Sampling Design and Machine Learning Optimization for the Application of Soil Sensing Data in Digital Soil Mapping

Dissertation

der Mathematisch-Naturwissenschaftlichen Fakultät

der Eberhard Karls Universität Tübingen

zur Erlangung des Grades eines

Doktors der Naturwissenschaften

(Dr. rer. nat.)

vorgelegt von

LEONARDO RAMIREZ-LOPEZ, M.Agr.

aus Fusagasugá, Kolumbien

Tübingen

2013

Tag der mündlichen Qualifikation:	08.02.2013
Dekan:	Prof. Dr. Wolfgang Rosenstiel
1. Berichterstatter:	Prof. Dr. Thomas Scholten
2. Berichterstatter:	Prof. Dr. Bas van Wesemael

Contents

1. Summary

The general aim of this thesis was to develop innovative methods to build and optimize empirical soil models based on soil sensing data. The combination of effective sampling schemes with geophysical sensing techniques is an active branch of soil scientific research. This approach aims to provide high resolution soil property data for flood forecasting and protection, agricultural management as well as for developing strategies to adapt to global climate change.

This thesis comprises four manuscripts. The first two manuscripts are dedicated to calibration sampling strategies. Sampling design is crucial in predictive modelling, since all results and interpretation are based on the selected samples. Hence, the first manuscript investigates the effect of the calibration set size and the calibration sampling strategy on the generalization error of visible and near infrared (vis–NIR) models. Furthermore, a method useful for identifying the optimal sample set size necessary for calibrating vis–NIR models of soil attributes is developed. Within the context of digital soil mapping, the second manuscript focuses on a comparison of different calibration sampling strategies for building predictive models of soil properties based on soil sensing. An improved version of the well-known conditioned Latin hypercube sampling algorithm, which is proposed in this manuscript, outperforms other approaches.

The third and fourth manuscripts are devoted to the development of novel methods and algorithms for dealing with large, heterogeneous and therefore complex soil sensing datasets. Generally in vis–NIR spectroscopy, there is a lack of methods for assessing the reliability of distance metrics for soil similarity analysis, required for building predictive models. In addition, the relationship between soil spectral similarity and soil compositional similarity has not been explored yet. For the third manuscript several distance metric algorithms for assessing the vis–NIR spectral similarity between soil samples are developed. The results show that some of the proposed algorithms outperform the standard methods significantly and adequately reflect the similarity in the compositional domain. The methods developed in the third manuscript are used in the fourth for developing an algorithm named spectrum based–learner (SBL). The SBL is inspired by memory–based learning (MBL). While a global target function may be very complex, MBL methods describe the target function as a collection of less complex local (or locally stable) approximations. The results presented in this manuscript show that in terms of predictive accuracy the SBL outperforms several other machine learning algorithms, which are usually employed in soil sensing.

2. Zusammenfassung

Ziel der vorliegenden Dissertation war die Entwicklung innovativer Ansätze zum Aufbau und zur Optimierung von Bodenprognosemodellen auf Basis geophysikalischer Naherkundungsdaten. Die Kombination effektiver Stichprobenverfahren mit geophysikalischen Naherkundungsverfahren stellt einen aktuellen Forschungszweig der Bodenkunde und der Geoinformatik dar, welcher kosteneffizient hochauflösende Bodeninformationen für die Anbauplanung, den Hochwasserschutz oder die Erarbeitung von notwendigen Anpassungen an den Klimawandel liefern kann.

Die Arbeit umfasst vier Manuskripte. Die ersten beiden behandeln die Entwicklung und den Vergleich von Stichprobenverfahren zum Aufbau von Prognosemodellen. Stichprobenverfahren stellen ein zentrales Glied im Rahmen der Bodenlandschaftsmodellierung dar, da alle weiteren Ergebnisse und Interpretationen auf der Auswahl der Stichprobe basieren. Im ersten Manuskript werden daher die Auswirkungen des Stichprobenumfangs und des jeweiligen Stichprobenverfahrens im Hinblick auf die Generalisierungsleistung von Modellen auf Basis von vis–NIR Spektroskopiedaten untersucht. Des Weiteren wird eine neue Methode zur Identifizierung des optimalen Stichprobenumfangs vorgestellt. Das zweite Manuskript behandelt neben der Einführung einer verbesserten Version des *Latin Hypercube Sampling*-Algorithmus schwerpunktmäßig den Vergleich von Stichprobenverfahren zum Aufbau von Bodeneigenschaftsmodellen mit Hilfe quasi-kontinuierlicher geophysikalischer Feldmessungen.

Das dritte und vierte Manuskript behandelt die Entwicklung neuer effizienter Modellierungsverfahren zur Bearbeitung großer, heterogener und komplexer vis–NIR Datensätze. In der vis–NIR Spektroskopie fehlen generell Ansätze zur Bewertung von Ähnlichkeitsmaßen, die die Grundlage für Prognosen darstellen. Darüber hinaus wurde der Zusammenhang zwischen der spektralen Ähnlichkeit und der Ähnlichkeit in der mineralischen und organischen Zusammensetzung der Böden bisher noch nicht untersucht. Im dritten Manuskript werden daher verschiedene Maße zur Abschätzung der spektralen Ähnlichkeit entwickelt und untersucht, die die Zusammensetzung des Boden adäquat widerspiegeln und im Vergleich bessere Ergebnisse als die bisher eingesetzten Standardmethoden liefern. Die vorgestellten Methoden dienen in der vierten Publikation der Entwicklung des sogenannten *Spectrum-Based-Learner*-Algorithmus (SBL). Hierbei handelt es sich um einen neuen leistungsfähigen Prognose-Algorithmus. Der SBL beruht auf der *memory-based learning*-Methode (MBL) und berücksichtigt auch die Ähnlichkeit in der Zusammensetzung der Böden. Dabei werden im Vergleich zu globalen Regressionsmodellen, viele jedoch weniger komplexe, lokaler Modelle erstellt. Die Ergebnisse zeigen, dass der SBL-Algorithmus anderen Regressionsverfahren überlegen ist.

3. General introduction

3.1 The role of soil information in ecosystem services

Soil is an essential source of ecosystem services such as food production and climate regulation (Sanchez *et al.*, 2009). Soil information is of fundamental importance for decision making on adequate land use planning and management and environmental protection which is in fact the motivation behind soil surveys (Rossiter, 2004).

Due to the rising levels of greenhouse gases, there is a great interest on monitoring the dynamics of gases such as CO_2 and NO_2 in the atmosphere. Soil is the most important global source of NO_2 (Bellamy *et al.*, 2005; Billings, 2008; Haygarth and Ritz, 2009) and also the largest terrestrial pool of C. Soil represents a key component in the global C cycle and has an important influence on the global CO_2 fluxes between terrestrial biosphere and atmosphere. Soil can be a source or sink of atmospheric C, therefore soil organic C monitoring is not only of fundamental importance for understanding the atmospheric C dynamics, but also for developing environmental policies.

Currently there is a growing demand for up-to-date soil information (Hartemink, 2008; McBratney *et al.*, 2006). This demand is especially critical for some vast areas of the globe (e.g. in the tropics) where the information about soils is very limited (Minasny and Hartemink, 2011). In general, these areas require massive information about soils for agricultural development and environmental sustainability issues.

One of the majors concerns in soil science relies on the fact that the conventional methods of soil analysis are too expensive and time-consuming, and soil legacy databases are often not adequate for assessing and mapping the soil condition (McBratney *et al.*, 2006; Minasny and Hartemink, 2011). In this sense, producing relevant soil information for improving the current soil legacy databases is one of the big goals of soil sensing and digital soil mapping (DSM). They can be used in a cost- and time- effective way for describing and monitoring the soil variability at different scales with high spatial and temporal resolutions. Furthermore,

these methods are important for bridging the gap between the digital revolution and soil science.

The GlobalSoilMap project (Sanchez *et al.*, 2009) aims to digitally map the global soil resources at a high spatial resolution in order to create solid basis on which end-users (e.g. agricultural extension workers, policy-makers, farmer associations, environmental extension services, agribusinesses, and nongovernmental and civil society organizations) can make decisions. In this project, soil sensing and DSM methods are being intensively used. For example, Odgers *et al.*, 2012, used DSM techniques and soil sensing products for creating maps of soil organic C, over the (contiguous) United States territory at a resolution of 100 m. However they acknowledge that there are still some methodological challenges which need to be addressed prior the full-scale production mapping.

On the other hand, despite the potential of soil sensing techniques for DSM has been demonstrated at the field-scale level; at regional or continental scales most of these techniques are not yet operational. In this respect, research efforts need to be addressed in the development of pedometric methods which deal the intrisic complexity problems of large and heterogeneous soil sensing datasets. Furthermore, adequate strategies for identifiying relevant data for modeling purpouses are also necesseray in order to improve the efficiency of soil sensing techniques in DSM.

3.2 Background of soil vis–NIR spectroscopy and sensing

Soil visible and near infrared (vis–NIR) spectroscopy is the study of the interactions between soil and electromagnetic radiation at wavelengths ranging from 400 nm to 2500 nm (Figure 1). Usually these interactions are studied in order to infer soil characteristics valuable in the assessment of the soil condition and to save additional labor costs.

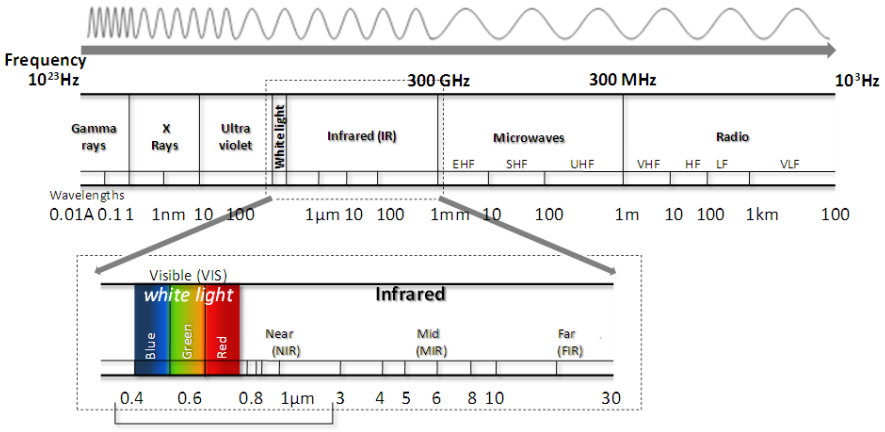

Figure 1. Regions of the electromagnetic spectrum.

Soil absorbs and reflects electromagnetic radiation as a function of its components, which in turn absorbs and reflects energy differently. For example, pedogenic oxides, silicates and carbonates usually exhibit contrasting spectral characteristics in the vis–NIR region as shown in Figure 2. However, since soil is a very complex mixture of mineral and organic constituents its vis–NIR characteristics are largely non-specific.

Soil visible and infrared spectra result from electronic and vibrational processes. In the visible region the spectral features are mainly due to electronic processes (which mostly occur in the ultraviolet region and rarely in the NIR region). Soil spectral features associated to electronic processes are related to minerals that contain iron (e.g. hematite, goethite). Despite fundamental vibration bands lie in the mid- and far-infrared regions, vibrational processes yield features in the NIR region due to the excitation of overtones and combination of tones of the fundamental modes of anion groups (e.g. OH, CO_3 and SO_4; Hunt and Salisbury, 1970). Therefore, soil constituents present weak, broad and in most of the cases overlapping and masking vis–NIR spectral responses. However, soil vis–NIR data contains important information regarding soil mineralogy. For instance, important diagnostic features of clay minerals due to the combination metal-OH bend plus OH stretch, can be found between 2200 nm and 2300 nm (Clark and Roush, 1984).

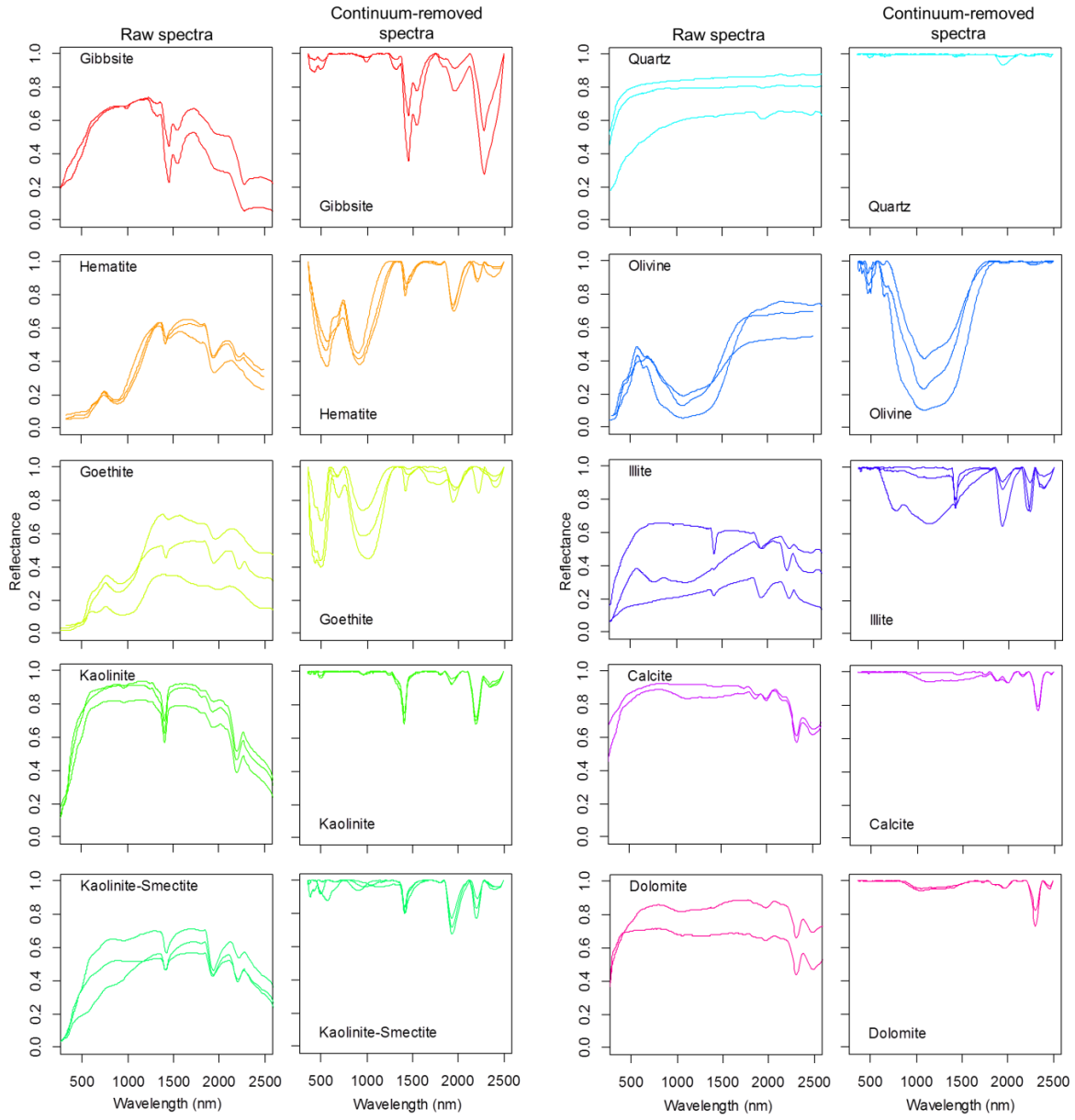

Figure 2. Spectra of samples of iron and aluminum oxides, silicates and sheet silicates and carbonates commonly present in soils. Columns 1, 3: non-preprocessed reflectance spectra. Columns 2, 4: Continuum-removed reflectance or band-depth normalized reflectance spectra.

