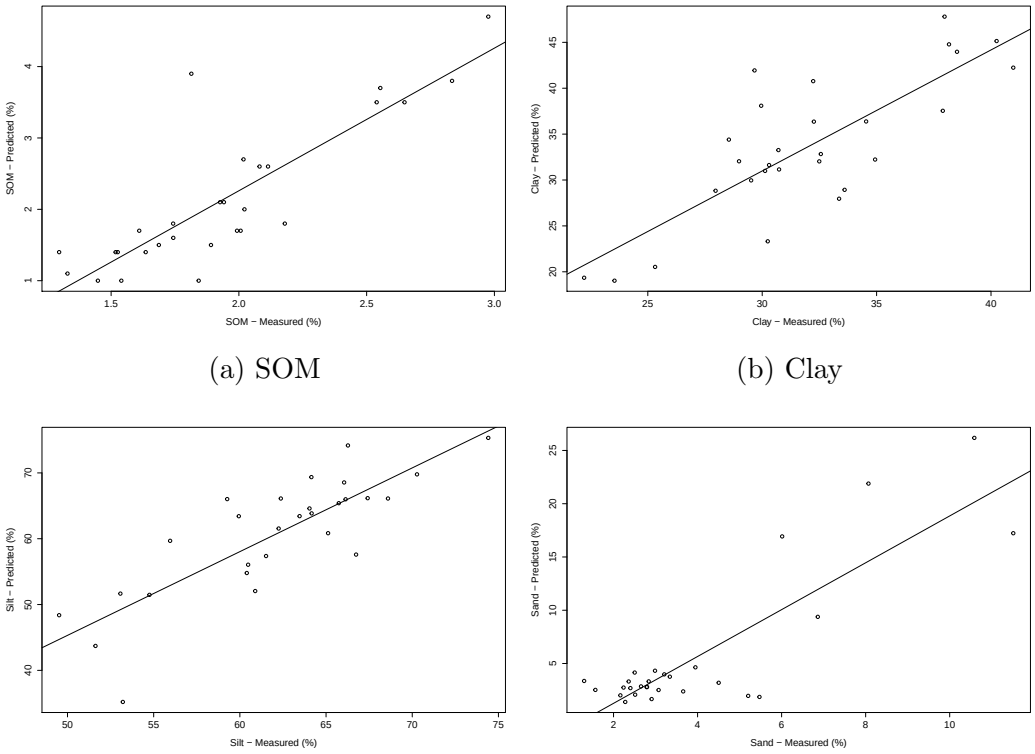
Figure 4.5: External validation of cubist for organic matter, clay, silt and sand.

Soil property	Gamma	Cost	Epsilon	Training dataset (n=111)		Independant validation (n=28)	
				R^2	RMSE	R^2	RMSE
SOM in %	0.001	1	0.1	0.89	0.59	0.85	0.53
Clay in %	0.00083	1	0.1	0.76	3.97	0.84	4.84
Silt in %	0.0015	1	0.1	0.74	4.50	0.89	5.38
Sand in %	0.00098	1	0.1	0.80	4.39	0.90	3.49

Table 4.3: Results for the calibration using SVM. The values for Gamma, Cost and Epsilon were automatically selected according to the ksvm function of the kernlab package in R (Karatzoglou et al., 2004).

tion methods. The results for the internal validation show the same trends than for the other models. We get the better results for SOM, followed by sand and then clay and silt. Especially, the cross validation for SOM gives an R^2 of 0.89 which is very low comparing to the other methods in this study. For the external validation, the prediction of SOM is good ($R^2 = 0.72$) but outperform as well the other methods. However, the RMSE decreases slightly comparing to the internal validation. For clay, the calibration gives good ($R^2 = 0.76$) to fair ($R^2 = 0.66$) results which could correspond to the large variability of the clay distribution. The RMSE is also high in comparison with the other calibration. The SVM is poorly capable of clay prediction. For silt, the internal validation shows good results ($R^2 = 0.74$) but with a high RMSE (RMSE = 4.5). The validation of the independent dataset shows an extremely high RMSE of 5.38 even the R^2 is comparable in term of goodness to the other models. In comparison, sand has good possibilities of prediction with SVM, given the good results for internal ($R^2 = 0.8$, RMSE = 4.39) as well as for external ($R^2 = 0.78$, RMSE = 3.49) validation.

The plot of the external validation gives us information about the ability of the model for further prediction. SOM have generally a homogeneous distribution along the regression line. Only one value seems to have a large deviation. The prediction performs well in this case. For clay, the regression line fits the data, there is no observation with high deviation although they seem to be scattered in three clusters. Silt contains only one observation that should count for most of the RMSE. However, the regression line does not seem to be affected by this value and performs well. Sand shows the same distribution as for the internal validation. There is no data far away from



(a) SOM

(b) Clay

(c) Silt

(d) Sand

Figure 4.6: External validation of SVM for organic matter, clay, silt and sand.

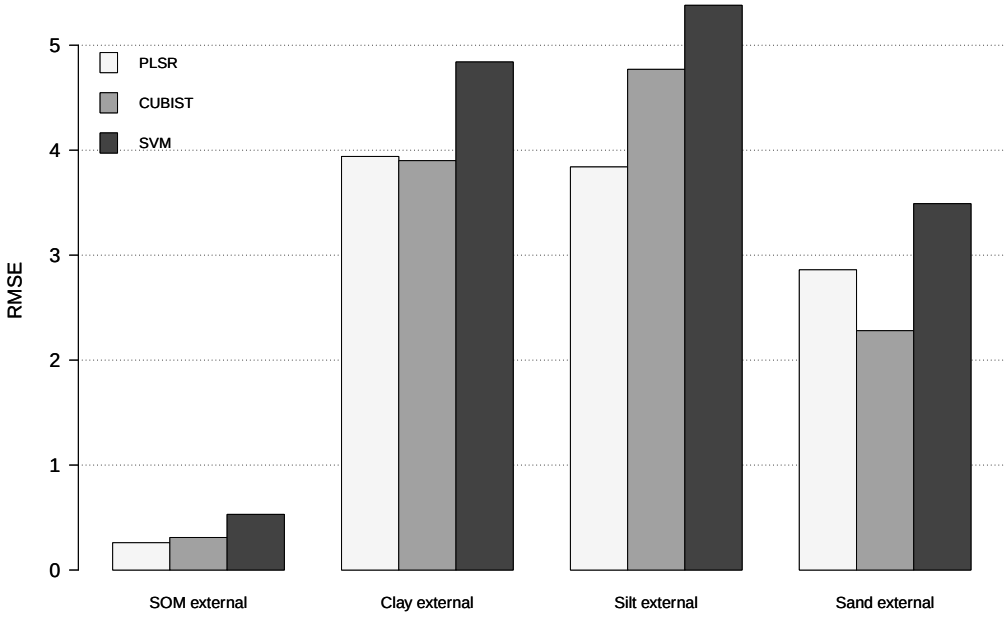


Figure 4.7: Comparison of the RMSE for the three calibration methods separately and for organic matter, clay, silt and sand.

the line; the slope of the regression line represents the trend.

4.1.2 Best model and comparison

4.1.2.1 Calibration methods performance

The results for all predicted target soil properties suggest that the best model is clearly Partial least square regression followed by cubist and support vector machine, Figure 4.7). However, we have high variability of the R^2 and the RMSE between the predicted soil property and also between the internal and the external validation. Regarding this, cubist appears to outperform PLSR and SVM according to the external validation. These results confirm Rossel and Behrens (2010), when the calibration methods using selection feature techniques outperform the methods using the full spectral range like SVM. More in detail, we can see that PLSR and cubist produce similar performance

for the prediction of SOM (R^2 >0.9) for both internal and external validation. SVM gives the lower results for both RMSE and R^2. It may be due to the fact that SVM uses all the data for the data for the calibration when cubist and PLSR allow respectively a band selection of the "best predictor wavelengths" (see Figure 4.4) and a "weight" given by the loadings values (see Figure 4.2). The variability between the internal and external validation is not large, excepted for the SVM calibration. As we can see in Figure 4.1a, SOM prediction is the result of band combinations with very strong positive or negative loadings. When selecting all the values for the calibration, SVM takes bands that can have negative correlation for a soil matter element and a positive relation to another SOM element, although they both are included in the soil organic category.
In contrast with the good prediction of the SOM, it appears difficult to validate clay estimation independently. We find again the same trend than for SOM prediction, with the best results given by PLSR, cubist and finally SVM. The variability is however larger between the validation of the calibration and the independent validation, assuming a greater variability of the clay content. We confirm this idea with the observation of the RMSE. In both cases SVM appears to badly represent the heterogeneity of the observations. The prediction of Silt gives also great variability between the internal and the external validation. The PLSR model gives the best results for R^2 as well as for RMSE, excepted for the RMSE of the independent validation where cubist outperform largely the two other models. It is apparent from this study and those conducted by others (Calderón et al., 2011; Rossel and McBratney, 2008; Wetterlind et al., 2013) that a method with feature selection is more suitable for predicting a target soil property with large variability. Especially in the MIR, a few bands represent most of the variability of the silt and this is therefore inappropriate to use the full spectral range. As for Silt, the prediction of sand reveals the efficiency of the band selection methods. SVM gives the lower results with a high RMSE. Despite the variability of the sand in our study area (see Figure 3.1), PLSR and cubist give good results for both internal and external validation.

4.1.2.2 Interpretability

For both PLSR and cubist calibration models, we get the possibility to extract the wavelengths that are "good predictor" for the target soil property. In terms of interpretability, SOM is related to the wavelengths of N-H

stretch of amines, the C=O bond of carboxylic acids and to the C-O stetching vibration of polysaccharides. the cubist model also gives us the bands near 1600cm^{-1} as good predictors, they can be related to the C=N or C=C stretching bond of amides as well as to the C-O stretch of polysaccharides for the bands near 860cm^{-1}. Clay is related to the wavelengths that represent absorption due to the OH stretching and bond vibration of kaolinite and to the OH bending of silicate structures. Generally, the clay minerals aborption wavelength are suitable for the soil texture classes. This is confirmed with the absorption bands of silt and sand that are mainly concentrated in the region around 3600cm^{-1}. Sand especially has absorption bands related to quartz structures with the absorption bands of SiO2 stretching and bending vibration.

4.2 Soil-terrain relationships through spectroscopy

4.2.1 Results

4.2.1.1 Precision

As stated in the methodology part, we compute for each spectral band along the range between 5300cm^{-1} and 850cm^{-1} its relationship to the terrain attributes. One part of the result is presented through the R^2 (see Figure 4.8) that shows how close the terrain attributes are related to the spectra.
The R^2 values spread between 0.02 (no correlation) and 0.51 (medium-good correlation) along the spectral range. There are large variation with, for example, a peak at 3620cm^{-1} followed by a low correlation at 3590cm^{-1}. However, the results are generally smoother in the range between 5300 and 2000cm^{-1} than between 2000 and 830cm^{-1} presenting high discontinuities. The largest R^2 corresponds to the peak at 3620cm^{-1} with a value of 0.51. This shows good relationships in the prediction of the spectra in this spectral band. A second peak corresponding to 4950cm^{-1} displays a R^2 of 0.47 that is a relatively good correlation. Two others high R^2 have to be highlighted; the first on at 900cm^{-1} with a R^2 of 0.45 and the second at 3030 with a value of 0.43. They exhibit a relative good relationship between the spectra and the terrain attributes. Other peaks with an inferior correlation are represented in the graph. At 1075cm^{-1}, the R^2 is of 0.35 when at 1639cm^{-1} the R^2 is 0.29. In the range between 4500 and 3700cm^{-1}, the R^2 value increases relatively constantly from 0.24 to 0.34. Lower than 0.25, we assume that the

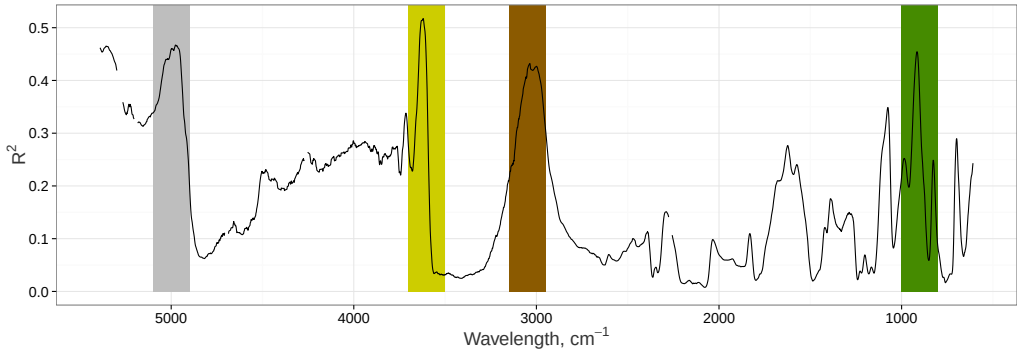


Figure 4.8: R^2 of the regression model between the terrain attributes and the spectra for each band. The colours represent the bands where the correlation is significant. This is to link with Figure 4.9.

R^2 explains a too low quantity of the variance to be significant and are thus not described here.

4.2.1.2 Significance

A second part of the results can be analysed through the P-value that is related to the significance of each predictor in the model (see Figure 4.9). We selected six representative terrain attributes (Elevation, LS-factor, Profile curvature, Plan curvature, Land use and Geology) to explain the goodness of the fit.

The P-value of the different terrain attributes spreads in the complete possible range, between 0 (significant influence of the parameter in the prediction) and 1 (the significance of the parameter cannot be confirmed).

Elevation has, for most of the bands, a P-value higher than 0.05 (low levels of significance). This terrain parameter exhibits however bands where the P-value is lower than 0.05 among which a few bands between 900 and 1800cm^{-1}, at 900, 1100 and 1750cm^{-1}; the bands between 3600 and 3750cm^{-1}; the bands between 4900 and 5050cm^{-1}.

LS-factor produces a different combination, with no significant P-values in the range between 850 and 3000cm^{-1}. In contrast, a full range between 3600 and 4600cm^{-1} appears to be either slightly above the 0.05 limit, or slightly underneath. Elevation has for these ranges very high values. Other bands with a clear P-value under 0.05 are at 3000cm^{-1} and 3600cm^{-1}.

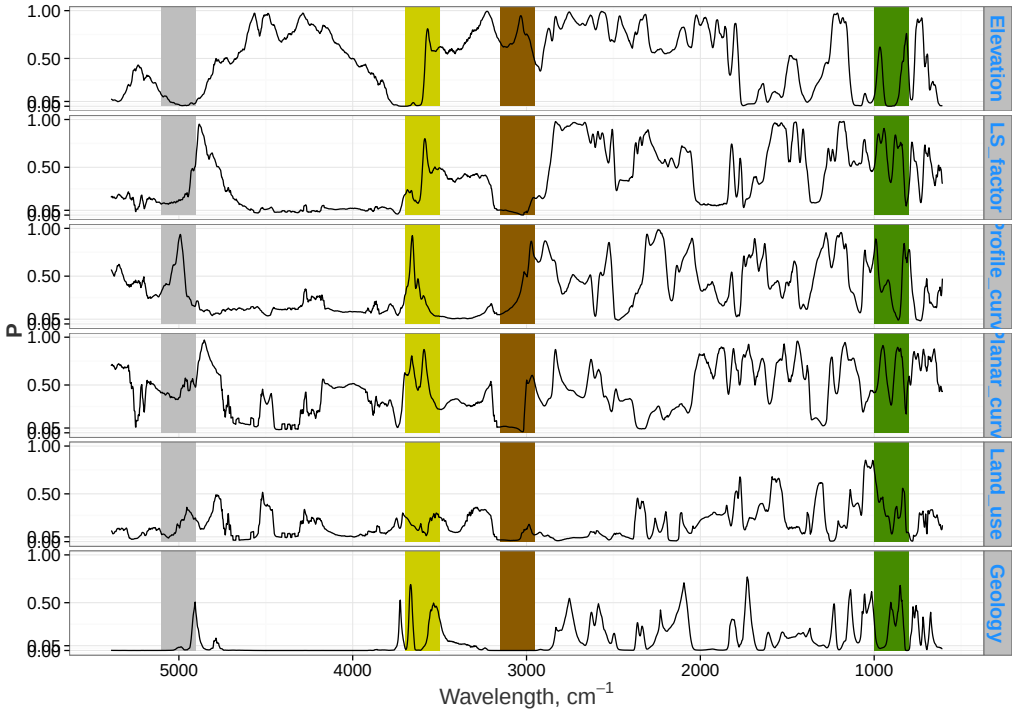


Figure 4.9: P-value of six representative terrain parameters used in the robust linear model. Colours exhibit the relation to the R^2 value in Figure 4.8.

Profile curvature has only a few peaks corresponding to a P-value lower than 0.05. They are located at 750, 850 and 2500cm^{-1} and show that this terrain covariate is in generally less significant than the other drawn parameters in the prediction of the spectra.
Plan curvature shows the same characteristics as profile curvature given that the terrain parameters has only a few wavelengths with a low P-value. These wavelengths are located at 2350, 3030, 4300 and 4400cm^{-1}.
Landuse exhibits generally low P-value, especially in the range between 2400 and 2750cm^{-1} as well as between 3050 and 3150cm^{-1} and 4100 and 4200cm^{-1}. The other significant band are positioned at 2150, 1250 and 800cm^{-1}. Geology has generally good prediction abilities because of its low P-value for large bands. All the bands comprises between 1870 and 2050cm^{-1}, 2850 and 3300cm^{-1} , 2750 and 4700cm^{-1} and 4950 and 5350cm^{-1} have a significant P-value. There is no P-value higher than 0.75 for the whole spectral range. The others bands are at 750, 975, 1200, 1650, 2400-2480, 3150 and 4750cm^{-1}.