On the other hand, despite the contrasting spectral characteristics between soil minerals, in most of the cases they present absorption features at the same spectral regions (Figure 2). Therefore, overlapping and masking effects between spectral features of minerals would be expected in a mixture of minerals. For example, in a mixture between hematite and goethite, overlapping would occur along the whole vis–NIR region. However, in such a case just some small spectral features of goethite and hematite (which are usually persistent in mineral mixtures)

would be determinant for their identification and quantification. On the other hand, in a mixture of olivine and quartz, the spectral features of olivine would not be affected since quartz is featureless (Figure 2).

The first studies on soil sensing in the visible region date back to 1958, when Kojima (1958a, 1958b) used a photo colorimeter to study the relationships between soil color and both moisture content and soil particle size. Later on, during the 60s two important works on soil spectroscopy were published. In both cases, the variation patterns of soil reflectance spectra as a function of several soil chemical and physical attributes were investigated. The first one corresponds to Obukhov and Orlov (1964) who studied the spectral reflectance of some soils in Russia. They suggested that soil reflectance spectroscopy could be used as a tool for soil survey. The second one corresponds to Bowers and Hanks (1965) who concluded that mineralogy, organic matter, particle size, moisture content are soil attributes that have a key influence on the absorption of radiated energy by the soil. After these seminal studies, several other important contributions to soil spectroscopy appeared during the 70s and 80s (e.g. Planet, 1970; Condit, 1970; Montgomery and Baumgardner, 1974; Stoner and Baumgardner, 1981; Coleman and Montgomery, 1987; Irons *et al.*, 1989; King and Clark, 1989).

During the 90s several researches began to address the problem of the quantification of soil attributes from soil visible and infrared data. For example, Ben-Dor and Banin (1990) modeled carbonate concentration in soils as a function of NIR features. Sudduth and Hummel (1991) compared several multivariate regression methods for calibrating soil organic matter to soil vis–NIR spectra. Palmborg and Nordgren (1993) used NIR data to model basal respiration in soils. Fritze *et al.* (1994) developed NIR models to predict microbial biomass and soil respiration. Ben-Dor and Banin (1995) used NIR data for simultaneous predictions of several soil attributes used in soil classification. Palacios-Orueta and Ustin (1996) used discriminant analysis to classify soil types according to their main vis–NIR features. Janik *et al.* (1998) discussed about the possibility to replace conventional routine analysis by using mid infrared reflectance analysis. Despite they did not included vis–NIR spectroscopy in their analyses, they stressed that soil spectros-

copy is in general a very powerful tool for enhancing soil information, especially in cases where spatially intensive sampling for soil analysis is needed. All these studies reported the possibility to predict accurately a wide range of soil attributes based only on the soil spectral characteristics. They also indicated that for extracting useful information from vis–NIR data (which is often complex), it is often necessary to use multivariate statistical methods (Viscarra-Rossel and Behrens, 2010). In addition, most of them pointed out some of the advantages that soil spectroscopy has over several conventional methods of soil analysis: it is time- and cost-efficient, non-invasive and non-destructive, requires minimal preparation of the sample, multiple soil attributes can be inferred from a single measurement of the spectral reflectance, etc. (McBratney *et al.*, 2006).

The promising results in terms of prediction performance and efficiency reported in soil spectroscopy studies triggered the development of soil vis–NIR spectral databases a.k.a soil spectral libraries. Nowadays, soil spectroscopy and soil spectral libraries have become powerful tools in soil science helping to analyze and store large amounts of soil information efficiently. Hence, the size of these databases has been increasing recently and some initiatives to create national and global spectral libraries emerged (e.g. Viscarra Rossel, 2009; Wetterlind and Stenberg, 2010; Terhoeven–Urselmans *et al.*, 2010).

3.3 Other soil sensing systems in the domains of the electromagnetic spectrum

Apart from vis–NIR spectroscopy, soil sensing also comprises the study of the interactions between soil and radiated energy at wavelengths belonging to other regions of the electromagnetic spectrum such as the mid infrared (2500 – 25000 nm), gamma-ray (<0.01 nm) and radio wave regions (1 mm – 100000 km).

Numerous studies have shown that soil mid infrared data can be used for predicting soil attributes accurately. Methods for soil mid infrared analysis have been developed in parallel to the development of vis–NIR spectroscopy. Moreover, in comparison to soil vis/NIR spectroscopy, models of soil attributes calibrated from

mid infrared data usually produce more accurate results (e.g. Viscarra Rossel *et al.*, 2006; Reeves, 2009). This is mainly due to the strong vibrational processes occurring in the mid infrared region. However, for in-situ applications, mid infrared spectroscopy still presents some drawbacks related with the portability and sensibility to external conditions of the mid infrared sensors (Reeves, 2009).

Soil gamma-ray spectroscopy is based on the natural radioactivity of the soil. It produces natural high frequency radiation in the gamma region of the electromagnetic spectrum. These soil gamma-rays present different energy levels and intensities. In the environment, K, U and Th are the only naturally occurring elements that produce gamma-rays of sufficient energy and intensity to be measured by gamma-ray sensors (Minty *et al.*, 1997). These radioisotopes produce well-defined peaks at specific areas of the gamma spectrum. They exhibit characteristic radioactive decay patterns which can be used to quantify their concentrations in the soil. These three radioisotopes have been present and continuously decaying (therefore their concentrations are continually decreasing) in rocks since their creation (Minty *et al.*, 1997). The concentration of these radioisotopes in soils strongly depends on pedogenic processes (Dickson and Scott, 1997). Therefore, from soil gamma-ray data it is possible to extract important information on quite specific soil/regolith properties (Wildford, 2012). Gamma-ray sensing has been used in soil science for soil-landscape formation and modeling (e.g. Stockmann *et al.*, 2012; Triantafilis, *et al.*, 2013), developing weathering indexes (e.g. Wildford, 2012), clay content modeling (e.g. Van Der Klooster *et al.*, 2011), etc. Since soil gamma-radiation is attenuated by bulk density and water content, it has also been used for modeling these soil attributes (e.g. Pires, 2009, de Groot *et al.*, 2009).

Electromagnetic induction (EMI) is another technique which has been largely used in proximal soil sensing. It is based on the measurement of the apparent soil electrical conductivity and works at the radio wave region of the electromagnetic spectrum. Its basic principle of operation is very simple: the system uses a transmitter coil which induces a specific electromagnetic field into the soil which generates a secondary electromagnetic field. This secondary field varies in inten-

sity depending on the soil. A receiver coil measures both the primary and secondary fields. The ratio between the primary and secondary fields is a linear function of the electrical conductivity (McNeill, 1992; Sudduth et al., 2001). It has been widely demonstrated that EMI can be used as an effective tool for assessing and mapping soil salinity (e.g. Cameron *et al.*, 1981; McKenzie *et al.*, 1989; Johnston *et al.*, 1996; Job *et al.*, 1999; Triantafilis *et al.*, 2000; Amezketa, 2007; McLeod *et al.*, 2010; Ganjegunte and Braun, 2011). Several researchers have also reported the use of EMI for predicting other soil attributes such as water content (e.g. Sheets *et al.*, 1995; Job *et al.*, 1999, Reedy and Scanlon, 2003) and soil texture (e.g. Hedley *et al.*, 2004).

Due to the efficiency benefits of soil sensing in the electromagnetic spectrum, this research area has been growing rapidly during the last years generating large and very complex volumes of data. However, pedometric techniques for dealing with such big and complex soil datasets have not been well explored yet. For example, search for useful information, data processing and analysis in big and complex datasets are challenging tasks in research efforts should be intensified.

3.4 Calibration sampling: size and predictor space coverage

Most of the soil sensing techniques have great potential for high resolution digital soil mapping because they are faster and more cost-effective compared to conventional methods (Bramley and Janik 2005; Kim *et al.*, 2009). For example, soil vis–NIR spectroscopy can be used as a tool for increasing the number of analyses (increasing the sampling density) and consequently the accuracy of digital soil maps without considerable increase in costs (Wetterlind *et al.*, 2010). In this respect, for one given study area in which vis–NIR data is available at high spatial resolution, it is possible to calibrate vis–NIR models of soil attributes by using a small but well designed set of soil spectral samples. Those models can be used to predict soil attributes efficiently over a large number of soil samples belonging to the area under study using only the soil vis–NIR spectra. This applies also for other soil sensing techniques such as electromagnetic induction gamma-ray spec-

troscopy. The question that frequently arises prior modeling soil sensing data is: how many observations (samples) should be included in a calibration set in order to efficiently produce relevant and generalizable (spatial) soil information from soil sensing models? In this context, selecting an adequate calibration set in terms of predictor space coverage and sample set size is important to ensure an accurate prediction performance. This is particularly fundamental when the number of samples that can be collected or analyzed is rather low, which in practice is often the case due to budget and/or time constraints.

Generally, a calibration sample set drawn from a population should reflect or cover the variability of the population. In this sense, two different approaches can be chosen for selecting calibration samples: the coverage of the predictor space or the coverage of the geographical space. If existing information is available in terms of relevant environmental covariates, the predictor space approach should be preferred over the geographical space approach (McKenzie and Ryan, 1999).

Concerning calibration sampling strategies for covering the predictor space, some methods such as fuzzy c–means–based sampling (de Gruijter *et al*., 2010), Latin hypercube sampling (McKay *et al*., 1979; Minasny and McBratney, 2006), Kennard-Stone sampling (Kennard and Stone, 1969) and response surface sampling (Lesch *et al*., 1995; Lesch, 2005) have been used in pedometrics research. Despite this, several works have shown that the strategies employed for covering the multivariate space can lead to different levels of prediction accuracies (e.g. Siano and Goicoechea, 2007; Rodionova and Pomerantsev 2008; Fu *et al*., 2011).

Since soil sensing is a crucial step in digital soil mapping, research on both sampling strategies and strategies for identifying adequate calibration set sizes have not received enough attention (Grinand *et al*., 2012; Kuang and Mouazen, 2012). In principle, the optimal calibration set size may vary depending on the soil variability of the area under study (Kuang and Mouazen, 2012). Due to this, strategies for identifying the optimal calibration set size based only on the predictive information, (i.e. without an explicit prior knowledge on the soil attributes) are of great importance for the practical application of soil sensing techniques at the field scale. On the other hand, Minasny and McBratney (2010) stress the im-

portance to investigate the relation between the calibration sampling strategy and the prediction accuracy of soil models.

3.5 Complexity in soil vis–NIR datasets

The suitability and reliability of soil spectroscopy at the field-scale level has been already demonstrated by several researchers. This is mainly due to the fact that often at the filed-scale level the soil variability in terms of mineralogical and organic matter compositions (which strongly affect the soil spectral features) is low. For instance, the soil particle size effect on the soil spectral features could be different in two areas with contrasting soil mineralogical composition. On the other hand if the mineralogical composition is rather similar, it is expected that the soil particle size effect on the spectral features will also be similar. This also explains why soil vis–NIR models calibrated from land-scape scale databases usually presents lower predictive performance in comparison to vis–NIR models calibrated from field-scale databases. Furthermore, for continental- and global-scale datasets the predictive performance of vis–NIR models usually do not reach an accuracy level required for practical applications (e.g. Brown *et al.*, 2006).

Overall, the accuracy of vis–NIR models usually decreases when the dataset contains very diverse samples in terms of geographical origin, mineralogy, parent material, environmental conditions, etc. Table 1 summarizes the results reported in several papers on vis–NIR modeling for clay content and soil organic carbon. The lowest predictive errors are reported for field-scale studies while for regional-scale studies the results are diverse. Furthermore, the reported prediction error presents a positive correlation with the standard deviation of the soil attribute in the calibration set used in the studies (Figure 3). Stenberg *et al.* (2010) observed the same effect of the standard deviation on the error. This tendency, suggest that soil variability affects the complexity of soil vis–NIR datasets.

Table 1. Main results reported in different papers on vis–NIR modeling for clay content and soil organic carbon using partial least squares regression (the standard regression method in soil spectroscopy). Summary statistics correspond to the statistics of either, the calibration o set or the whole set used in each study. *n* indicates the number of samples used for validating the models. The root mean square error (RMSE) and the R^2 correspond to the results of the validation of the predictions.