4.2.1.3 Interelation

The aim of combining the R^2 with the P-value is to highlight the parameters that affect the prediction for each spectral band. Thus, the good R^2 values have to be put in association the P-value to show the ones that are relevant. In Figure 4.9 and 4.8, the colours represent the four higher values of R^2 associated to the P-value of the terrain parameters. The black colour represents the 4950cm^{-1} band with a R^2 of 0.47. This band is linked to the low P-value of the parameters elevation and geology and nearly to the parameter landuse. The yellow colour is associated to the 3620cm^{-1} band with the higher R^2 at 0.52. This R^2 seems to be due to the elevation, LS-factor and geology. We could also associate plan curvature because of its P-value on the limit of 0.05. The red colour at 3030cm^{-1} is combined with a R^2 of 0.43. The terrain parameters able of prediction in this spectral range are the LS-factor, plan curvature and geology. Finally, the R^2 of 0.45 at 900cm^{-1} is expressed by the green colour. In this case, only elevation can clearly be significant for the prediction although profile curvature seems to have a low P-value in the same range.

4.2.2 Interpretation and discussion

4.2.2.1 Band assignement for Soil Organic Matter

Soil MIR spectra is well suited for analysis of soil organic matter (Stenberg and Rossel, 2010). As described in Russell and Fraser (1994), the band absorption in the MIR region can be divided between the fingerprint, the double-bond, the triple-bond and the X-H stretch regions depending on the fundamental molecular vibration of the soil properties. Thus, we can assign bands to the R^2 in order to analyse what kind of soil properties are better detected than others. This analyse in based on a literature review.
In the fingerprint region, band near 1075cm^{-1} may be attributed to CO stretching vibrations of carboxylic acids in SOM (Simonescu, 2012). The Polysaccharides are almost exclusively detectable in this region due to the C-O stretch at 1050cm^{-1} (Skjemstad and Dalal, 1987; Haberhauer and Gerzabek, 1999), 1160cm^{-1} (Calderón et al., 2011) and 1170cm^{-1} (Rossel and Behrens, 2010). Badly predicted, the vibrational bond of carbohydrates at 1050cm^{-1} (Rossel and Behrens, 2010) and 1051cm^{-1} (Janik et al., 2007) are created by the CO and -COH stretch, respectively. A large absorption at around 1075cm^{-1} may be caused by the CO stretching bond of alcohols (Simonescu, 2012).
The double-bond region is particularly suited for soil organic group detections. Although the R^2 shows generally low values, many groups in SOM are due to fundamental vibrations in this region. Aromatics –CH exhibits a plane deformation at 1238cm^{-1} (Janik et al., 2007) as well as at 1930cm^{-1} (Haberhauer and Gerzabek, 1999) corresponding to the C=O stretch and in the region comprises between 1587 and 1639cm^{-1} (Ziechmann, 1964) for the -C=C- stretch. The bands near 1404cm^{-1} and 1610cm^{-1} are associated with the N-H bending of Amines (Rossel and Behrens, 2010; Simonescu, 2012).
The alkyls assymetric-symmetric doublet with the CH_2 deformation are mainly concentrated in the bands between 1400 and 1450cm^{-1} (Janik et al., 2007), but also at 1530 and 1639cm^{-1} (Skjemstad and Dalal, 1987) due to the C=N and C=O stretch.
C=O carboxylic acids has characteristic bands near 1730cm^{-1} (Rossel and Behrens, 2010; Haberhauer and Gerzabek, 1999), 1708cm^{-1} (Janik et al., 2007) according to the -COOH stretch band at 1629cm^{-1} (Simonescu, 2012) due to the C=O stretching bond. Major peaks or R^2 are present in the bands related to the amides. Near 1640cm^{-1}, C=O stretch band combination of

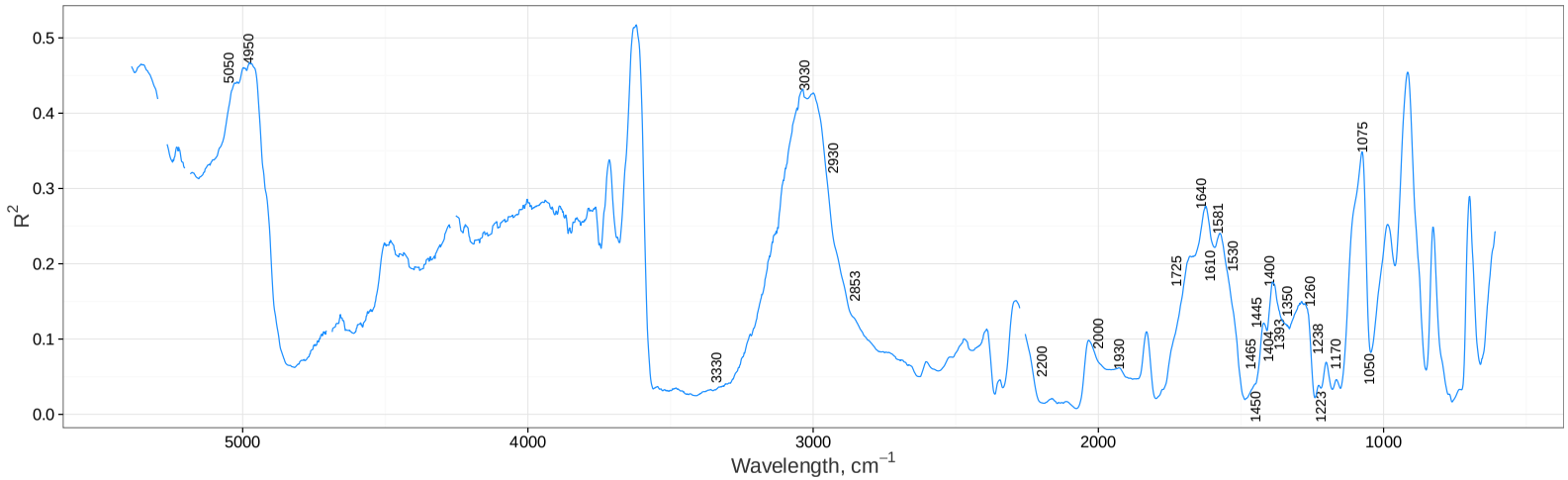


Figure 4.10: Band assignement for soil organic matter in the R^2 curve. Each number represents a soil organic element.

amides (Rossel and Behrens, 2010; Skjemstad and Dalal, 1987) are correlated with a high R^2. Around 1510-1530cm^{-1}, the bands can be likely assigned to the C=N and C=C stretching bands of amides (Haberhauer and Gerzabek, 1999) as well as at 1640cm^{-1} (Rossel and Behrens, 2010), 1929cm^{-1} (Simonescu, 2012) and 1660cm^{-1} (Janik et al., 2007). Aliphatic-CH stretch are detectable at 1465cm^{-1} and 1350cm^{-1} (Baes and Bloom, 1989; Rossel and Behrens, 2010). The absorption bands between 1350 and 1445cm^{-1} are most likely due to methyl (Rossel and Behrens, 2010) when ether at 1260cm^{-1} (Ziechmann, 1964) may be attributed to the COC stretching vibrations. The phenolic group is assigned to the bands near 1275cm^{-1} (Baes and Bloom, 1989) because of the C-O stretch and at 1581, 1393, 1223cm^{-1} (Janik et al., 2007) by the C=C skeletal vibrations. It has also to be noted the weak absorption band of the carboxylate anion near 1393cm^{-1} (Janik et al., 2007). The triple-bond region is not very useful for the detection of organic matter. The carbohydrate groups are however notable, with a –COH stretch in the bands between 2000 and 2200cm^{-1} (Janik et al., 2007).
Many bands in the X-H stretch regions are assigned to SOM. The broad band near 3030cm^{-1} may be attributed to the C-H aromatics (Rossel and Behrens, 2010) when amine has N-H stretch vibrations near 3330cm^{-1} (Rossel and Behrens, 2010). The region between 2850 and 2930cm^{-1} are useful for the detection of Alkyl asymmetric-symmetric doublet (Stenberg et al., 2010) given that they are free of overlaps from other vibrations. They are linked to the CH_2-alkyls for 2853 and 2922cm^{-1} (Stenberg and Rossel, 2010; Janik et al., 2007) and at the C-H stretching at 2850 and 2930cm^{-1} (Rossel and Behrens, 2010). Alcohols with the –OH bonded phenol is predictable near 3333cm^{-1} (Ziechmann, 1964).
Finally, a small range of the NIR spectra is present. It corresponds to the combination and the first overtone. These bands are useful at 4950cm^{-1} for the prediction of amides and at 5050cm^{-1} for phenolic (Viscarra Rossel et al., 2006).