Source	**Scale**	***N***	**Mean**	**S.D.**	**Min.**	**Max.**	**R^2**	**RMSE (%)**
			– Soil organic carbon (%) –					
Kuang and Mouazen (2011)	Field	62	1.48	0.20	1.06	2.16	0.12	0.19
Sudduth *et al.* (2010)	Field	74	1.22	0.20	-	-	0.55	0.13
Wetterlind and Stenberg (2010)	Field	58	2.30	0.20	1.80	2.80	0.70	0.12
Viscarra Rossel *et al.* (2006)	Field	116	1.34	0.28	0.81	1.98	0.60	0.18
Cañasveras *et al.* (2012)	Regional	55	0.85	0.33	0.09	1.96	0.77	0.18
This thesis (Manuscript 4)	Regional	1050	0.64	0.39	0.06	4.00	0.48	0.28
Wetterlind and Stenberg (2010)	Field	81	1.80	0.40	1.20	3.40	0.57	0.27
Summers *et al.* (2011)	Regional	228	1.50	0.53	0.31	2.90	0.57	0.35
Cambule *et al.* (2012)	Regional	137	0.90	0.60	0.00	2.70	0.65	0.37
Wetterlind and Stenberg (2010)	Field	112	2.30	0.60	1.30	4.50	0.85	0.22
Kuang and Mouazen (2011)	Field	38	1.38	0.70	0.70	3.51	0.75	0.30
Sarkhot *et al.*, (2011)	Field	154	1.04	0.76	0.08	3.72	0.86	0.29
Wetterlind and Stenberg (2010)	Field	65	4.20	0.80	3.10	8.20	0.71	0.53
Northup and Daniel (2012)	Field	139	2.49	0.87	1.45	4.87	0.86	0.30
McCarty *et al.* (2002)	Regional	60	7.10	1.10	4.30	8.80	0.82	0.55
Kuang and Mouazen (2011)	Regional	122	1.50	1.20	0.70	14.43	0.83	0.54
Nocita *et al.* (2012)	Regional	36	1.35	1.25	0.18	6.03	0.87	0.33
Leone *et al.* (2012)	Regional	93	1.54	1.46	0.04	13.50	0.80	0.47
Stenberg (2010)	Regional	50	3.00	1.60	0.40	6.90	0.71	0.88
This thesis (Manuskript 4)	World	900	1.1	1.93	0.00	45.80	0.50	1.08
Viscarra Rossel and Behrens (2010)	Regional	1104	2.5	2.28	0.01	13.90	0.82	0.96

Kuang and Mouazen (2011)	Field	21	1.74	2.50	0.74	14.43	0.96	0.62
				– Clay content (%) –				
Sudduth *et al.* (2010)	Field	13	17.50	2.40	-	-	0.15	2.68
Viscarra Rossel *et al.* (2006)	Field	118	14.23	3.04	8.00	24.14	0.60	1.91
Wetterlind and Stenberg (2010)	Field	61	45.00	5.00	37.00	58.00	0.61	3.50
Wetterlind and Stenberg (2010)	Field	65	28.00	5.00	11.00	34.00	0.50	3.60
Summers *et al.* (2011)	Regional	237	16.32	5.42	4.97	35.98	0.66	3.13
Wetterlind and Stenberg (2010)	Field	112	46.00	8.00	25.00	66.00	0.82	3.70
Wetterlind and Stenberg (2010)	Field	81	24.00	9.00	12.00	52.00	0.81	4.30
Genot *et al.*, (2010)	Regional	150	21.2	9.65	1.50	70.60	0.71	6.74
Cañasveras *et al.* (2012)	Regional	55	27.50	9.70	5.00	72.00	0.80	4.07
This thesis (Manuscript 4)	Regional	1050	23.51	12.48	1.00	81.10	0.78	6.10
Leone *et al.* (2012)	Regional	93	28.20	13.14	0.48	66.33	0.82	5.29
Stenberg (2010)	Regional	50	24.00	18.00	0.00	67.00	0.89	5.38
Viscarra Rossel and Behrens (2010).	Regional	1104	33.95	18.84	2.80	79.20	0.83	7.70
This thesis (Manuscript 4)	World	900	33.08	22.49	0.00	96.80	0.71	12.95

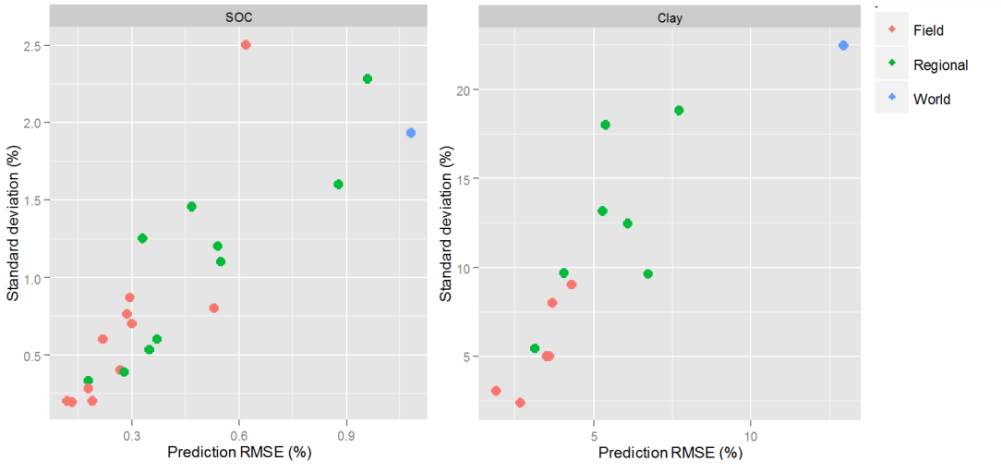

Figure 3. Reported root mean square error (RMSE) of vis–NIR based predictions against the standard deviation of the soil attribute in the calibration sets. Right: clay content. Left: soil organic carbon (SOC). Created from data presented in Table 1.

One reasonable approach for reducing the complexity of a given soil vis–NIR dataset (X), which is very heterogeneous, is to split X into *c* partitions or clusters, so that samples in the same partition share similar soil characteristics. In this sense, the complexity in each partition must be lower than the global complexity contained in X. In general in soil science and specifically in soil spectroscopy, several studies have demonstrated that models based on (either spectrally or geographically) local partitions perform better than single or global models. In many cases the use of geographical information for partitioning a spectral dataset results in reduction of the soil (spectral) variability within each partition in comparison to the global soil (spectral) variability. Stevens *et al.* (2010) observed that vis–NIR local models of soil organic carbon perform better than global models when the soil dataset is partitioned into different soil texture classes and agro–pedological regions. They also showed that the organic carbon variability within each partition is lower than the organic carbon variability of the entire area. Guerrero *et al.* (2010) used different regional calibration sets for predicting soil attributes in each region. For modeling soil attributes in different agricultural fields, Wetterlind and Stenberg (2010) used models calibrated with a national soil vis–NIR library, and models calibrated only with local samples taken from the fields under study. They observed that the local models outperformed the national soil vis–NIR models. Janik *et al.* (2007) suggested that local calibrations of soil spectroscopic models based on the minimization of changes in soil type may be more accurate than global calibrations. Similar conclusions are reported on the analysis of soil data for digital soil mapping. When the variability patterns of a given soil attribute differs between geomorphological or pedological regions, they should be modeled separately (McBratney *et al.*, 1991; Schmidt *et al.*, 2010).

In this respect memory–based learning (MBL, Mitchel, 1997) offers a plausible approach for solving such problems. In contrast to the commonly used approaches for modeling vis–NIR spectra such as partial least squares (PLS), principal component regression (PCR), support vector machines (SVM), decision trees (DT), artificial neural networks (ANN), etc., the MBL approaches do not derive an explicit global function or model. In MBL for each new sample or spectrum from which a given attribute has to be predicted, a certain number of spectrally simi-

lar samples are searched in a reference set (memory) and retrieved for calibrating one spectral model in order to predict the attribute specifically for the new sample. Put in another way, since it is possible to derive numerous soil characteristics from the soil vis–NIR features, by using an adequate similarity/dissimilarity measure is possible to retrieve samples from a reference set (e.g. a soil vis–NIR library) which share similar vis–NIR features and therefore similar compositional characteristics (e.g. mineralogy, organic matter composition). If the new sample is actually similar to the retrieved samples, then inferences about the new sample can be done by using the retrieved samples. Apart from being a very coherent strategy for soil vis–NIR modeling, MBL can also be viewed as one strategy for managing soil spectral libraries for soil inference tasks since eventually a small subset of the entire library is used i.e. library samples which are not similar to the new samples are ignored.

Two typical examples of MBL are the *k*–nearest neighbor algorithm which has been widely used in several research fields and locally weighted regression (LWR, Naes *et al.*, 1990) which has been applied specifically in vis–NIR spectroscopy. In the literature MBL is also referred to as local modeling, nevertheless local modeling comprises other approaches such as cluster–based modeling and geographical segmentation–based modeling, etc. Hence, MBL is one type of local modeling.

In general, modeling soil attributes using large and diverse soil vis–NIR libraries still remains a challenging task, and methods for dealing with complexity problems in vis–NIR datasets are necessary. Such methods would be useful for turning large-scale vis–NIR libraries in operational tools.

4. Objectives

The objectives of the investigations carried out throughout this doctoral thesis are divided according to two main research topics which are: *i.* calibration sampling in soil sensing datasets and *ii.* soil vis–NIR modeling in complex datasets using similarity/dissimilarity search methods.

For calibration sampling the main objectives were:

- Investigate on the effect of both the calibration set size and the sampling algorithm on the predictive performance of soil vis–NIR models.
- Analyze the sample predictor space coverage on the basis of different calibration sampling algorithms.
- Propose a straightforward method for identifying the optimal calibration set size for modeling soil sensing data in linear datasets.
- Evaluate the interaction between different machine learning methods and sampling algorithms and its effect of the accuracy of soil models based on sensing technologies.
- Evaluate and compare the effect of different calibration sampling strategies on digital soil maps produced by using data predicted from soil sensors.

For soil vis–NIR modeling in complex datasets using similarity/dissimilarity search methods, the main objectives were:

- Investigate the relationship between soil compositional similarity and soil vis–NIR similarity.
- Explore and develop suitable approaches for performing similarity/dissimilarity measurements between samples in soil vis–NIR datasets.
- Provide a method to evaluate the reliability of soil vis–NIR distance measurements.
- Introduce new methods for measuring soil vis–NIR distances.
- Develop a high-performance memory–based learning algorithm for modeling complex soil spectral data using reliable soil spectral similarity/dissimilarity measures.

5. Results and discussion

The main results and conclusions presented in this section are subdivided according to the four manuscripts that comprise this doctoral thesis. The first two man-

uscripts are dedicated to calibration sampling. In the first one, the effect of the calibration set size and the calibration sampling strategy on the generalization error of laboratory vis–NIR models is investigated. One of the most relevant contributions of this manuscript is the development of a straightforward method useful for identifying optimal calibration set sizes. The second manuscript focuses (in the context of digital soil mapping) on calibration sampling strategies for proximal soil sensing (EMI and gamma-rays) data modeling at very low calibration set sizes.

The third and fourth manuscripts are devoted to the development of novel methods and algorithms for dealing with complexity in vis–NIR datasets. The methods developed for the third manuscript are used in the fourth manuscript for developing a new high performance algorithm for modeling vis–NIR data.

6.1 Calibration sampling

6.1.1 Set size and methods

(Manuscript 1, Geoderma, submitted on July 2012)

This manuscript presents the results of the investigation on the effect of the calibration set size and three different calibration sampling strategies on the error of vis–NIR models. Furthermore, analyses of representativeness for identifying optimal calibration set sizes were also carried out. The calibration sampling strategies were based on the following sampling algorithms:

- Kennard-Stone (KSS, Kennard and Stone, 1969).
- Fuzzy c–means–based sampling (FCMS, de Gruijter *et al.*, 2010).
- Conditioned Latin hypercube (cLHS, Minasny and McBratney 2006).

For carrying out this study, two soil vis–NIR datasets from Brazil were used. The first one corresponds to a field-scale dataset which covers an area of 5 km^2. This dataset comprises samples collected at two fixed depths in points distributed on a spatially dense grid with 459 nodes (Figure 4).

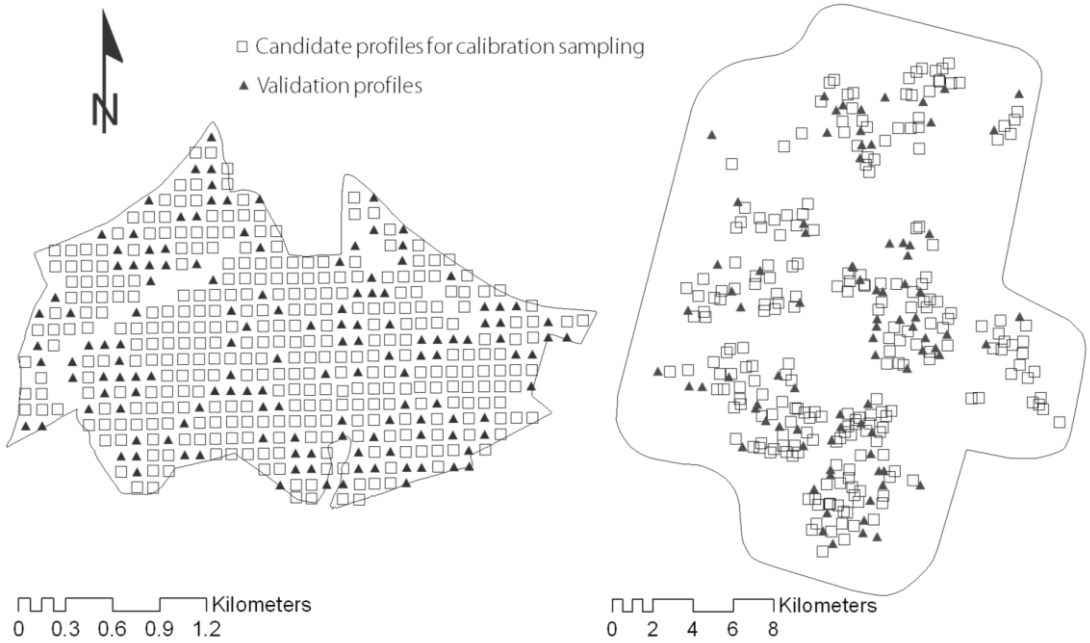

Figure 4. Spatial distribution of both the candidate profiles for calibration sampling and the validation profiles in the field scale dataset (left) and regional scale dataset (right).

The second dataset is a regional-scale dataset which comprises an area of approximately 300 km^2. It includes 318 profiles sampled at three fixed depths (Figure 4). These datasets were selected for this study due to the high variability in terms of soil types. In both cases, the main driving factor of soil variability should be topography. The clay minerals found in these areas are relatively constant. Due to this, the datasets are not expected to be strongly affected by non-linear relationships between soil vis–NIR features and soil mineralogy.

The soil attributes for which soil vis–NIR models were calibrated were clay content and exchangeable calcium (Ca^{++}).

The three sampling algorithms were used separately in each dataset to select a given number of samples which were then used to calibrate models of the studied soil attributes. This process was repeated several times varying the number of selected samples from 10 to 380. These models were used to predict the target soil attributes in an independent set of samples. Finally the prediction errors were compared at the different calibration set sizes.

The results found showed that the error of the soil vis–NIR models depends on the calibration set size. Particularly for low calibration set sizes the errors are higher. This is probably due to insufficient coverage of the predictor space. In this respect, when the number of calibration samples is relatively low the sampling algorithm plays a critical role on the accuracy of the vis–NIR models.

The highest training errors were returned by the KSS. However this algorithm tends to select samples with a wider range of soil attribute values in comparison to the cLHS and the FCMS algorithms. This is due to the fact the KSS selects extreme samples. In this sense, it is possible that the inclusion of extreme samples in the calibration set can be beneficial when the dataset does not contain outlier samples. On the other hand, in the case of proximal soil sensing measurements where many outlier samples can arise (due to uncontrolled conditions) the KSS would not be a good choice since outlier samples would be included in the calibration set. In this case, FCMS or cLHS should be preferred over the KSS algorithm. In order to illustrate the outlier sensitivity of the three algorithms, two synthetic grids comprising two variables (x1 and x2) were created (Figure 5). The first grid does not contain outliers while the second grid contains one. The sampling algorithms were used for selecting 9 samples. The KSS and FCMS show uniform coverage of the grid while the coverage of the cLHS points is irregular. However, the KSS selected the outlier present in the second grid. For the KSS Figure 5 shows that the inclusion of one outlier in the grid produces a displacement of the sampling locations of three points. Despite the FCMS algorithm includes a random initialization (for the selection of the centers of the clusters), comparing the locations at which the FCMS points are placed in both datasets, the distribution patters are very similar. This means that the FCMS, is not affected by the inclusion of the outlier. In the case of the cLHS, despite the distributions of the points is different in both grids (which is due to the steps related with the random search of samples in the cLHS algorithm), the outlier sample is not selected.

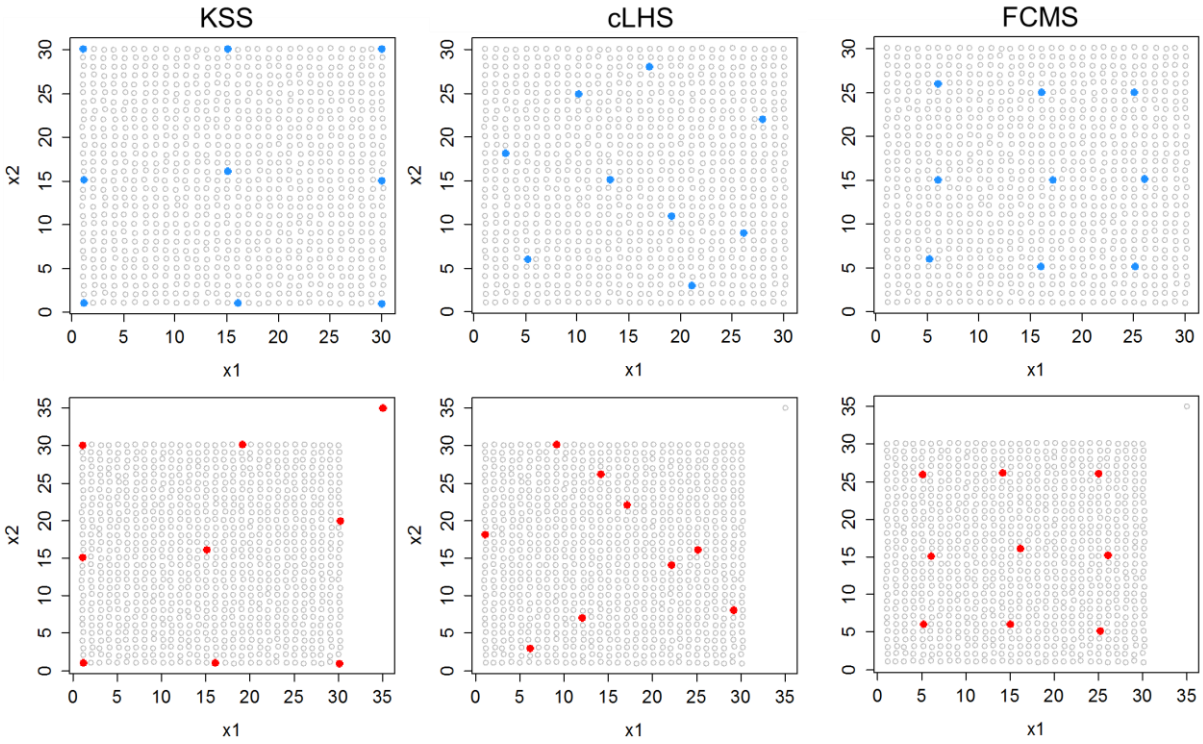

Figure 5. Distribution of nine sampling points selected by the sampling algorithms (KSS, FCMS and cLHS) on two synthetic grids formed by two variables (x1 and x2). The first grid does not contain outlier points (top) and the second grid contains one outlier point (bottom).

On the other hand, the sample representativeness on the basis of the above sampling strategies was also analyzed. In this respect, a straightforward method for identifying the optimal calibration set size based only on the analysis of the vis–NIR data (i.e. without prior knowledge on the soil attributes) was proposed in this thesis. It consists in comparing the statistics of the sample set against the (original) statistics of the population at different calibration set sizes. This is carried out in the standardized principal component (PC) space of the soil sensing data. The sample mean ($\overline{x}$) and the sample variance (s^2) of the PC variables are compared to the original mean (μ) and the original variance (σ^2) of the PCs. Note that σ^2 and μ, are equivalent to 1 and 0 respectively since the PC variables previously standardized to zero mean and unit variance. Both the absolute difference between variances ($|s^2 - \sigma^2|$) and the absolute difference between means ($|\overline{x} - \mu|$) were computed as (eqs. 8 and 9):

$$|s^2 - \sigma^2| = |s^2 - 1| = \left|\frac{1}{k}\left(\sum_{j=1}^{k} s^2_{\mathrm{pc}\,j}\right) - 1\right|; \quad (1)$$

$$|\overline{x} - \mu| = \left|\frac{1}{k}\left(\sum_{j=1}^{k} \overline{x}_{\mathrm{pc}\,j}\right) - 0\right|, \quad (2)$$

where $s^2_{\mathrm{pc}\,j}$ and $\overline{x}_{\mathrm{pc}\,j}$ are the sample variance and the sample mean of the jth PC, and k is the total number of PCs retained in the analysis.

It was found for all the algorithms that the original distribution of the vis–NIR data in the principal component (PC) space can be better replicated by increasing the calibration set size (Figure 6). The results showed that the samples selected by the cLHS and the FCMS algorithms better replicate the original distribution of the PCs in comparison to those selected by the KSS algorithm. For low calibration set sizes the cLHS better replicated the original distribution of the PCs in comparison to the FCMS. However at calibration set sizes ≥ 130 the cLHS and the FCMS produced comparable results.

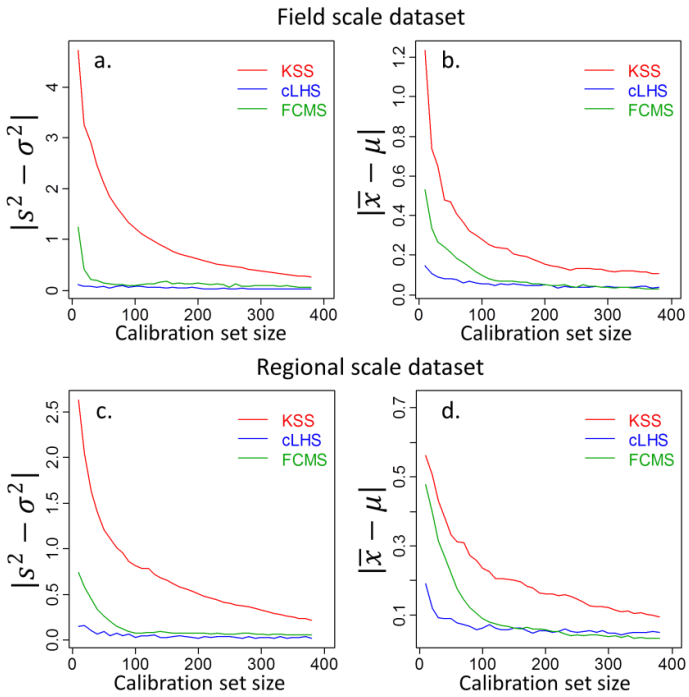

Figure 6. Calibration set size against the absolute difference between the sample variance (s^2) and the original variance (σ^2); and absolute difference between the sample mean ($\overline{x}$) and the original mean (μ).

The reason why cLHS reproduces adequately the statistics of the population is because the selection of samples is based on the probability distribution of the variables and not on the distances between points. Figure 7 shows an example of the localization of the 9 points selected from the synthetic grid (without outliers) in Figure 5 in the cumulative probability of the variables (x1 and x2). The cLHS algorithm selects samples which cover well the cumulative probability of the var-

iables, while the samples selected by KSS and FCMS are clustered at three locations in the cumulative probabilities. A good coverage of the probability distribution of the variables ensures a good representation of the original statistics of the population.

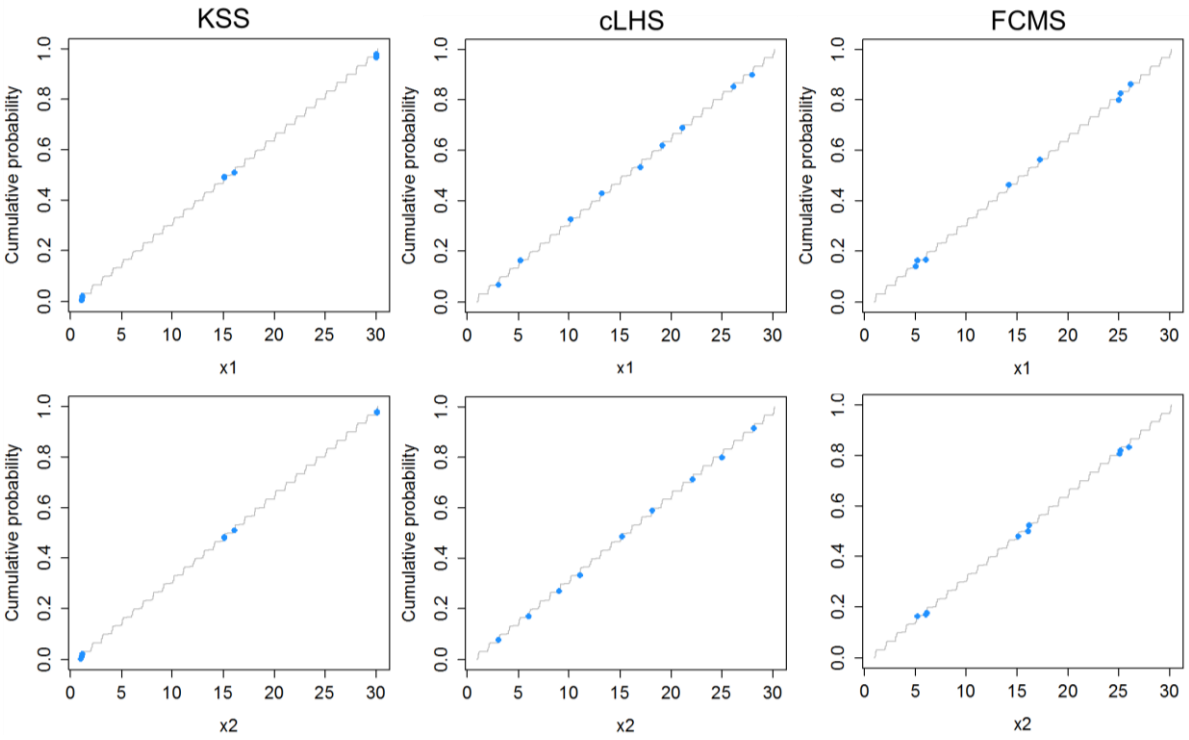

Figure 7. Distribution of nine sampling points selected by the sampling algorithms (KSS, FCMS and cLHS) on the cumulative probability of the variables x1 (top) and x2 (bottom) of the synthetic grid in Figure 5 without outliers.

Overall, the comparison between the distribution of the calibration set and the original distribution of the population of samples is an adequate strategy for identifying an optimal calibration set size based only on the predictive information. Furthermore, for the calibration of models it can be beneficial to select a calibration sample set whose distribution is close or equal to the distribution of the population.

6.1.2 Sampling for digital soil mapping at field scale

(Manuscript 2, Geoderma, submitted on June 2011)

This study was developed in the context of proximal soil sensing modeling for digital soil mapping. For this manuscript, the generalization error of predictive proximal soil sensing models was studied on the basis of three different calibration

sampling algorithms: *i.* A weighted conditioned Latin hypercube sampling with extremes (wecLHS) which is a modified version of the original algorithm proposed by Minasny and McBratney (2006); *ii.* Fuzzy c-means sampling (FCMS, Gruijter *et al.*, 2010); and *iii.* Response surface sampling (RSS, Lesch, 2005).

The interaction of these three sampling algorithms with other components of digital soil mapping was jointly analyzed in this study. These components are: *i.* Calibration set size; *ii.* Regression strategy; and *iii.* Estimation of the generalization error of predictive models based on cross-validation strategies. These components are summarized as follows:

- Relative small calibration set sizes:
    - For wecLHS, $n = 29$.
    - For FCMS, $n = 28$.
    - For RSS, $n = 20$.
- Two different regression approaches.
    - Multiple linear regression (MLR).
    - Random forest regression (RF) .
- Four cross-validation strategies based on resampling:
    - 10-fold cross validation (10cv).
    - Leave-group-out cross-validation (lgocv).
    - Bootstrapping (boot).
    - Bootstrapping 632 (.632boot).

The study area comprises 0.36 km² and it is located in Dessau-Rosslau near to the Elbe River in Saxony-Anhalt, Germany. Despite the low topographical variation within the study area, there is a high variation in soil texture.

A mobile geophysical platform equipped with both a frequency-domain electromagnetic-induction sensor and a portable gamma-ray spectrometer was used to perform on-the-go (proximal) soil sensing measurements within the study area. The use of this platform resulted in a (spatially) very dense set of soil sensing data. In order map the measured soil sensing variables (predictive data), they were spatially interpolated using kriging. A total of three different electrical con-

ductivity maps were produced as well as maps of K, Th and U derived from the gamma-ray measurements.

The maps of predictive variables were used to identify the sampling points according to each calibration sampling algorithm. Soil samples were collected from the identified sampling locations at two depths (0-10 cm and 10-30 cm). They were submitted to soil organic carbon (OC), pH and particle size (sand, silt and clay contents) analyses. The predictive data at these sampling points was extracted from the maps and used as input data for the two regression algorithms tested. Predictive models of OC, pH and particle size at each sampling depth according to each sampling algorithm were developed.

In order to estimate the generalization error of predictive models, the four cross-validation strategies were used. Furthermore, each model calibrated with the sample set selected by each calibration sampling algorithm, was used to predict the selected soil attributes in the calibration sets selected with the two remaining calibration sampling algorithms (i.e. they were used as independent validation sets). Figure 8 summarizes the methodology followed in this study.

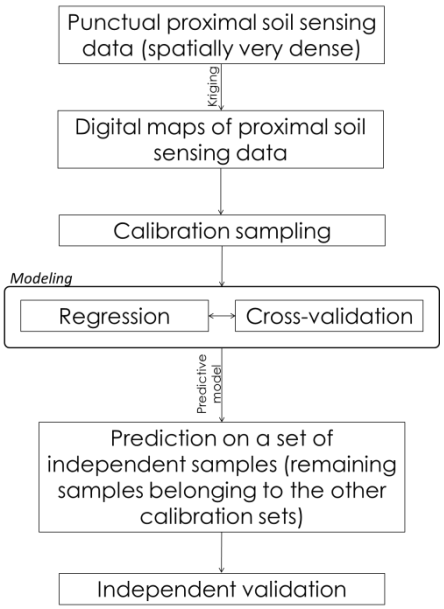

Figure 8. Methodological framework employed in the study.

In general, the wecLHS sampling algorithm proposed in this study in combination with Random Forests produced the best results in terms of soil predictive accuracy. Furthermore, in terms of accuracy, the difference between the results produced by the calibration sampling algorithms is larger than the difference between the results produced by the regression algorithms. In other words, the calibration sampling strategy presented a higher impact on the predictive performance of the soil models than the regression algorithms.

The results also showed that high predictive accuracy can be obtained with a small calibration set size as long as a reliable sampling strategy is employed. For example, in this study, only 20 samples are enough for explaining 70 % of variance of pH, SOC, and soil texture.

The best estimates of the generalization error of predictive models were produced by the LGOCV and the .632boot strategies. Nonetheless, the use of independent validation sets for further estimations of the uncertainty is recommended since resampling validation can result in either under- or over estimation of the accuracy.