4.2.2.2 Band assignement for soil mineralogy

The MIR spectra is also suitable for the prediction of soil mineralogy and water. As for Figure 4.10, Figure 4.11 highlights the band assignment for minerals and free water. In contrast to SOM, most of the useful bands of mineral groups are concentrated in the X-H stretch and fingerprint regions although the double bond (DB) appears to be good predictor, but with a lot

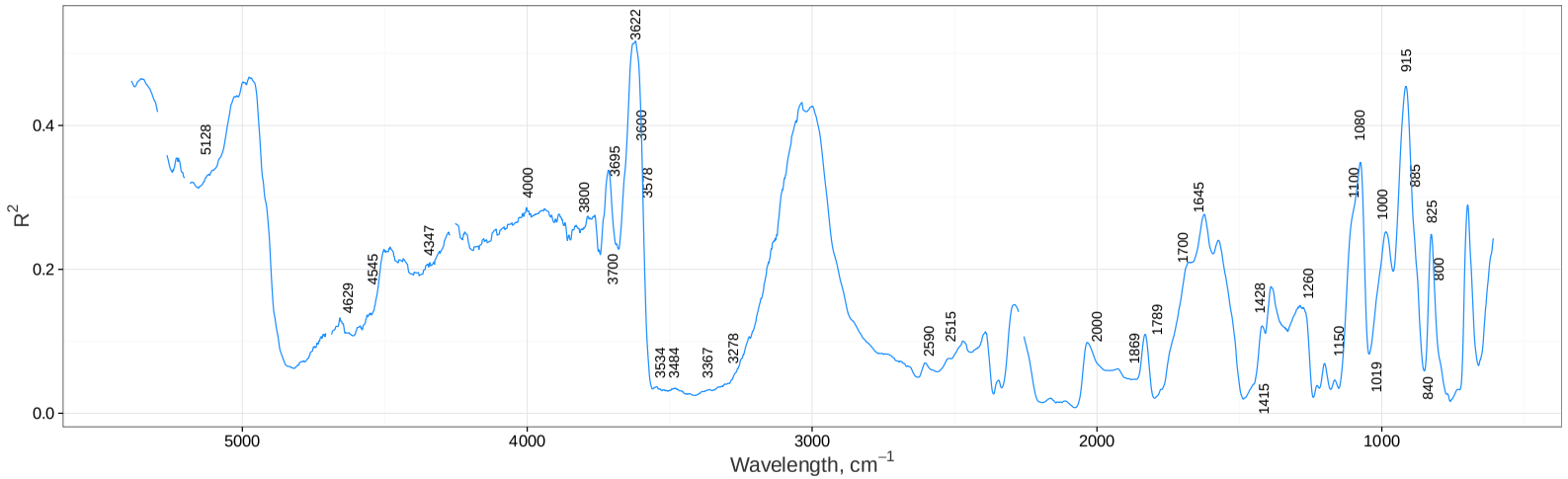


Figure 4.11: Band assignement for soil mineralogy in the R^2 curve. Each number represents a soil mineral element.

of overlap with other soil properties.
In contrast with SOM, many absorption bands of minerals are present in the Fingerprint region. The bands at 915cm^{-1} (Rossel and Behrens, 2010) and 1150cm^{-1} (Russell and Fraser, 1994) are particularly suitable for the detection of kaolinite with the Al-Oh bending vibrations also studied in Nguyen et al. (1991) with the bands near 920 and 1019cm^{-1}. Smectite exhibits a few peaks at 840cm^{-1} (Russell and Fraser, 1994), 885cm^{-1} and 915cm^{-1} (Rossel and Behrens, 2010), due to the vibration of alFe-OH bond. For the same molecular bond, illite has absorption bands between 825 and 890cm^{-1} (Russell and Fraser, 1994). Bands near 800 and 1080cm^{-1} are due to the fundamental SiO2 stretching and bending vibrations of quartz (Russell and Fraser, 1994; Chukanov, 2013; Madejová and Komadel, 2001). Si-O stretching or bending OH at 1100cm^{-1} (Calderón et al., 2011) is suitable to detect silicate structures as well as Serpentine minerals such chrysotile and antigorite near 1000cm^{-1} (Russell and Fraser, 1994). Olivine-group minerals have absorption bands near 1000cm^{-1} (Russell and Fraser, 1994; Chukanov, 2013). Carbonates have a weak absorption band at 878cm^{-1} (Russell and Fraser, 1994).
The double-bond (DB) region is rather difficult to interpret for clay minerals. Minerals like kaolinite have a broad vibrational bond between 1700 and 2000cm^{-1} (Stenberg et al., 2010), that is also the absorption band for several organic matter (see part 4.2.2.1). Carbonates CO_3^{-2} have is characteristic of the band near 1415cm^{-1} (Rossel and Behrens, 2010) when calcite is typical of the band at 1428cm^{-1} (Chukanov, 2013; Madejová and Komadel, 2001). Bentonite can have a weak combination band at 1260cm^{-1} (Calderón et al., 2011). In contrast with clay minerals, quartz mineral are characteristic of the double-bond bands with absorption feature at 1789, 1869, 2000cm^{-1} (Nguyen et al., 1991) and between 1700 and 2000cm^{-1} (Stenberg et al., 2010). Absorption bands for silicates are detectable at 2000 and 1790cm^{-1} (Haberhauer and Gerzabek, 1999).
The triple-bond region has not mineral absorption bands specific to minerals. The presence of minerals instead of organic matter can be confirmed by a view on the continuum-removed spectra (Figure 4.12) that exhibits low reflectance values, especially near 915, 1100, 1400 and 1650cm^{-1}.
The X-H stretch region highlights the presence of kaolinite with the OH vibrational bond at 3620cm^{-1} (Rossel and Behrens, 2010; Russell and Fraser, 1994), 3695cm^{-1} (Rossel and Behrens, 2010) and 3700cm^{-1} (Russell and Fraser, 1994; Chukanov, 2013; Madejová and Komadel, 2001). Another clay

mineral closely linked to the high R^2 is the smectite with its absorption bands between 3600 and 3800cm^{-1} (Stenberg et al., 2010) by the OH stretching vibrations and more specifically at 3620cm^{-1} (Rossel and Behrens, 2010) and 3622cm^{-1} (Russell and Fraser, 1994). Both kaolinite and smectite have close vibrational OH bonds but its broader absorption bands (Stenberg et al., 2010) can distinguish smectite. The low value in the Figure 4.12 around 3600cm^{-1} attests to the high mineral absorption of clay minerals in this region of the spectra. Illite is also largely represented in the X-H stretch region with the OH bond at 3620cm^{-1} (Rossel and Behrens, 2010) that can be coupled with 3625cm^{-1} (Stenberg et al., 2010) or like smectite between 3600 and 3800cm^{-1} with the OH stretching vibrations. Carbonates have typical absorption bands at 2515 and 2590cm^{-1}(Stenberg and Rossel, 2010) when the bentonite vibrational bond overlaps kaolinite, smectite and illite at 3622cm^{-1} (Calderón et al., 2011). Glauconite shows a weak absorption at band 3534cm^{-1}and 3578cm^{-1} (Russell and Fraser, 1994) due to the Fe-OH bond. Absorptions near 3367cm^{-1} shows the presence of olivine-group mineral by the OH stretching band.
The NIR range combination can help to differentiate between the elements avoiding overlapping into the clay minerals. Kaolinite exhibits an Al-OH bend plus O-H stretch at 4545cm^{-1} near to the metal O-H bend of smectite at 4347cm^{-1}and to the O-H stretch of illite at 4347cm^{-1} (Viscarra Rossel et al., 2006). Carbonates have a broad absorption region between 4000 and 4347cm^{-1} and two typical bands at 4629 and 5405cm^{-1} (Hunt, 1970).

Water has specific absorption bands in the double-bond (DB) region at 1645cm^{-1} (Rossel and Behrens, 2010) with the HOH vibrational bond and in the X-H stretch region the OH stretch at 3278 and 3484cm^{-1} (Rossel and Chen, 2011). Note the OH bond in the combination region of the NIR range at 4545 and 5128cm^{-1} (Viscarra Rossel et al., 2006).

4.2.2.3 Influence attributes on soil composition

The relationships between terrain and soil property is hard to define with accuracy. This link is often unknown and very noisy, (Moore et al., 1993; Zhang et al., 2012) because soil properties is the results of many environmental factors as described in Jenny et al. (1941) and McBratney et al. (2003). Terrain for prediction of soil organic matter or soil mineral gives generally bad results. The distribution characteristics have been studied as an in-

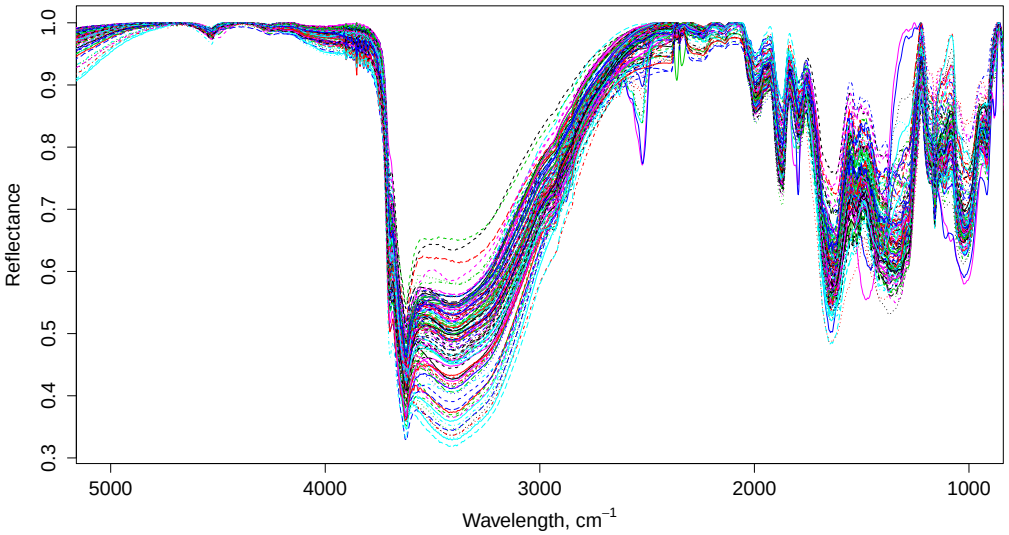


Figure 4.12: Continuum Removal of the reflected spectra in order to highlight the mineral absorption.