6.2 Soil similarity and complexity in vis–NIR datasets

6.2.1 Distances and similarity search

(Manuscript 3, Geoderma, accepted on August 2012, doi: 10.1016/j.geoderma.2012.08.035)

In spectroscopy in general, there is a lack of methods for assessing the reliability of distance metrics. For this manuscript nine distance metric algorithms for assessing the vis–NIR spectral similarity/dissimilarity between soil samples were evaluated. They are as follows:

- Euclidean distance (ED).
- Mahalanobis distance (MD).
- Spectral angle mapper (SAM).
- Spectral information divergence (SID).

- Spectral difference surface (SDS) distance.
- Principal component Mahalanobis (PC–M) distance
- Optimized principal component Mahalanobis (oPC–M) distance.
- Locally linear embedding Mahalanobis (LLE–M) distance.
- Sigma locally linear embedding Mahalanobis (σLLE–M) distance.

The SDS, oPC–M and σLLE–M correspond to novel methods. The ED, MD, SAM, SID and SDS operate directly in the spectral vis–NIR space, while PC–M, oPC–M, LLE–M and σLLE–M use the Mahalanobis distance in a low dimensional space with uncorrelated variables derived from the original (and highly correlated) vis–NIR data. The PC–M corresponds to the standard method in soil spectroscopy for computing distances between vis–NIR spectra.

For this manuscript, a novel method for evaluating the reliability of the similarity/dissimilarity measures was also proposed. It is based on a nearest neighbor search. Apart from the vis–NIR data, this method also uses side or compositional information, i.e. information about one soil compositional variable which is available for a group of samples. It is assumed that there is a correlation (or at least an indirect or secondary correlation) between this side information and soil spectra. In other words, this approach is based on the assumption that the similarity measures between the spectra of a given group of soil samples should be able to reflect their similarity also in terms of the side information (e.g. compositional similarity). The side information approach works as follows:

1. A distance matrix is derived from the spectral matrix X. This spectral matrix has a side information Y.
2. By using the distance matrix, for each sample in X select its closet (most spectrally similar) sample.
3. A comparison between the side information of each sample in X and the side information of its corresponding closest sample is performed. The statistics of these comparisons is evaluated in order to assess the reliability of the distance metric algorithm.

Figure 9 summarizes the side information approach for evaluating the distances computed between samples in a spectral matrix Xu and samples in a spectral matrix Xr.

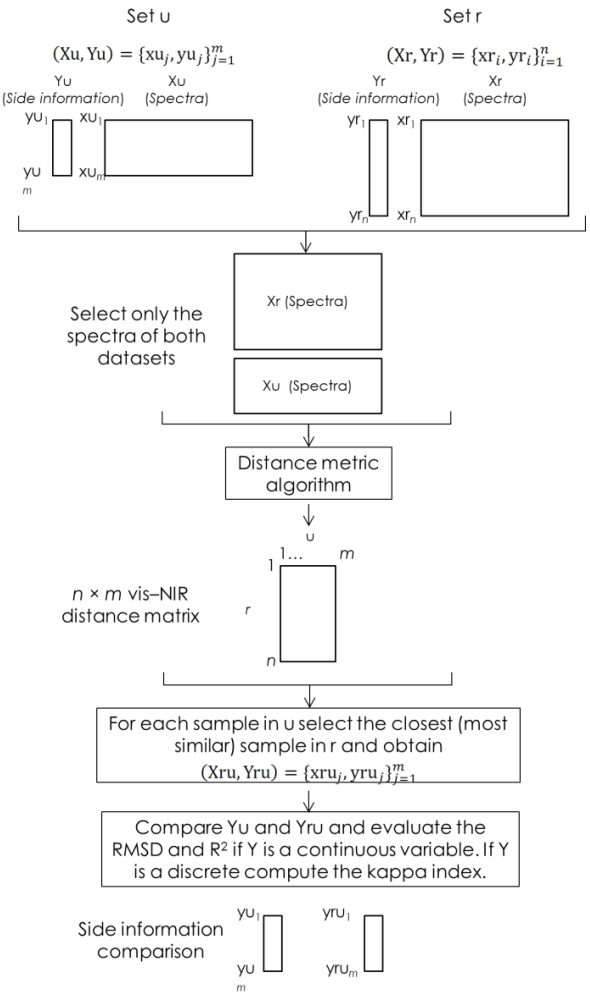

Figure 9. Methodological framework of the side information approach for soil compositional similarity search using soil vis–NIR distances. In the side information comparison step, the root mean square of differences (RMSD) is used if the side variables are continuous. Optionally, if the side variables are discrete the kappa index is used. For this manuscript clay content and pH were used as side information. Therefore the RMSD was used as measure of similarity.

In order to test the distance metric algorithms, they were used in a global soil spectral library (GSSL) developed by the World Agroforestry Centre (ICRAF) and the ISRIC - World Soil Information (2010). From the GSSL a total of 3643 samples were used in this study. The GSSL was split into two groups: a group of 700 "unknown" samples (Xu,Yu), and a group of 2943 samples which was used as reference set (Xr,Yr). Note that Xr and Yr represent the spectra and their corre-

spondent soil attribute (side information) in the reference set and Xu and Yu represent the spectra and their correspondent soil attribute (side information) in the "unknown" set. The soil attributes used in this study were clay content and pH. The distance algorithms were used to find in Xr the most spectrally similar samples of Xu (Figure 9). In order to evaluate the compositional similarity, the clay content and pH values of the Xu were compared to the clay content and pH values of the samples found in Xr by each algorithm.

It was found that information on the compositional similarity is useful for obtaining reliable distance measurements. The best distance metric approaches are those that better reflect the soil compositional similarity. In general, the results indicate that the distances computed in the spectral vis–NIR space have a lower performance in comparison to the ones computed in the low dimensional projected spaces.

The conventional methods (ED, MD, SAM and SID) commonly used in remote sensing did not present satisfactory performance when used in soil vis–NIR spectroscopy. One probable reason for that relies on the fact that in high dimensional spaces such the notion of similarity becomes less accurate (Abou–Moustafa and Ferrie, 2008).

The worst results were obtained by using the MD method. This is attributed to the fact that in this method the covariance matrix is computed in the vis–NIR spectral space which does not reflect well the relationships in the spectral variables. For this reason the classical estimates of the covariance matrix in the original vis–NIR space should be avoided.

In comparison to the standard method used in soil spectroscopy (i.e. the PC–M) for computing distances between vis–NIR spectra, its improved version (i.e. the oPC–M) proposed in this manuscript returned much better results. The oPC–M distance method only differs from the standard PC–M distance in the way in which the adequate number of PC variables to retain is calculated. The number of PCs retained in the PC–M method is based on the explained variance of the components while in the oPC–M the number of PCs is selected based on the maximization of the compositional similarity between most similar samples. The PC–

M tends to select a lower number of PC variables than the oPC–M. This indicates that the conventional selection of the number of PC variables may lead to a loss of information which is important for soil similarity analysis. However in the oPC–M method this important information is captured and used for computing more reliable distances. For example, when principal component analysis is used for regression, usually a larger number of PC variables are required in comparison to the ("adequate") number of PC variables that is indicated by the explained variance of the components. For regression of soil infrared data several works report that the number of components usually varies between 10 and 25 (e.g. Viscarra Rossel *et al.*, 2006; Viscarra Rossel *et al.*, 2008; Vasques *et al.*, 2009; Terhoeven-Urselmans *et al.*, 2010). However when principal component analysis is used only for projecting the soil infrared spectra into a lower dimensional space (i.e. compression) usually less than 6 PC variables are retained since they tend to explain a large part of the variance (e.g. Savvides *et al.*, 2010; Viscarra Rossel *et al.*, 2011). This also indicates that there is information about the soil composition in some of the PCs that are usually ignored when the conventional method of selection of PCs for dimensionality reduction purposes is employed.

Overall, it was found that the oPC–M, the LLE–M and the oLLE–M methods outperformed largely the current approaches used for soil vis–NIR distance measurements and they can be used for computing reliable vis–NIR similarity measurements.

These reliable methods would be very useful for integrating soil spectral libraries into proximal soil sensing for in the field soil predictions. For instance, given a soil spectral library (Xr) and a set of soil vis–NIR spectra measured in the field (Xu), it is possible to use a distance metric algorithm for searching the samples in Xr which are most similar to the Xu samples. Once the most similar samples have been found, specific soil models representing the field data can be calibrated. By using this procedure, redundant information as well as noisy or non–informative samples (regarding the field spectral variability) in the soil spectral library can be removed in order to infer the target soil attribute in the field.

6.2.2 Memory–based learning

(Manuscript 4, Geoderma, 2013, vol. 195–196, p. 268–279
doi: 10.1016/j.geoderma.2012.12.014)

In contrast to pure component systems, soil is a very complex mixture of mineral and organic constituents. Soil vis–NIR datasets are particularly complex and therefore the prediction performance of soil vis–NIR models calibrated from regional spectral libraries is usually low. In this respect, the main goal in this manuscript was to develop a suitable memory–based learning algorithm for calibrating soil vis–NIR models of soil attributes in large and heterogeneous datasets.

In this this manuscript, the spectrum–based learner (SBL) is introduced. It is a new algorithm which exploits both the vis–NIR features and the information of local distance matrices. As any other memory–based learning (MBL) method, the SBL does not yield a global function; instead it performs local interpolations which are based on a reference set or spectral library. The SBL is a three step approach which comprises: 1.Nearest neighbor search (recovering), 2. Training and testing, and 3. Fitting and predicting. These steps are described as follows:

1. *Nearest neighbor search (recovering)*: The main goal of this step is to discover which samples in a reference set "resemble" the samples to be predicted. Recovering similar samples from a set of samples stored in a "memory" (reference set) implies that similarity or dissimilarity measurements must be carried out. For these measurements a distance matrix can be used. In the SBL the nearest neighbor search process is carried out by using an optimized principal component distance method. The distances computed with this method are used to evaluate how similar or dissimilar the vis–NIR spectra are to the samples to be predicted.
2. *Training and testing*: Training and testing are carried out in the spectral space. For each sample to be predicted a model must be fitted by using its most similar samples i.e. its k-nearest neighbors. However prior to the fitting process, an adequate number of neighbors (k) (to be used in each calibration) must be identified. In this respect, k must be optimized since it can affect the fitting process. In this step k is optimized based on the minimization of the

prediction error of the set of closest samples (first nearest neighbors) to the prediction set which are found in the reference set.

This subset of closest samples can be viewed as the subset in the reference set that better reproduces the soil variability of the prediction samples therefore it can be exploited for optimizing k. Local predictions are carried out by using a linear Gaussian process (GP) regression algorithm which does not present internal parameters to be optimized. The SBL does not use (distance-based) weigthting functions for local regressions. The predictors at each local regression are a combination of local distance matrices and the vis–NIR features.

3. *Fitting and predicting*: Once the optimal k is found, a new local GP regression model is fitted for each sample in the prediction set with its corresponding k-nearest neighbors found in the reference set. As explained previously, the predictors in this step are also a combination of local distance matrices and vis–NIR features.

The above steps carried out by the SBL algorithm are summarized in Figure 10.

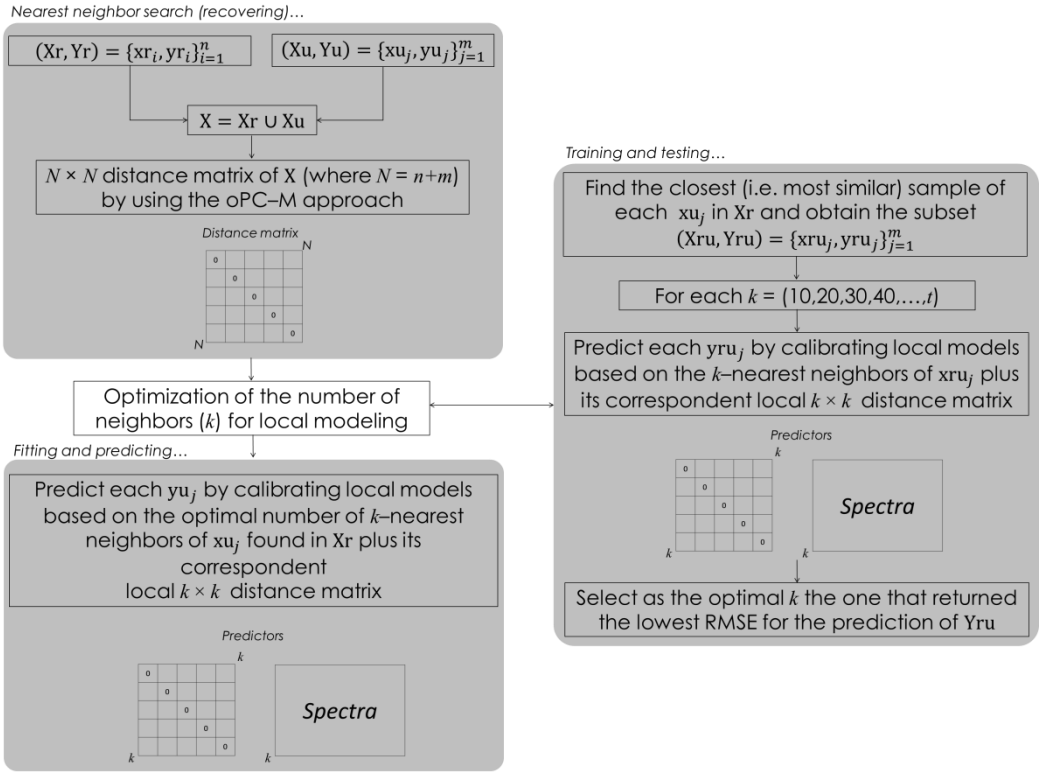

Figure 10. Description of the recovering, training and testing, and fitting and predicting steps in the SBL algorithm.

In order to test the SBL algorithm, two vis–NIR soil libraries were used: a regional soil spectral library (R–SSL, n = 4200) of the State of São Paulo (Brazil) and a soil spectral library of the world (G–SSL, n = 3643). As validation sets, 350 soil profiles (1050 samples) from the R–SSL and 125 soil profiles (900 samples) from the G–SSL were randomly sampled. The remaining samples were used as training sets in each spectral library.

The analyses were carried out separately for each soil spectral library. The SBL algorithm was used to calibrate vis–NIR local models and perform vis–NIR–based predictions of clay content (CC), soil organic carbon (OC) and exchangeable Ca^{++} in the validation samples. The SBL performance was evaluated on the basis of the accuracy of these predictions.

In addition, the following machine learning algorithms were used for predicting CC and OC: partial least squares regression (PLS), support vector regression machines (SVM), locally weighted PLS (LWR) and LOCAL. The PLS and SVM are global approaches (i.e produce a global function) while (like SBL) LWR and LOCAL are memory–based learners. The results obtained with these algorithms were compared with those obtained with the SBL algorithm.

Overall, SBL outperformed the global calibration models (PLS and SVM) and the other memory–based learning approaches (LWR and LOCAL) in both spectral libraries. In all cases, the SBL produced the lowest training and prediction errors as well as the highest prediction R^2. It was observed a trend in which the variability of the errors produced by the algorithms increases with the variability of the soil attribute. Comparing the prediction errors of CC in both libraries, it was found that the differences between algorithms are much higher in the R–SSL than in the G–SSL.

For the results obtained in the G–SLL, it was found that the SBL can produce competitive results in comparison with other approaches applied in global soil spectral libraries reported in the literature.

The good prediction performance of the SBL results from the combination of two important characteristics: *i.* an appropriate neighbor selection is carried out by using the distance matrix computed with the optimized principal components distance method, and *ii.* the inclusion in each local model of a $k \times k$ distance matrix as a source of additional predictor variables.

In contrast to other memory–based learning methods, in the SBL algorithm the distances between the target sample and its neighbor samples are not used for assigning weights to the neighbors. The distance information is used differently. The SBL uses the distance matrix as a source of additional predictor variables. At each (local) neighborhood, the local distance matrix between the all the neighbors (which is squared and symmetric) is used as source of additional predictors. For example, for the SBL predictions of CC in the R–SSL 360 neighbors at each local model were used. This means that each local model of CC was calibrated with 360 new predictor variables in addition to the spectral features.

The reason why the neighbor weighting approach is not used in the SBL algorithm is twofold. First, weighting implies the modification of all the spectral variables. Therefore if a distance score does not represent properly the similarity/dissimilarity between the samples, then it will affect the entire set of predictors of the sample which was weighted with the "noisy" distance score. Secondly, in that approach the information about the position of the samples within the neighborhood is missing since only the information about the distance to the target sample is employed.

It is assumed that the more similar two samples are in terms of their vis–NIR spectra, the more similar they are in terms of soil compositional characteristics. This means that in a given set of samples, the variability of a soil attribute could be explained in part by the variability of the spectral similarity/dissimilarity scores with respect to a reference point (spectrum).

Each sample in the neighborhood is used as a reference point within the same neighborhood. The similarity/dissimilarity between the reference point and all the samples is estimated. Each new similarity/dissimilarity variable (or column

of the distance matrix) represents new information about the position of the samples in the multivariate space. The exact position of the target sample within the neighborhood is known since the number of reference points is equal to the number of neighbor samples.

Figure 11, is used to illustrate the information (on the position of samples in a neighborhood) contained in distance matrices. It shows a one-dimensional example of distances between a target sample (red point) and two neighbors (triangles). In the example, two situations are presented: in the first one the target sample is located at one of the extremes and the distance between its neighbor samples is represented by the letters "a" and "b", while in the second situation the sample is located between its neighbors and the distance to them are also "a" and "b". The example 2 also presents two situations: in the first one, the target sample (located at the left extreme) is separated by two identical neighbors by a distance "b". In the second situation the neighbors are different; however target sample is located in the middle of them by a distance "b". When only the distance information between the target sample and its neighbor samples is used (as in the neighbor weighting approach) the information on the position of the samples is lost.

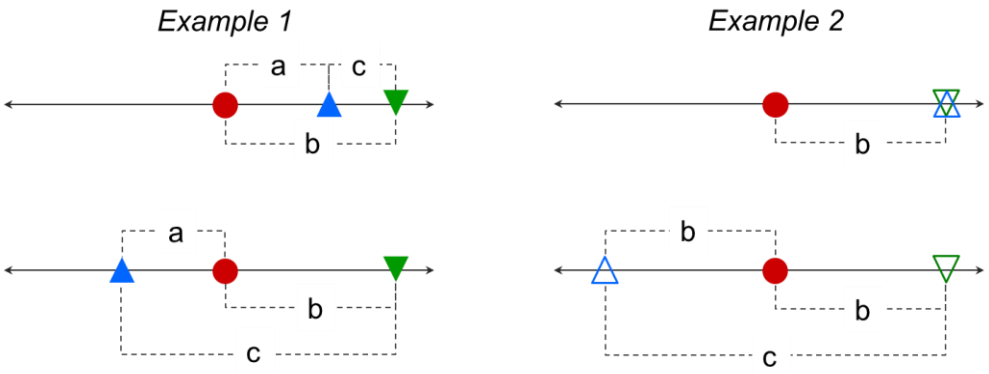

Figure 11. One-dimensional examples of sample position within neighborhoods.

Probably the information about the position of the target sample within the neighborhood could capture information about the variability of the samples which cannot be easily captured by the regression algorithm when it is applied

only to the vis–NIR variables. In spectroscopy, Zerzucha *et al*. (2012) showed that non-linear modeling problems can be resolved by simply applying partial least square regression on distance matrices.

In general, some of the concluding remarks drawn from this study are:

- The SBL is a new and reliable algorithm which produces more accurate predictions than global calibration models and the other memory–based learning approaches.
- The optimized principal component distances used in the SBL algorithm probably represent better the compositional similarity between samples than the conventional distance matrices used in the LWR and LOCAL algorithms.
- The use of local distance matrices as source of additional predictor do not degrade the prediction performance, instead it can result in an increment of it.
- The application potential of the SBL algorithm is not restricted to soil spectroscopy; its use could be extended to other research areas in which complex spectral libraries are being used.

Finally, this manuscript points out that soil spectroscopy research should be focused on bridging the gap between modeling algorithms and theories of the interactions between soil components and electromagnetic radiation. With the development of the SBL algorithm, this manuscript attempts to stimulate the use of memory–based learning which represents a straightforward strategy for integrating theory and algorithms.

7. References

Abou-Moustafa, K., Ferrie, F. 2008. Regularized minimum volume ellipsoid metric for query-based learning. Proceedings - 7th International Conference on Machine Learning and Applications, ICMLA 2008 , art. no. 4724974 , pp. 188-193.

Amezketa, E. 2007. Use of an electromagnetic technique to determine sodicity in saline-sodic soils. Soil Use and Management 23, 278-285.

Bellamy, P. H., Loveland, P. J., Bradley, R. I., Lark, R. M., Kirk, G. J. 2005. Carbon losses from all soils across England and Wales 1978–2003. Nature 437, 245-248.

Ben-Dor, E., Banin, A. 1990. Near-infrared reflectance analysis of carbonate concentration in soils. Applied Spectroscopy, 1064-1069.

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

Billings, S.A., 2008. Biogeochemistry: nitrous oxide in flux. Nature 456, 888–889.

Bowers S. A., Hanks, R.J. 1965. Reflection of radiant energy from soils. Soil Science 100, 130-138.

Bramley, R. G. V. E., Janik L.J. 2005. Precision agriculture demands a new approach to soil and plant sampling and Analysis – Examples from Australia. Communications in Soil Science and Plant Analysis 36, 9–22.

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

Cambule, A.H., Rossiter, D.G., Stoorvogel, J.J., Smaling, E.M.A. 2012. Building a near infrared spectral library for soil organic carbon estimation in the Limpopo National Park, Mozambique. Geoderma 183-184, 41-48.

Cameron, D.R., E. De Jong, D.W.L. Read, M. Oosterveld. 1981. Mapping salinity using resistivity and electromagnetic inductive techniques. Canadian Journal of Soil Science 61, 67–78.

Cañasveras, J.C., Barrón, V., del Campillo, M.C., Viscarra Rossel, R.A. 2012. Reflectance spectroscopy: a tool for predicting soil properties related to the incidence of Fe chlorosis. Spanish Journal of Agricultural Research 10, 1133-1142.

Clark, R.N., Roush, T.L. 1984. Reflectance Spectroscopy: Quantitative Analysis Techniques for Remote Sensing Applications. Journal of Geophysical Research 89, 6329–6340.

Coleman, T.L., Montgomery, L. 1987. Soil moisture, organic matter, and iron content effect on the spectral characteristics of selected vertisols and alfisols in Alabama. Photogrammetric Engineering and Remote Sensing 53, 1659-1663.

Condit, H.R. 1970. The spectral reflectance of American soils. Photogrammetric Engineering and Remote Sensing 36, 955-966.

de Groot, A.V., van der Graaf, E.R., de Meijer, R.J., Maučec, M. 2009. Sensitivity of in-situ γ-ray spectra to soil density and water content. Nuclear Instruments and Methods in Physics Research, Section A: Accelerators, Spectrometers, Detectors and Associated Equipment 600 (2), 519-523.

de Gruijter, J.J., McBratney, A. 2010. Sampling for High-Resolution Soil Mapping. In: Proximal Soil Sensing, Progress in Soil Science, edited by R. A. Viscarra Rossel, A. B. McBratney, B. Minasny, Springer Netherlands, Netherlands, p. 3–14.

Dickson, B. L., Scott, K. M. 1997. Interpretation of aerial gamma-ray surveys-adding the geochemical factors. AGSO Journal of Australian Geology and Geophysics, 17, 187-200.

Fritze, H., Jarvinen, P., Hiukka, R. 1994. Near-infrared characteristics of forest humus are correlated with soil respiration and microbial biomass in burnt soil. Biology and Fertility of Soils 18, 80-82.

Fu, X. , Ying, Y. , Yang, D. 2011. A comparative study of representative subset selection for NIR model updating. American Society of Agricultural and Biological Engineers Annual International Meeting 2011, 3411-3421

Ganjegunte, G.K., Braun, R.J. 2011. Delineating salinity and sodicity distribution in major soil map units of El Paso, Texas, using electromagnetic induction technique. Soil Science 176, 441-447.

Genot, V., Colinet, G., Bock, L., Vanvyve, D., Reusen, Y., Dardenne, P. 2011. Near infrared reflectance spectroscopy for estimating soil characteristics valuable in the diagnosis of soil fertility. Journal of Near Infrared Spectroscopy 19, 117–138.

Grinand, C., Barthes, B.G., Brunet, D., Kouakoua, E., Arrouays, D., Jolivet, C., Caria, G., Bernoux, M. 2012. Prediction of soil organic and inorganic carbon contents at a national scale (France) using mid-infrared reflectance spectroscopy (MIRS). European Journal of Soil Science 63, 141–151.

Guerrero, C., Zornoza, R., Gómez, I., Mataix–Beneyto, J. 2010. Spiking of NIR regional models using samples from target sites: Effect of model size on prediction accuracy. Geoderma 158, 66–77.

Hartemink, A. E., McBratney, A. 2008. A soil science renaissance. Geoderma, 148, 123-129.

Haygarth, P. M., Ritz, K. 2009. The future of soils and land use in the UK: Soil systems for the provision of land-based ecosystem services. Land use policy, 26, S187-S197.

Hedley, C.B., Yule, I.J., Eastwood, C.R., Shepherd, T.G., Arnold, G. 2004. Rapid identification of soil textural and management zones using electromagnetic induction sensing of soils. Australian Journal of Soil Research 42, 389-400.

Hunt, G.R., Salisbury, J.W., 1970. Visible and near infrared spectra of minerals and rocks. I. Silicate minerals. Modern Geology 1, 283-300.

Irons, J.R.; Weismiller, R.A.; Petersen, G.W. Soil reflectance. In: ASRAR, G. (Ed.). Theory and application of optical remote sensing. New York: Wiley, 1989. p. 429-473.

Janik, L.J., Merry, R.H., Skjemstad, J.O. 1998. Can mid infrared diffuse reflectance analysis replace soil extractions? Australian Journal of Experimental Agriculture 38, 681-696.

Janik, L.J., Skjemstad, J.O., Shepherd, K.D., Spouncer, L.R., 2007. The prediction of soil carbon fractions using mid-infrared-partial least-square analysis. Australian journal of soil research 45, 73–81.

Job, J.O., Gonzalez Barrios, J.L., Rivera Gonzalez, M. 1999. Effect of soil moisture on the determination of soil salinity using electromagnetic induction. European Journal of Environmental and Engineering Geophysics 3, 187-199.

Johnston, M.A., Savage, M.J., Moolman, J.H., Du Plessis, H.M. 1996. Calibration models for interpretation of soil salinity measurements using an electromagnetic induction technique. South African Journal of Plant and Soil 13, 110-114.

Kennard, R.W., Stone, L. 1969. Computer aided design of experiments. Technometrics 11, 137–148.

Kim, H. J., Sudduth, K. A., Hummel, J. W. 2009. Soil macronutrient sensing for precision agriculture. Jounal of environmental monitoring 11, 1810–1824.

King, T. V., Clark, R. N. 1989. Spectral characteristics of chlorites and Mg-serpentines using high-resolution reflectance spectroscopy. Journal of Geophysical Research 94(B10), 13997-14.

Kojima, M. 1958a. Relationship between size of soil particles and soil colors. Soil and Plant Food, 3, 204.

Kojima, M. 1958b . On the relation between soil color and its moisture content. Soil and Plant Food 3, 206.

Kuang B., Mouazen A. M. 2011. Calibration of visible and near infrared spectroscopy for soil analysis at the field scale on three European farms. European Journal of Soil Science 62, 629-636.

Kuang, B., Mouazen, A.M. 2012. Influence of the number of samples on prediction error of visible and near infrared spectroscopy of selected soil properties at the farm scale. European Journal of Soil Science 63, 421-429.

Leone, A.P., Viscarra-Rossel, R.A., Amenta, P., Buondonno, A. 2012. Prediction of soil properties with PLSR and vis–NIR spectroscopy: Application to mediterranean soils from southern Italy. Current Analytical Chemistry 8, 283-299.

Lesch, S.M., 2005. Sensor-directed response surface sampling designs for characterizing spatial variation in soil properties. Computers and Electronics in Agriculture 46, 153-180.

Lesch, S.M., Strauss, D.J., Rhoades, J.D., 1995. Spatial prediction of soil salinity using electromagnetic induction techniques: 2. An efficient spatial sampling algorithm suitable for multiple linear regression model identification and estimation. Water Resources Research 31, 387-398.

McBratney, A. B., Hart, G. A., McGarry, D. 1991. The use of region partitioning to improve the representation of geostatistically mapped soil attributes. Journal of Soil Science 42, 513–532.

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

McCarty, G.W., Reeves III, J.B., Reeves, V.B., Follett, R.F., Kimble, J.M. 2002. Mid-infrared and near-infrared diffuse reflectance spectroscopy for soil carbon measurement. Soil Science Society of America Journal 66, 640-646.

McKay, M.D., Conover W. J., Beckman, R. J. 1979. A comparison of three methods for selecting values of input variables in the analysis of output from a computer code. Technometrics, v. 21, p. 239-245.

McKenzie, N.J., Ryan, P.J., 1999. Spatial prediction of soil properties using environmental correlation. Geoderma 89, 67-94.

McKenzie, R.C., Chomistek, W., Clark. N.F. 1989. Conversion of electromagnetic inductance readings to saturated paste extract values in soils for different tem-

perature, texture and moisture conditions. Canadian Journal of Soil Science 69, 25–32.

McLeod, M.K., Slavich, P.G., Irhas, Y., Moore, N., Rachman, A., Ali, N., Iskandar, T., Hunt, C., Caniago, C. 2010. Soil salinity in Aceh after the December 2004 Indian Ocean tsunami. Agricultural Water Management 97, 605-613.