fluence of a coupled effects on terrain, soil texture and soil genetics types (Zhang et al., 2012). Around 5000cm^{-1}, SOM through amides and phenolic seem to be explained by elevation, geology and to a lesser extent by landuse. Phenolic in soil in commonly linked to decomposing plant litter. The decomposition of dead plant materials conducts the oxidation of the humus matter and the transformation into less complex forms. In this case, landuse is fully linked to the phenol in soil by the release of organic materials and by the change in the humus degradation rate. The link between elevation and phenol is largely unexplored and remains unclear. It could be explain indirectly ba the underlying change in solar irradiation of by soil particles movement. Spielvogel et al. (2007) gives some explanation about it when he highlighted the link between grain size and phenols. Phenols bind to clay minerals and therefore coarser grain could explain the variability of this element. In our specific study area, geology is linked to elevation (geological layers correlated with altitude). It has been also widely studied the impact of the bedrock properties on the grain size distribution Kiem and Kögel-Knabner (2003). It seems that phenol can be predicted by elevation and geology by the latent

factor of grain size distribution. Amides are derived from microbial proteins, suggesting that the litter is rapidly transformed by soil fauna (Whalen and Sampedro, 2010). Landuse as a main role on it, but we are limited by the resolution of this terrain derivative. An important part of the amide concentration in the soil is due to the amount of residues from plants and the velocity of the decomposition. In our catchment area, shrubs and abandoned terraces are increasing with altitude and could explain the role of elevation as good predictor for amides. However, the link with geology is not clear. Might be with the underlying grain size because amides have physical protection within micro aggregates due to clay minerals (Whalen and Sampedro, 2010). Around 3030cm^{-1} aromatics and alkyls are as main predictor the LS-factor, plan curvature, landuse and geology. Aromatics and alkyls are, in contrast with phenolic and amides, linked to substrates for a large number of phungi and bacteria, and not directly linked to soil microbial activity and plant litterfall quantity. Vegetation type seems to have the main influence on the aromatic and alkyl occurrence, given that this factor drives soil respiration (Sjögersten and Wookey, 2004). Geology also have a main role by the underlying grain size distribution. Aromatic and alkyls have relationships with clay particles. Clay minerals by association with organic C can slow the degradation by protection from decomposition. The relation between slope properties of the LS-factor and the plan curvature in association with alkyls and aromatics are not clearly defined. The slope properties have influences on the topsoil, among which the soil erosion by water. Organic matter is largely impacted by accumulation and deposition of topsoil organic substances (Berhe, 2012) and lead to soil aggregates that can contain a higher rate of alkyls and aromatics comparing to the total soil organic carbon value. There is a potential biological protection of organic C by with the fine grain size elements (mainly clay) that are moved with the downslope movement of organic-rich clay (Huang et al., 2011). Soil minerals are generally well linked to the terrain attribute such as elevation, LS-factor and geology. At around 3600cm^{-1}, clay minerals are represented through kaolinite, smectite and illite. Kaolinite can be found in most soil into the clay fraction. It is the result of a highly weathered soil even the mineral is itself very resistant to chemical degradation. Therefore, it is found in high quantity and it is easily detectable with infrared spectroscopy. Smectite refers to soil that can be expansible due to the high capacity of water retention of the mineral. Soil with high smectite content can expend until 30%. Illite are the result of the decomposition of the micas and feldspars (composition of granite). The cor-

relation of these terrain attributes with these clay minerals is in agreement with the soil particle redistribution and the erosion-deposition processes developed in Brady and Weil (2010). Kaolinite, smectite and illite are moved downward according to the movement of water. This is why they are linked in a clear way to the slope properties of the ls-factor and to the elevation. For bands near 915cm^{-1}, elevation and geology have the largest influence. These bands correspond to kaolinite ad smectite. Their interrelation are interpreted above.

Chapter 5

Conclusion

In this thesis we have argued that the MIR range of a soil spectra contains enough information to be used for describing soil-terrain relationships quickly. The results bring us to the following conclusions:

There is high correlation between soil and spectra. This has been highlighted with multivariate statistics and the wavelengths corresponding to the main soil elements were extracted. Notably, we note that the bands "good predictors" for soil organic matter and soil mineralogy are clearly linked to specific spectral range. The facilities of interpretation offered by the models brought us to the conclusion that the spectra contains all the information needed for soil monitoring.
The MIR range, when linked to topographic secondary information, is an useful tool for describing the influence of terrain on the soil. MIR gives reliable view on the soil formation and evolution factors. In particular, we can understand in which proportion the terrain analysis is suitable to explain the spatial distribution of soil organic matter or soil mineralogy.

Our results support earlier works analyzing the possibility of soil spectroscopy for describing the spatial variability of the soil. The main advantages of Vis-NIR and MIR spectroscopy are (i) the relative cheap measurements comparing to conventional laboratory analysis (ii) the possibility to measure directly in the field and therefore avoiding to carry the samples and (iii) the specific-wavelength absorption gives us information about soil composition. The development of techniques to derive soil maps from the spectra itself would decrease significantly the costs related the classical soil surveys. Espe-

cially in our study area, we show that the spectra is a powerful tool to gather information about soil characteristics without any prior study and with lack of information.

This study has been limited by the lack of auxiliary information such as high resolution climate data that we could have incorporated. In the same way, the resolution of digital elevation model have a direct impact on the results. We could therefore include higher resolution dem, landcover and geological covariates and improve significantly the model. The size of the study area (4.2km^2) also brings some limitations concerning the use of remote sensing data. Notably, hyperspectral images were not present in this region and we did therefore not used them. Another limit is that there is not spatial link between the observations. We implemented a robust linear model that does not take into consideration the spatial distribution of the samples.

Future projection of this work would involved combining not only topographic related covariates, but all the parameters of the famous equation described in Jenny et al. (1941). It would include bedrock properties, climate, temporal change and all the environmental covariates that affect soil variability. This should therefore be conducted in a larger study area in order to reduce the error affected to the DEM and to climate information. The use of remote sensing as auxiliary source of data has to be considered. Another possibility to improve the evaluation would be use a model that takes into account the spatial distribution of the samples. Particularly, when introducing all this characteristics into a new model, we can expect to get results that will allow a spatial prediction of a spectra. This could modify the use of traditional soil analysis and allow high resolution mapping in area with few soil information. Especially, the launch of new hyperspectral sensors such as HYMAP would support enhancement, comparison or validation of the prediction.

Bibliography

Andersen, R. (2008). *Modern methods for robust regression*. Number 152. Sage.

Baes, A. and Bloom, P. (1989). Diffuse reflectance and transmission fourier transform infrared (drift) spectroscopy of humic and fulvic acids. *Soil Science Society of America Journal*, 53(3):695–700.

Barnes, R., Dhanoa, M., and Lister, S. J. (1989). Standard normal variate transformation and de-trending of near-infrared diffuse reflectance spectra. *Applied spectroscopy*, 43(5):772–777.

Baveye, P. (2006). A future for soil science. *Journal of soil and water conservation*, 61(5):148A–151A.

Beckett, P.H.T., B. S. (1978). Use of soil and land-system maps to provide soil information in australia. *Division of soils technical paper*, 33.

Behrens, T. and Scholten, T. (2006). Digital soil mapping in germany—a review. *Journal of Plant Nutrition and Soil Science*, 169(3):434–443.

Ben-Dor, E. and Banin, A. (1995). Near-infrared analysis as a rapid method to simultaneously evaluate several soil properties. *Soil Science Society of America Journal*, 59(2):364–372.

Berhe, A. A. (2012). Decomposition of organic substrates at eroding vs. depositional landform positions. *Plant and soil*, 350(1-2):261–280.

Bivand, R. S., Pebesma, E. J., and Gómez-Rubio, V. (2008). *Applied spatial data analysis with R*, volume 747248717. Springer.

Brady, N. C. and Weil, R. R. (2010). *Elements of the nature and properties of soils*. Pearson Educational International Upper Saddle River, NJ.

Breiman, L., Friedman, J., Stone, C. J., and Olshen, R. A. (1984). *Classification and regression trees*. CRC press.

Brus, D., Kempen, B., and Heuvelink, G. (2011). Sampling for validation of digital soil maps. *European Journal of Soil Science*, 62(3):394–407.

Burgess, T. and Webster, R. (1980). Optimal interpolation and isarithmic mapping of soil properties. *Journal of Soil Science*, 31(2):333–341.

Burrough, P. A., McDonnell, R., Burrough, P. A., and McDonnell, R. (1998). *Principles of geographical information systems*, volume 333. Oxford university press Oxford.

Calderón, F., Haddix, M., Conant, R., Magrini-Bair, K., and Paul, E. (2013). Diffuse-reflectance fourier-transform mid-infrared spectroscopy as a method of characterizing changes in soil organic matter. *Soil Science Society of America Journal*, 77(5):1591–1600.