McNeill, J.D., 1992. Rapid, accurate mapping of soil salinity by electromagnetic ground conductivity meters. In: Advances in Measurement of Soil Physical Properties: Bringing Theory Into Practice. Spec. Publ. 30, SSSA, Madison, WI, pp. 209–229.

Minasny, B., Hartemink, A. E. 2011. Predicting soil properties in the tropics. Earth-Science Reviews, 106, 52-62.

Minasny,B., McBratney, A. 2010. Conditioned Latin hypercube sampling for calibrating soil sensor data to soil properties. In: Proximal Soil Sensing, Progress in Soil Science, edited by R. A. Viscarra Rossel, A. B. McBratney, B. Minasny, Springer Netherlands, Netherlands, p. 15–28.

Minasny. B., Mcbratney. A. B. 2006. A conditioned Latin hypercube method for sampling in the presence of ancillary information. Computers and Geosciences 32, 1378–1388.

Minty, B., 1997. Fundamentals of airborne gamma-ray spectrometry. AGSO Journal of Australian Geology and Geophysics 17, 39-50.

Mitchell, TM. Machine Learning. McGraw-Hill, New York, 1997.

Montgomery, O. L., Baumgardner, M. F. The effects of the physical and chemical properties of soils on the spectral reflectance of soils. 1974. LARS Technical Reports. Paper 134.

Naes, T., Isaksson, T., Kowalski, B. 1990. Locally weighted regression and scatter correction for nearinfrared reflectance data. Analytical Chemistry 62, 664–673

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

Northup, B.K., Daniel, J.A. 2012. Near infrared reflectance-based tools for predicting soil chemical properties of Oklahoma grazinglands. Agronomy Journal 104, 1122-1129.

Obukhov, E.D.A.I., Orlov, S. 1964. Spectral reflectivity of the major soil groups and possibility of using diffuse reflection in soil investigations. Soviet Soil Science 2, 174-184.

Odgers, N. P., Libohova, Z., Thompson, J. A. 2012. Equal-area spline functions applied to a legacy soil database to create weighted-means maps of soil organic carbon at a continental scale. Geoderma 189, 153-163.

Palacios-Orueta, A., Ustin, S.L. 1996. Multivariate statistical classification of soil spectra. Remote Sensing of Environment 57, 108-118.

Palmborg, C., Nordgren, A. 1993. Modelling microbial activity and biomass in forest soil with substrate quality measured using near infrared reflectance spectroscopy. Soil Biology and Biochemistry, 25, 1713-1718.

Pires, L.F., Rosa, J.A., Pereira, A.B., Arthur, R.C.J., Bacchi, O.O.S. 2009. Gamma-ray attenuation method as an efficient tool to investigate soil bulk density spatial variability. Annals of Nuclear Energy 36 (11-12), 1734-1739

Planet, W. G. 1970. Some comments on reflectance measurements of wet soils. Remote Sensing of Environment 1, 127-129.

Reedy, R.C., Scanlon, B.R. 2003. Soil water content monitoring using electromagnetic induction. Journal of Geotechnical and Geoenvironmental Engineering 129, 1028-1039.

Reeves III, J.B. 2009. Near- versus mid-infrared diffuse reflectance spectroscopy for soil analysis emphasizing carbon and laboratory versus on-site analysis: Where are we and what needs to be done? Geoderma 158, 3-14.

Rodionova, O.Y. , Pomerantsev, A.L. 2008. Subset selection strategy. Journal of Chemometrics 22, 674-685.

Rossiter, D. G. 2004. Digital soil resource inventories: status and prospects. Soil use and management 20, 296-301.

Sanchez, P. A., Ahamed, S., Carré, F., Hartemink, A. E., Hempel, J., Huising, J., ..., Zhang, G. L. 2009. Digital soil map of the world. Science 325, 680-681.

Sarkhot, D.V., Grunwald, S., Ge, Y., Morgan, C.L.S. 2011. Comparison and detection of total and available soil carbon fractions using visible/near infrared diffuse reflectance spectroscopy. Geoderma 164, 22-32.

Savvides,A., Corstanje, R., Baxter, S.J., Rawlins, B.G. Lark, R.M. 2010. The relationship between diffuse spectral reflectance of the soil and its cation exchange capacity is scale-dependent. Geoderma 154, 353–358.

Schmidt, K., Behrens, T., Friedrich, K., Scholten, T. 2010. A method to generate soilscapes from soil maps. Journal of Plant Nutrition and Soil Science 173, 163–172.

Sheets, K.R., Hendrickx, J.M.H. 1995. Noninvasive soil water content measurement using electromagnetic induction. Water Resources Research 31, 2401-2409.

Siano, G.G., Goicoechea, H.C. 2007. Representative subset selection and standardization techniques. A comparative study using NIR and a simulated fermentative process UV data. Chemometrics and Intelligent Laboratory Systems 88, 204–212.

Stenberg, B. 2010. Effects of soil sample pretreatments and standardised rewetting as interacted with sand classes on Vis–NIR predictions of clay and soil organic carbon. Geoderma 158 (1-2) , pp. 15-22 Geoderma 158, 15-22.

Stenberg, B., Viscarra Rossel, R.A., Mouazen, A.M., Wetterlind, J. 2010. Visible and near infrared spectroscopy in soil science. In: Donald L. Sparks, Ed. Advances in Agronomy, Vol. 107, Burlington: Academic Press, pp. 163–215.

Stevens, A., Udelhoven, T., Denis, A., Tychon, B., Lioy, R.,Hoffmann, L., van Wesemael, B. 2010. Measuring soil organic carbon in croplands at regional scale using airborne imaging spectroscopy. Geoderma 158, 32–45.

Stockmann, U., Minasny, B., McBratney, A. B., Hancock, G. R., Willgoose, G. R. 2012. Exploring short-term soil landscape formation in the Hunter Valley, NSW, using gamma ray spectrometry. In: Digital Soil Assessments and Beyond: Proceedings of the 5th Global Workshop on Digital Soil Mapping 2012, Sydney, Australia (p. 77). CRC Press.

Stoner, E.R., Baumgardner, F. 1981. Characteristic variations in reflectance of surface soils. Soil Science Society of America Journal, 45. 1161-1165.

Sudduth, K.A., Drummond, S.T., Kitchen, N.R. 2001. Accuracy issues in electromagnetic induction sensing of soil electrical conductivity for precision agriculture. Computers and Electronics in Agriculture 31, 239-264.

Sudduth, K.A., Hummel, J.W. 1991. Evaluation of reflectance methods for soil organic matter sensing. Transactions of the American Society of Agricultural Engineers 34, 1900-1909.

Sudduth, K.A., Kitchen, N.R., Sadler, E.J., Drummond, S.T., Myers, D.B. 2010. VNIR spectroscopy estimates of within-field variability in soil properties. In: Proximal Soil Sensing, Progress in Soil Science, edited by R. A. Viscarra Rossel, A. B. McBratney, B. Minasny, Springer Netherlands, Netherlands, p. 153–163.

Summers, D., Lewis, M., Ostendorf, B., Chittleborough, D. 2011. Visible near-infrared reflectance spectroscopy as a predictive indicator of soil properties. Ecological Indicators 11, 123-131.

Terhoeven–Urselmans, T., Vagen,T.G., Spaargaren,O. 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, 1792–1799.

Triantafilis, J., Gibbs, I., Earl, N. 2013. Digital soil pattern recognition in the lower Namoi valley using numerical clustering of gamma-ray spectrometry data. Geoderma 192, 407-421.

Triantafilis, J., Laslett, G.M., McBratney, A.B. 2000. Calibrating an electromagnetic induction instrument to measure salinity in soil under irrigated cotton. Soil Science Society of America Journal 64, 1009-1017.

Van Der Klooster, E., Van Egmond, F.M., Sonneveld, M.P.W. 2011. Mapping soil clay contents in Dutch marine districts using gamma-ray spectrometry. European Journal of Soil Science 62, 743-753.

Vasques, G.M., Grunwald, S., Sickman, J.O. 2009. Modeling of soil organic carbon fractions using visible-near-lnfrared spectroscopy. Soil Science Society of America Journal 73, 176-184.

Viscarra Rossel, R. 2009. The Soil Spectroscopy Group and the development of a global soil spectral library. NIR news, 20: 14–15

Viscarra Rossel, R., Behrens, T. 2010. Using data mining to model and interpret soil diffuse reflectance spectra. Geoderma 158, 46–54.

Viscarra Rossel, R.A., Chappell, A., De Caritat, P., Mckenzie, N.J. 2011. On the soil information content of visible-near infrared reflectance spectra. European Journal of Soil Science 62, 442-453.

Viscarra Rossel, R.A., Jeon, Y.S., Odeh, I.O.A., McBratney, A.B. 2008. Using a legacy soil sample to develop a mid-IR spectral library. Australian Journal of Soil Research 46, 1-16.

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

Wetterlind, J., Stenberg, B. 2010. Near–infrared spectroscopy for within–field soil characterization: Small local calibrations compared with national libraries spiked with local samples. European Journal of Soil Science 61, 823–843.

Wetterlind, J., Stenberg, B. Soderstrom, M. 2010. Increased sample point density in farm soil mapping by local calibration of visible and near infrared prediction models. Geoderma 156, 152–160.

Wilford, J. 2012. A weathering intensity index for the Australian continent using airborne gamma-ray spectrometry and digital terrain analysis. Geoderma 183-184, 124-142.

World Agroforestry Centre (ICRAF) and ISRIC – World Soil Information. 2010. ICRAF–ISRIC Soil vis–NIR spectral Library. Nairobi, Kenya: World Agroforestry Centre (ICRAF).

Zerzucha, P., Daszykowski, M., Walczak, B. 2012. Dissimilarity partial least squares applied to non-linear modeling problems, Chemometrics and Intelligent Laboratory Systems 110, 56–162.

8. Ergänzungsblatt zur Eigenleistung

Erklärung nach § 5 Abs. 2 Nr. 7 der Promotionsordnung der Math.-Nat. Fakultät
-Anteil an gemeinschaftlichen Veröffentlichungen-

Declaration according to § 5 Abs. 2 No. 7 of the PromO of the Faculty of Science
-Share in publications done in team work-

Name: Leonardo Ramirez Lopez

List of publications

1. Ramirez-Lopez, L, Schmidt, K., Behrens, T., Demattê, J.A.M., van Wesemael, B., Scholten, T. Calibration sampling and calibration set size for soil vis–NIR modeling. Geoderma, submitted on July 2012.

2. Schmidt, K., Behrens, T., Daumann, J., Ramirez-Lopez, L., Werban, U., Dietrich, P., Scholten, T. A comparison of calibration sampling schemes at the field scale. Geoderma, submitted on June 2012.

3. Ramirez-Lopez, L., Behrens, T., Schmidt, K., Viscarra Rossel, R., Demattê, J.A.M., Scholten, T. Distance and similarity-search metrics for use with soil vis–NIR spectra. Geoderma, doi: 10.1016/j.geoderma.2012.08.035. Accepted.

4. Ramirez-Lopez, L., Behrens, T., Schmidt, K., Stevens, A., Demattê, J.A.M., Scholten. The spectrum-based learner: a new local approach for modeling soil vis–NIR spectra. Geoderma, doi: 10.1016/j.geoderma.2012.12.014. Accepted.

Nr.	Accepted for publication /Submitted	Number of all authors	Position of the candidate in list of authors	Scientific ideas of the candidate (%)	Data generation by the candidate (%)	Analysis and interpretation by the candidate (%)	Paper writing by the candidate (%)
1	Submitted	6	1st	80	60	80	90
2	Submitted	7	4th	20	30	20	20
3	Accepted	6	1st	80	10	80	90
4	Accepted	6	1st	90	80	90	90

I certify that the above statement is correct.

_____

Date, Signature of the candidate

I/We certify that the above statement is correct.

_____

Date, Signature of the doctoral committee or at least of one of the supervisors

Manuscipt 1: Calibration sampling and calibration set size for soil vis–NIR modeling

Geoderma, submitted on July 2012

Leonardo Ramirez-Lopez[a,b], Karsten Schmidt[a], Thorsten Behrens[a], Bas van Wesemael[b], Jose A. M. Demattê[c], Thomas Scholten[a]

[a]Institute of Geography, Physical Geography and Soil Science, University of Tübingen,
Rümelinstraße 19–23, 72070, Tübingen, Germany.

[b]Georges Lemaître Centre for Earth and Climate Research, Earth and Life Institute,
Université Catholique de Louvain,
3 Place Louis Pasteur – 1348, Louvain la Neuve, Belgium.

[c]Soil Science Department, Escola Superior de Agricultura "Luiz de Queiroz"
University of São Paulo.
Av.: Pádua Dias, 11 CP 9. Piracicaba – SP 13418–900. Brazil.

Abstract

By using a small and representative number of soil samples from a given area it is possible to calibrate vis–NIR models of soil attributes. Those models can be used to predict soil attributes over a large number of soil samples of the area by only using the spectral information. Despite the well-known potential of vis–NIR spectroscopy for obtaining high resolution soil information, research on the adequate size of the calibration set has not received enough attention. In this respect, we investigated the effect of both the calibration set size and the calibration sampling strategy on the predictive performance of vis–NIR models of clay content and exchangeable Ca (Ca^{++}). We evaluated the following calibration sampling algorithms: Kenard–Stone (KSS), conditioned Latin hypercube (cLHS) and fuzzy c-means (FCMS). These algorithms were tested separately in a field–scale dataset and in a regional scale dataset. For each dataset we randomly selected a validation subset and the remaining samples were used as candidate samples for

calibration. The accuracy of vis–NIR models of clay content and Ca^{++} were compared on the basis of the sampling algorithms used for selecting the calibration samples. We also tested different calibration set sizes varying from 10 to 380 samples in steps of 10 samples. The models were calibrated by using the support vector regression machines (SVM) algorithm. The training root mean square error (RMSE), the normalized RMSE and the prediction RMSE were used to evaluate the sensibility of the models to both the sampling algorithm and the calibration set size. In addition, we investigate on the sample representativeness of each algorithm. The results show that the sampling strategy is critical at low calibration set sizes. Furthermore, based on the sample representativeness analysis we suggest a method to identify an adequate calibration set size.

Keywords: soil spectroscopy; sampling strategy; calibration set size; Kennard–Stone sampling; Latin hypercube sampling; fuzzy c–means.

1. Introduction

During the last two decades a growing interest on the quantification of soil attributes by means of soil sensing techniques has emerged. Most of these techniques such as soil infrared soil spectroscopy have great potential for high resolution soil sampling and mapping because they are faster and cost-effective compared to conventional methods (Bramley and Janik 2005; Kim *et al.*, 2009). For example, soil visible and near infrared spectroscopy (vis–NIR, 400-2500 nm) can be used as a tool for increasing the number of analyses (increasing the sampling density) and consequently the accuracy of soil maps without considerable increase in costs (Wetterlind *et al.*, 2010). In this respect, for one given area in which vis–NIR data is available at high spatial resolution, it is possible to calibrate vis–NIR models of soil attributes by using a small but representative number of soil spectral samples. Those models can be used to predict soil attributes efficiently over a large number of soil samples belonging to that area by only using the soil vis–NIR spectra. In this context, the strategy for selecting an adequate calibration set in terms of representativeness and size (number of samples) is of fundamental importance to ensure accurate prediction performances.