Calderón, F. J., Mikha, M. M., Vigil, M. F., Nielsen, D. C., Benjamin, J. G., and Reeves III, J. B. (2011). Diffuse-reflectance mid-infrared spectral properties of soils under alternative crop rotations in a semi-arid climate. *Communications in soil science and plant analysis*, 42(17):2143–2159.

Chukanov, N. V. (2013). *Infrared spectra of mineral species: extended library*. Springer Science & Business Media.

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

Dalrymple, J., L. R. and Conacher, A. (1968). A hypothetical nine unit landsurface form. *Zeitschrift fur Geomorphologie*, 12:60–76.

Demattê, J. A., Campos, R. C., Alves, M. C., Fiorio, P. R., and Nanni, M. R. (2004). Visible–nir reflectance: a new approach on soil evaluation. *Geoderma*, 121(1):95–112.

Desmet, P. and Govers, G. (1996). A gis procedure for automatically calculating the usle ls factor on topographically complex landscape units. *Journal of soil and water conservation*, 51(5):427–433.

Di Gregorio, A. (2005). *Land cover classification system: classification concepts and user manual: LCCS*. Number 8. Food & Agriculture Org.

Donoho, D. L. (1982). Breakdown properties of multivariate location estimators. Technical report, Technical report, Harvard University, Boston. URL http://www-stat. stanford. edu/˜ donoho/Reports/Oldies/BPMLE. pdf.

Donoho, D. L. and Huber, P. J. (1983). The notion of breakdown point. *A Festschrift for Erich L. Lehmann*, pages 157–184.

Dunn, B., Batten, G., Beecher, H., and Ciavarella, S. (2002). The potential of near-infrared reflectance spectroscopy for soil analysis—a case study from the riverine plain of south-eastern australia. *Animal Production Science*, 42(5):607–614.

Durner, W. and Nieder, R. (2006). Bodenkundliches praktikum i. *Skript. Institut für Geoökologie, Abteilung Bodenkunde und Bodenphysik, TU Braunschweig*.

ESRI (2011). Arcgis desktop: release 10.

Filzmoser, P., Serneels, S., Maronna, R., and Van Espen, P. J. (2009). *Robust multivariate methods in chemometrics*. na.

Freeman, T. G. (1991). Calculating catchment area with divergent flow based on a regular grid. *Computers & Geosciences*, 17(3):413–422.

Guisan, A., Weiss, S. B., and Weiss, A. D. (1999). Glm versus cca spatial modeling of plant species distribution. *Plant Ecology*, 143(1):107–122.

Haaland, D. M. and Thomas, E. V. (1988). Partial least-squares methods for spectral analyses. 1. relation to other quantitative calibration methods and the extraction of qualitative information. *Analytical Chemistry*, 60(11):1193–1202.

Haberhauer, G. and Gerzabek, M. (1999). Drift and transmission ft-ir spectroscopy of forest soils: an approach to determine decomposition processes of forest litter. *Vibrational Spectroscopy*, 19(2):413–417.

Hampel, F. (1971). A general definition of qualitative robustness. *Ann. Math. Stat*, 42:1887–1896.

Hampel, F. R. (1974). The influence curve and its role in robust estimation. *Journal of the American Statistical Association*, 69(346):383–393.

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J., and Stahel, W. A. (1986). Linear models: Robust estimation. *Robust Statistics: The Approach Based on Influence Functions*, pages 307–341.

Hartemink, A. E. and McBratney, A. (2008). A soil science renaissance. *Geoderma*, 148(2):123–129.

Hengl, T., Rossiter, D. G., and Stein, A. (2004). Soil sampling strategies for spatial prediction by correlation with auxiliary maps. *Soil Research*, 41(8):1403–1422.

Holmes, G., Hall, M., and Prank, E. (1999). *Generating rule sets from model trees*. Springer.

Hong-Yu, T. (2005). On change in mean maximum temperature, minimum temperature and diurnal range in china during 1951—2002. *Climatic and Environmental Research*, 4.

Höskuldsson, A. (1988). Pls regression methods. *Journal of chemometrics*, 2(3):211–228.

Huang, P. M., Li, Y., and Sumner, M. E. (2011). *Handbook of soil sciences: resource management and environmental impacts*. CRC Press.

Huber, G. P. (1984). The nature and design of post-industrial organizations. *Management Science*, 30(8):928–951.

Huber, P. J. (1967). The behavior of maximum likelihood estimates under nonstandard conditions. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 221–233.

Huber, P. J. (2011). *Robust statistics*. Springer.

Huber, P. J. et al. (1964). Robust estimation of a location parameter. *The Annals of Mathematical Statistics*, 35(1):73–101.

Hunt, G. R. (1970). Visible and near-infrared spectra of minerals and rocks: I silicate minerals. *Modern Geology*, 1:283–300.

Hutchinson, M. (1989). A new procedure for gridding elevation and stream line data with automatic removal of spurious pits. *journal of Hydrology*, 106(3):211–232.

Ivanciuc, O. (2007). Applications of support vector machines in chemistry. *Reviews in computational chemistry*, 23:291.

Janik, L. J., Skjemstad, J., Shepherd, K., and Spouncer, L. (2007). The prediction of soil carbon fractions using mid-infrared-partial least square analysis. *Soil Research*, 45(2):73–81.

Jarmer, T., Vohland, M., and Heuer, A. (2009). Spatial assessment of soil depth from laboratory reflectance measurements and hyperspectral imagery. In *Proceedings of the 6th EARSeL Workshop on Imaging Spectroscopy, Tel Aviv, March*, pages 16–18.

Jenny, H. et al. (1941). Factors of soil formation. a system of quantative pedology.

Jenson, S. and Domingue, J. (1988). Extracting topographic structure from digital elevation data for geographic information system analysis. *Photogrammetric engineering and remote sensing*, 54(11):1593–1600.

Jiang J., Xiang W, Z. W. R. (2012). Research on the water-rock (soil) interaction mechanism of huangtupo riverside landslide in three gorges reservoir. *Journal of Geotech. Eng.*, 34(1):1209–1216.

Karatzoglou, A., Smola, A., Hornik, K., and Zeileis, A. (2004). kernlab-an s4 package for kernel methods in r.

Kiem, R. and Kögel-Knabner, I. (2003). Contribution of lignin and polysaccharides to the refractory carbon pool in c-depleted arable soils. *Soil Biology and Biochemistry*, 35(1):101–118.

Koethe, R. and Lehmeier, F. (1996). System zur automatischen reliefanalyse. 93(4):11–21.

Kuhn, M. and Johnson, K. (2013). *Applied predictive modeling*. Springer.

Kuhn, M., Weston, S., Keefer, C., and Coulter, N. (2012). Cubist models for regression.

Kuhn, M., Weston, S., Keefer, C., and Kuhn, M. M. (2014). Package 'cubist'.

Li, H., Liang, Y., and Xu, Q. (2009). Support vector machines and its applications in chemistry. *Chemometrics and Intelligent Laboratory Systems*, 95(2):188–198.

Liu, J., Liu, M., Tian, H., Zhuang, D., Zhang, Z., Zhang, W., Tang, X., and Deng, X. (2005). Spatial and temporal patterns of china's cropland during 1990–2000: an analysis based on landsat tm data. *Remote Sensing of Environment*, 98(4):442–456.

Madejová, J. and Komadel, P. (2001). Baseline studies of the clay minerals society source clays: infrared methods. *Clays and clay minerals*, 49(5):410–432.

Maronna, R., Martin, D., and Yohai, V. (2006). *Robust statistics*. John Wiley & Sons, Chichester. ISBN.

Martens, H. and Martens, M. (2001). *Multivariate analysis of quality: an introduction*. John Wiley & Sons.

McBratney, A. B., Mendonça Santos, M. d. L., and Minasny, B. (2003). On digital soil mapping. *Geoderma*, 117(1):3–52.

McBratney, A. B., Minasny, B., and Viscarra Rossel, R. (2006). Spectral soil analysis and inference systems: A powerful combination for solving the soil data crisis. *Geoderma*, 136(1):272–278.

Merry, R. and Janik, L. (2001). Mid infrared spectroscopy for rapid and cheap analysis of soils. In *Proc. 10th Australian Agronomy Conf., CD-ROM. Hobart, Australia: Australian Society of Agronomy*.

Mevik, B.-H. and Cederkvist, H. R. (2004). Mean squared error of prediction (msep) estimates for principal component regression (pcr) and partial least squares regression (plsr). *Journal of Chemometrics*, 18(9):422–429.

Mevik, B.-H. and Wehrens, R. (2007). The pls package: principal component and partial least squares regression in r. *Journal of Statistical Software*, 18(2):1–24.

Miklos, M., Short, M. G., McBratney, A. B., and Minasny, B. (2010). Mapping and comparing the distribution of soil carbon under cropping and grazing management practices in narrabri, north-west new south wales. *Soil Research*, 48(3):248–257.

Minasny, B. and McBratney, A. B. (2006). A conditioned latin hypercube method for sampling in the presence of ancillary information. *Computers & Geosciences*, 32(9):1378–1388.

Minasny, B. and McBratney, A. B. (2008). Regression rules as a tool for predicting soil properties from infrared reflectance spectroscopy. *Chemometrics and Intelligent Laboratory Systems*, 94(1):72–79.

Minasny, B., Tranter, G., McBratney, A. B., Brough, D. M., and Murphy, B. W. (2009). Regional transferability of mid-infrared diffuse reflectance spectroscopic prediction for soil chemical properties. *Geoderma*, 153(1):155–162.

Möller, M., Volk, M., Friedrich, K., and Lymburner, L. (2008). Placing soil-genesis and transport processes into a landscape context: A multiscale terrain-analysis approach. *Journal of Plant Nutrition and Soil Science*, 171(3):419–430.

Moore, I. D., Gessler, P., Nielsen, G., and Peterson, G. (1993). Soil attribute prediction using terrain analysis. *Soil Science Society of America Journal*, 57(2):443–452.

Mosteller, F. and Tukey, J. W. (1977). Data analysis and regression: a second course in statistics. *Addison-Wesley Series in Behavioral Science: Quantitative Methods*.

Mouazen, A. M., De Baerdemaeker, J., and Ramon, H. (2005). Towards development of on-line soil moisture content sensor using a fibre-type nir spectrophotometer. *Soil and Tillage Research*, 80(1):171–183.

Nguyen, T., Janik, L. J., and Raupach, M. (1991). Diffuse reflectance infrared fourier transform (drift) spectroscopy in soil studies. *Soil Research*, 29(1):49–67.

Næs, T. (1987). The design of calibration in near infra-red reflectance analysis by clustering. *Journal of Chemometrics*, 1(2):121–134.

Olaya, V. and Conrad, O. (2008). Geomorphometry in saga. *in: T. Hengl, H.I. Reuter (Eds.), Geomorphometry: Concepts, Software, Applications, Developments in Soil Science*, 33:293—308.

Olson, D. M., Dinerstein, E., Wikramanayake, E. D., Burgess, N. D., Powell, G. V., Underwood, E. C., D'amico, J. A., Itoua, I., Strand, H. E., Morrison, J. C., et al. (2001). Terrestrial ecoregions of the world: A new map of life on earth a new global map of terrestrial ecoregions provides an innovative tool for conserving biodiversity. *BioScience*, 51(11):933–938.

Quinlan, J. R. (1993a). *C4. 5: programs for machine learning*, volume 1. Morgan kaufmann.

Quinlan, J. R. (1993b). Combining instance-based and model-based learning. In *ICML*, pages 236–243.

Quinlan, J. R. et al. (1992). Learning with continuous classes. In *Proceedings of the 5th Australian joint Conference on Artificial Intelligence*, volume 92, pages 343–348. Singapore.

Quinn, P., Beven, K., Chevallier, P., and Planchon, O. (1991). The prediction of hillslope flow paths for distributed hydrological modelling using digital terrain models. *Hydrological processes*, 5(1):59–79.

R, C. T. et al. (2012). R: A language and environment for statistical computing.

Reeves, J. B. (2012). Mid-infrared spectral interpretation of soils: Is it practical or accurate? *Geoderma*, 189:508–513.

Riley, S.J., D. G. S. E. R. (1999). Terrain ruggedness that quantifies topographic heterogeneity. *Intermountain Journal of Science*, 5:23–27.

Rossel, R. and Behrens, T. (2010). Using data mining to model and interpret soil diffuse reflectance spectra. *Geoderma*, 158(1):46–54.

Rossel, R. and Chen, C. (2011). Digitally mapping the information content of visible–near infrared spectra of surficial australian soils. *Remote Sensing of Environment*, 115(6):1443–1455.

Rossel, R. V. and McBratney, A. (2008). Diffuse reflectance spectroscopy as a tool for digital soil mapping. In *Digital Soil Mapping with Limited Data*, pages 165–172. Springer.

Rousseeuw, P. and Yohai, V. (1984). Robust regression by means of s-estimators. In *Robust and nonlinear time series analysis*, pages 256–272. Springer.

Rousseeuw, P. J. and Croux, C. (1993). Alternatives to the median absolute deviation. *Journal of the American Statistical Association*, 88(424):1273–1283.

Rousseeuw, P. J. and Leroy, A. M. (2005). *Robust regression and outlier detection*, volume 589. John Wiley & Sons.

Rubel, F. and Kottek, M. (2010). Observed and projected climate shifts 1901–2100 depicted by world maps of the köppen-geiger climate classification. *Meteorologische Zeitschrift*, 19(2):135–141.

Russell, J. and Fraser, A. (1994). Infrared methods. In *Clay Mineralogy: Spectroscopic and chemical determinative methods*, pages 11–67. Springer.

SAGA, G. (2013). System for automated geoscientific analyses. *Online) http://www. sagagis. org/en/index. html*.

Savitzky, A. and Golay, M. J. (1964). Smoothing and differentiation of data by simplified least squares procedures. *Analytical chemistry*, 36(8):1627–1639.

Shibusawa, S., Anom, S. M., Hache, C., Sasao, A., Hirako, S., Stafford, J., Werner, A., et al. (2003). Site-specific crop response to temporal trend of soil variability determined by the real-time soil spectrophotometer. In *Precision agriculture: Papers from the 4th European Conference on Precision Agriculture, Berlin, Germany, 15-19 June 2003.*, pages 639–643. Wageningen Academic Publishers.

Simonescu, C. M. (2012). *Application of FTIR spectroscopy in environmental studies*. INTECH Open Access Publisher.

Sjögersten, S. and Wookey, P. A. (2004). Decomposition of mountain birch leaf litter at the forest-tundra ecotone in the fennoscandian mountains in relation to climate and soil conditions. *Plant and Soil*, 262(1-2):215–227.

Skjemstad, J. and Dalal, R. (1987). Spectroscopic and chemical differences in organic matter of two vertisols subjected to long periods of cultivation. *Soil Research*, 25(3):323–335.

Spielvogel, S., Prietzel, J., and Kögel-Knabner, I. (2007). Changes of lignin phenols and neutral sugars in different soil types of a high-elevation forest ecosystem 25 years after forest dieback. *Soil Biology and Biochemistry*, 39(2):655–668.

Stenberg, B. and Rossel, R. V. (2010). Diffuse reflectance spectroscopy for high-resolution soil sensing. In *Proximal Soil Sensing*, pages 29–47. Springer.

Stenberg, B., Viscarra Rossel, R. A., Mouazen, A. M., and Wetterlind, J. (2010). Chapter five-visible and near infrared spectroscopy in soil science. *Advances in agronomy*, 107:163–215.

Stevens, A. and Ramirez-Lopez, L. (2013). An introduction to the prospectr package. *R package Vignette R package version 0.1.3.*

Stuart, C. (2011). Robust regression.

Tarboton, D. G. (1997). A new method for the determination of flow directions and upslope areas in grid digital elevation models. *Water resources research*, 33(2):309–319.

Tenenhaus, M. (1998). *La régression PLS: théorie et pratique*. Editions Technip.

Terhoeven-Urselmans, T., Vagen, T.-G., Spaargaren, O., and Shepherd, K. D. (2010). Prediction of soil fertility properties from a globally distributed soil mid-infrared spectral library. *Soil Science Society of America Journal*, 74(5):1792–1799.

Travis, M. R., Elsner, G. H., Iverson, W. D., Johnson, C. G., et al. (1975). Viewit: computation of seen areas, slope, and aspect for land-use planning.

Tukey, J. W. (1960). A survey of sampling from contaminated distributions. *Contributions to probability and statistics*, 39:448–485.

Tukey, J. W. (1962). The future of data analysis. *The Annals of Mathematical Statistics*, pages 1–67.

Vågen, T.-G., Shepherd, K. D., and Walsh, M. G. (2006). Sensing landscape level change in soil fertility following deforestation and conversion in the

highlands of madagascar using vis-nir spectroscopy. *Geoderma*, 133(3):281–294.

Vapnik, V. N. and Vapnik, V. (1998). *Statistical learning theory*, volume 2. Wiley New York.

Viscarra Rossel, R., Walvoort, D., McBratney, A., Janik, L. J., and Skjemstad, J. (2006). Visible, near infrared, mid infrared or combined diffuse reflectance spectroscopy for simultaneous assessment of various soil properties. *Geoderma*, 131(1):59–75.

Viscarra Rossel, R. A., Chappell, A., De Caritat, P., and McKenzie, N. J. (2011). On the soil information content of visible–near infrared reflectance spectra. *European Journal of Soil Science*, 62(3):442–453.

Walvoort, D., Brus, D., and De Gruijter, J. (2010a). An r package for spatial coverage sampling and random sampling from compact geographical strata by k-means. *Computers & Geosciences*, 36(10):1261–1267.

Walvoort, D., Brus, D., de Gruijter, J., Walvoort, M. D., and Java, S. (2010b). Package 'spcosa'.

Wang, Y. and Witten, I. H. (1997). Inducing model trees for continuous classes. In *Proceedings of the Ninth European Conference on Machine Learning*, pages 128–137.

Wasserman, L. (2004). *All of statistics: a concise course in statistical inference*. Springer.

Webster, R. (1977). Canonical correlation in pedology: how useful? *Journal of Soil Science*, 28(1):196–221.

Wetterlind, J., Stenberg, B., and Rossel, R. A. V. (2013). Soil analysis using visible and near infrared spectroscopy. In *Plant Mineral Nutrients*, pages 95–107. Springer.

Whalen, J. K. and Sampedro, L. (2010). *Soil ecology and management*. CABI.

Wold, S., Sjöström, M., and Eriksson, L. (2001). Pls-regression: a basic tool of chemometrics. *Chemometrics and intelligent laboratory systems*, 58(2):109–130.

Yohai, V. J. (1987). High breakdown-point and high efficiency robust estimates for regression. *The Annals of Statistics*, pages 642–656.

Yohai, V. J. and Maronna, R. A. (1979). Asymptotic behavior of m-estimators for the linear model. *The Annals of Statistics*, pages 258–268.

Yurui, S., Lammers, P. S., Daokun, M., Jianhui, L., and Qingmeng, Z. (2008). Determining soil physical properties by multi-sensor technique. *Sensors and Actuators A: Physical*, 147(1):352–357.

Zhang, S., Huang, Y., Shen, C., Ye, H., and Du, Y. (2012). Spatial prediction of soil organic matter using terrain indices and categorical variables as auxiliary information. *Geoderma*, 171:35–43.

Ziechmann, W. (1964). Spectroscopic investigations of lignin, humic substances and peat. *Geochimica et Cosmochimica Acta*, 28(10):1555–1566.

Appendix A

Additional figures

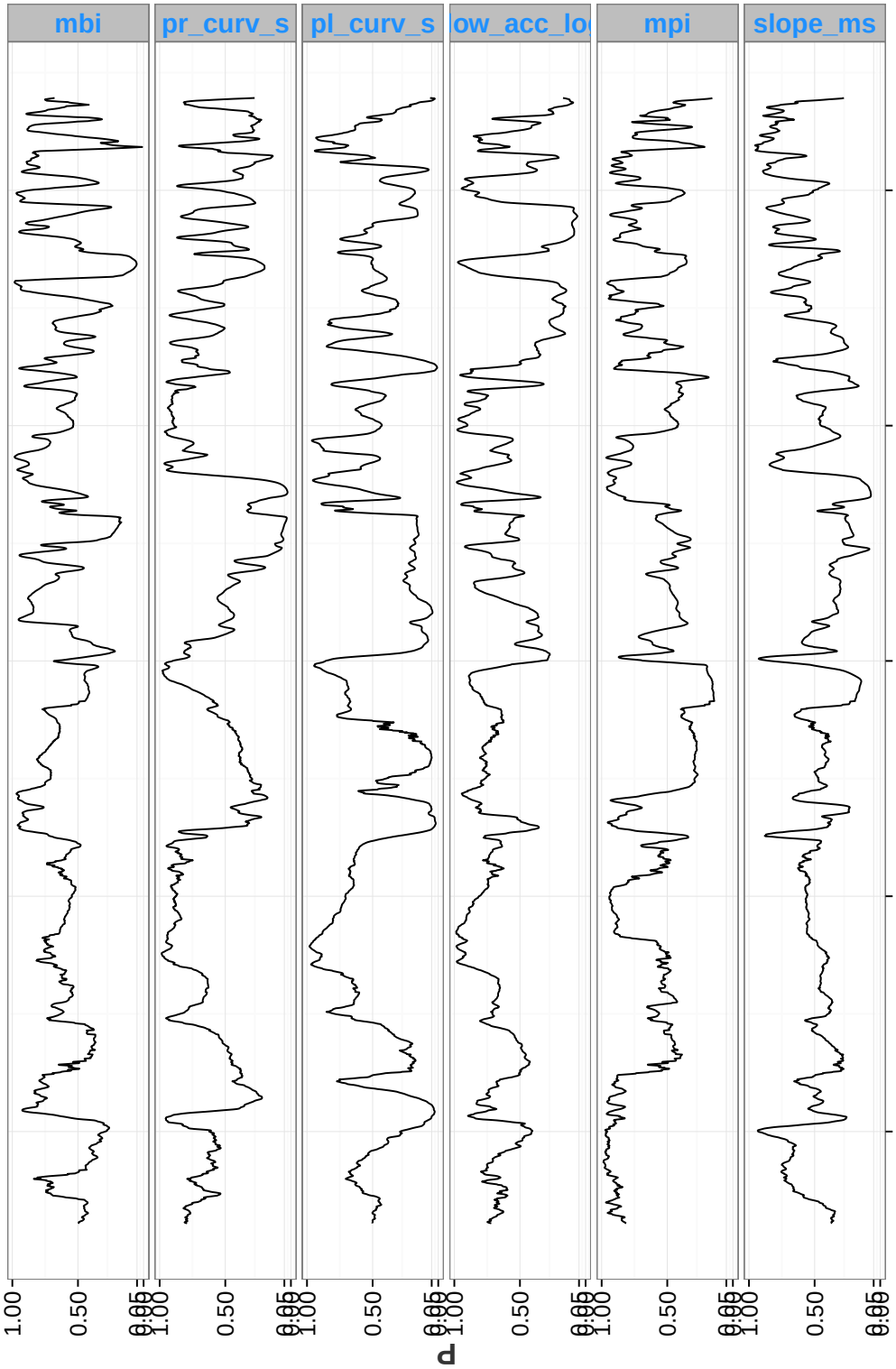


Figure A.1: P-value for the terrain attributes

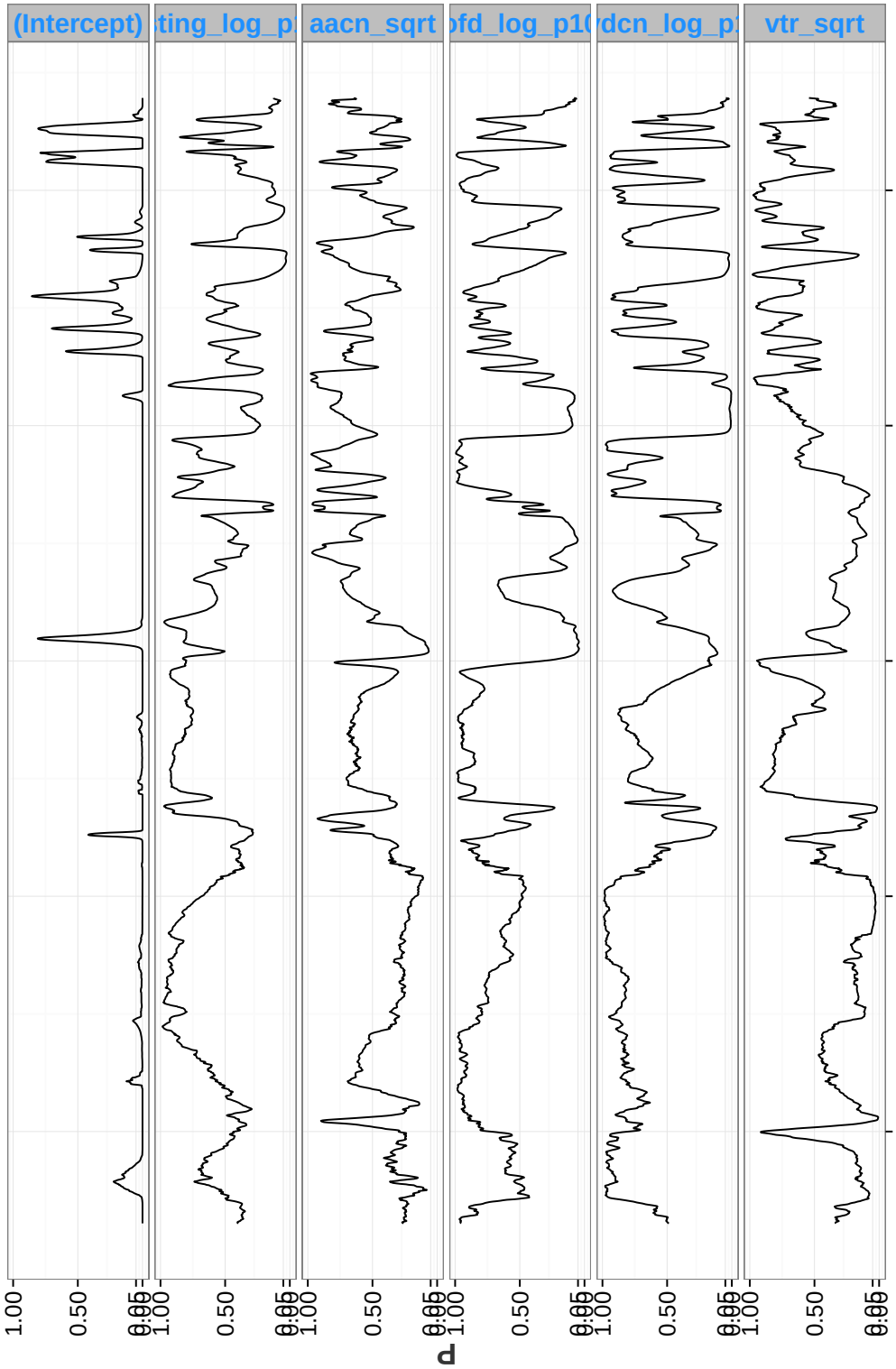


Figure A.2: P-value for the terrain attributes -2