Despite the well-known potential of vis–NIR spectroscopy for obtaining high spatial resolution soil information, research on both the sampling strategy and the adequate calibration set size have not received enough attention (Grinand *et al*., 2012; Kuang and Mouazen, 2012). In principle, the optimal calibration set size could vary depending on the soil variability of the area under study (Kuang and Mouazen, 2012). Due to this, strategies for identifying the optimal calibration set size without an explicit prior knowledge of the soil attributes to be predicted are of great importance for the practical application of soil spectroscopy at the field scale. On the other hand, Brown *et al*, (2005) indicate that the calibration sampling strategy is crucial when only few samples can be included in the calibration set. Furthermore, Minasny and McBratney (2010) stress the importance to investigate the relation between the calibration sampling strategy and the prediction accuracy of soil models.

Concerning calibration sampling algorithms, the most common ones applied in pedometrics are the fuzzy c–means based sampling (de Gruijter *et al*., 2010) and the Latin hypercube sampling (McKay *et al*., 1979; Minasny and McBratney, 2006). Another calibration sampling algorithm widely employed in chemometrics (Daszykowski *et al*., 2002) and often used in soil spectroscopy is the Kennard-Stone sampling (Kennard and Stone, 1969). All these algorithms attempt to cover adequately the multivariate space of a set of predictors. Despite this, several works have shown that the strategies employed for covering the multivariate space can lead to different levels of prediction accuracies (eg. Siano and Goicoechea, 2007; Rodionova and Pomerantsev 2008; Fu *et al*., 2011).

In this context the main objectives of this paper were: *i*. Investigate on the effect of both the calibration set size and the sampling algorithm on the predictive performance of soil vis–NIR models for predicting clay content and exchangeable calcium (Ca^{++}) and *ii*. Analyze the sample representativeness on the basis of three different calibration sampling algorithms.

2. Theory

2.1 Kennard–Stone sampling (KSS)

The KSS (Kennard and Stone, 1969) has been widely used in quantitative spectroscopy showing good performance in terms of calibration sampling (eg. Wu *et al.*, 1996; Daszykowski *et al.*, 2002; Zhu *et al.*, 2009; Gogé *et al.*, 2012). The KSS is a deterministic sequential approach which was initially called uniform mapping algorithm since it attempts to select samples uniformly distributed in the predictor space. In KSS, the procedure to select a training or calibration subset of n samples ($X_{tr} = \{x_{trj}\}_{j=1}^{n}$) from a given set of N samples ($X = \{x_i\}_{i=1}^{N}$, note that $n<N$) consists in:

1. Find in X the sample x_{tr1} which is closest to the mean (μ), allocate it in X_{tr} and remove it from X.
2. Find in X the sample x_{tr2} which is the most dissimilar to x_{tr1}, and allocate x_{tr2} in X_{tr} and remove it from X.
3. Find in X the sample x_{tr3} which is the most dissimilar to the ones already allocated in X_{tr}. Allocate x_{tr3} in X_{tr} and then remove it from X. Note that the dissimilarity between X_{tr} and each x_i is given by the minimum distance of any sample allocated in X_{tr} to each x_i.
4. Repeat the step 3, n–4 times in order to select the remaining samples ($x_{tr4}, \ldots, x_{tr\,n}$).

For distance computations in the KSS algorithm, the Euclidean distance is commonly used, however the Mahalanobis distance (MD) is also an adequate alternative.

2. 2 Conditioned Latin hypercube sampling (cLHS)

In soil spectroscopy, the cLHS (Minasny and McBratney 2006) has been used for calibration sampling and uncertainty analyisis (e.g. Viscarra Rossel *et al.*, 2008; McBratney *et al.*, 2006). Basically, the cLHS attempts to cover the multivariate space of the predictor variables by using a stratified random sampling scheme

based on the cumulative distributions of those variables. In a one–dimensional space, the cumulative distribution of the variables of X is divided into n (number of sampling points) strata and the idea is to select one sample per stratum. However in a multivariate space this task becomes more complex. In cLHS the training subset X_{tr} with n samples taken from a set X with N samples (where $n<N$) must form a Latin hypercube. In the case of continuous variables, the objective function (O) of the cLHS integrates two objective functions O_1 and O_2, so that (eq. 1):

$$O = O_1 + O_2 \tag{1}$$

In this respect O_1 is given by eq. 2:

$$O_1 = \sum_{i}^{n}\sum_{j=1}^{m}\left|\eta(q_j^i \leq \mathrm{X}_{\mathrm{tr}\,j} < q_j^{i+1}) - 1\right| \tag{2}$$

where m is the number of variables, $\eta(q_j^i \leq x_{\mathrm{tr}\,j} < q_j^{i+1})$ is the number of samples in X_{tr} whose cumulative distribution values at the jth variable fall in the stratum that comprises q_j^i and q_j^{i+1}. On the other hand, O_2 is based on the differences between C and A which are the correlation matrix for X and the correlation matrix for X_{tr} respectively. The O_2 is calculates as follows (eq. 3):

$$O_2 = \sum_{i=1}^{m}\sum_{j=1}^{m}\left|\mathrm{C}_{ij} - \mathrm{A}_{ij}\right| \tag{3}$$

A simulated annealing scheme is carried out in order to find a subset Xtr that returns an O as close as possible to zero where the cumulative distribution of Xtr is representative for the original cumulative distribution of X. The reader is referred to Minasny and McBratney (2006) for additional details on the cLHS algorithm.

2.3 Fuzzy c–means based sampling

In soil science, the fuzzy c–means based sampling (FCMS) has been proposed as a calibration sampling method (de Gruijter *et al.*, 2010). The FCMS works on the basis of the fuzzy c–means clustering algorithm (Dunn, 1973; Bezdek, 1981) and

a nearest neighbor search. The cluster algorithm basically creates sample partitions of a given dataset. Each sample in the dataset is assigned to one cluster; however each sample also has a degree of membership to each cluster. It is expected that samples belonging to each cluster will share similar characteristics while the dissimilarity between samples in different clusters will be maximized. For each cluster construction a centroid is calculated. The optimal fuzzy c-partitions and centroids are found by using the following objective function (eq. 4):

$$J(U,V;C) = \sum_{i=1}^{N}\sum_{j=1}^{c} u_{ij}^{m}\ d(x_i, \mathrm{v}_j)^2 \qquad (4)$$

where V= $\{\mathrm{v}_j\}_{i=1}^{c}$, is a matrix of prototypes of cluster centroids, $d(x_i, \mathrm{v}_j)^2$ is the squared distance between each sample and each prototype, m is a fuzzy exponent, c is the number of clusters and u_{ij} is given by eq. 5:

$$u_{ij} = \frac{d\left(x_i, \mathrm{v}_j\right)^{-2/(m-1)}}{\sum_{q=1}^{c} d(x_i, \mathrm{v}_q)^{-2/(m-1)}} \qquad (5)$$

Although different metrics for similarity measurements can be used the most common ones are the Euclidean distance and the Mahalanobis distance. In fuzzy c–means clustering the only two parameters that need to be set are m (which controls the fuzziness of the cluster model) and c. Values of m may vary between 1 (which corresponds to a hard clustering) and infinity (soft clustering) and a typical choice is m=2 (Odeh *et al.*, 1999).

In fuzzy c–means clustering for calibration sampling purposes a nearest neighbor search is applied to select the nearest sample to each cluster centroid. The set of nearest samples is then the final calibration set. In this sense the number of clusters determines the number of calibration samples to be selected.

3. Material and Methods

3.1 Field scale dataset

3.1.1 Field sampling

The field scale dataset covers an area of 5 km² and it is located in the State of São Paulo (Brazil, 22°24'30"S and 48°29'58"W) at altitudes ranging from 500 to 710 m. This area has been historically cultivated with sugarcane. A set of basaltic flows alternated with sandstones of the Serra Geral Formation underlies the area. The predominant soils are: Arenosols, Ferralsols, Acrisols, Cambisols and Nitisols (IUSS Working Group WRB, 2006).

The field sampling scheme was based on a dense regular grid of 100 × 100 m where soil samples were collected at depth intervals of 0–0.20 m (459 samples) and 0.80–1.00 m (452 samples).

3.1.2 Soil vis–NIR scanning

Soil samples were oven–dried for 24 hours at 45°C, and sieved (2 mm mesh) prior to the spectral scanning. In order to obtain their bidirectional reflectance vis–NIR spectra, an Infrared Intelligent Spectroradiometer (Geophysical and Environmental Research Corporation, Buffalo, Ney York) was used. Each spectrum resulted from an average of 100 scans of the same sample. The spectra were obtained in the form of absorbance (log 1/Reflectance) with spectral resolution of 2 nm in the range from 400 to 1000 nm and 4 nm for the range from 1004 to 2500 nm. The final spectra comprised 830 spectral bands.

3.2 Regional scale dataset

3.2.1 Area and samples

For this dataset the study are covers an area of approximately 464 km² and it is located in the central–eastern portion of the state of São Paulo (Brazil, 22°51'51"S and 47°36'08"W). This area has been used for sugarcane production.

Sandstone, siltstone, and shale dominate with inclusions of limestone, basalt, and colluvial deposits. Elevations range from 489 to 709 m. The soils are classified as Arenosols, Ferralsols, Acrisols, Alisols, Nitisols, Cambisols and Lixisols (IUSS Working Group WRB, 2006).

In this dataset, soil samples correspond to 318 soil profiles collected over the past 10 years in different soil surveys. These profiles were sampled at three depth intervals: 0-0.2 m (318 samples), 0.4-0.6 m (317 samples), and 0.8-1.0 m (291 samples).

3.2.2 Soil vis–NIR scanning

Samples were air-dried and sieved (2 mm). Their soil vis–NIR (400-2500 nm) reflectance spectra were scanned using a FieldSpec Pro sensor (Analytical Spectral Devices Inc., Boulder, CO) which is characterized by a full width half maximum of 3 nm for the 350-1000 nm region and 10 nm for the 1000-2500 nm region. The final spectrum of each sample was an average of 100 scans. The reflectance spectra were resampled to a spectral resolution of 4 nm obtaining a total of 526 spectral features.

3.3 Soil analyses

The soil attributes evaluated in this study were clay content and exchangeable calcium (Ca^{++}). For samples of both the field and the regional datasets the ion exchange resin method (Raij *et al.*, 1987) was used for Ca^{++} analysis. The densimeter method (Camargo *et al.*, 1986) was used to measure the clay content.

3.4 Calibration: sampling, set size and SVM modeling

Here we describe the general framework followed for analyzing both datasets separately. All the algorithms were implemented in R 2.14.1 (R Development Core Team, 2011).

In order to avoid multi-colinearity and high dimensionality problems inherent to the vis–NIR spectra, all the sampling procedures were carried out on the princi-

pal component (PC) space of the vis–NIR spectra. The number of PCs retained in the analysis was based on the cumulative amount of spectral variance explained. The PCs that accounted for less than 0.1% of the total spectral variance were ignored. Each retained PC variable was standardized dividing it by its standard deviation.

We used KSS, FCMS and cLHS for selecting the calibration samples. In the case of FCMS we used a fuzzy exponent of 2. For KSS and FCMS the Euclidean distance was used. As we standardized the PCs, in this case the Euclidean distance is equivalent to the Mahalanobis distance (De Maesschalck *et al.*, 2000).

As validation sets, we randomly sampled 138 profiles (275 samples) from the field dataset and 83 profiles (249 samples) from the regional dataset. The remaining samples were used as candidate samples for the calibration sampling algorithms. We sampled entire profiles as a validation sets instead individual samples in order to avoid pseudo-replication of samples (Terhoeven-Urselmans *et al.*, 2010). Figure 1 shows the spatial distribution of the candidate profiles for calibration sampling and also the validation profiles for both datasets. For each calibration sampling approach (KSS, cLHS and FCMS) we selected different calibration sets with sizes varying from 10 to 380 samples in steps of 10 samples.

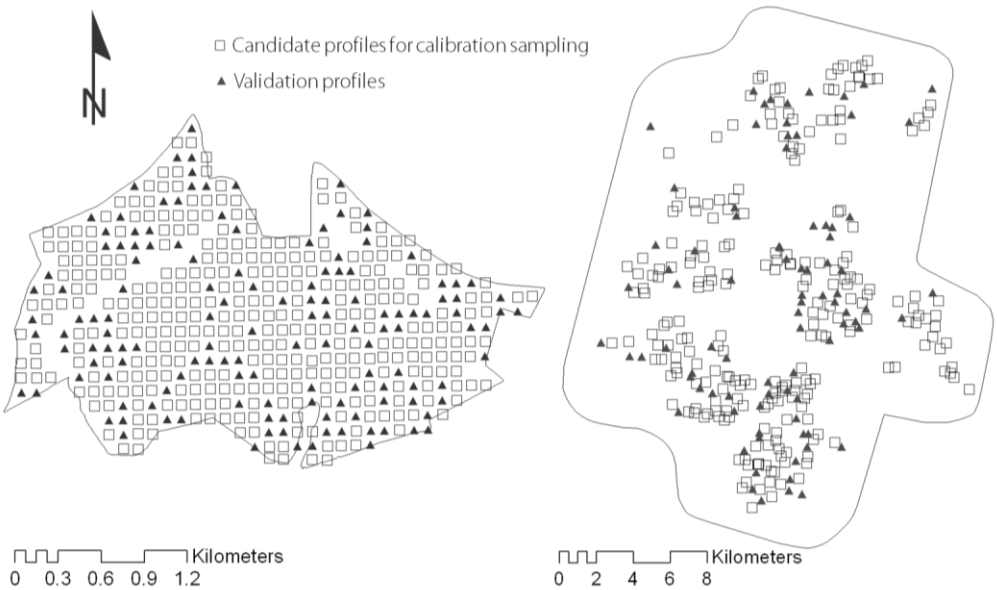

Figure 1. Spatial distribution of both the candidate profiles for calibration sampling and the validation profiles in the field scale dataset (left) and regional scale dataset (right).

In each calibration set selected with each sampling approach, the support vector regression machines (SVM) algorithm (Drucker, 1996) was used for calibrating models of clay content and Ca^{++}. The reason for using SVM instead the classical partial least squares (PLS) regression was to ensure good prediction performance. Viscarra Rossel and Behrens (2010) showed that SVM outperforms several machine learning algorithms including PLS. Briefly, the SVM uses the kernel trick (Aizerman *et al.*, 1964) to perform a non–linear transformation of the original predictor space into a high dimensional space (without computing it explicitly) with a linear or nearly linear structure. In this work, we used the linear basis function (LBF) kernel in order to keep the models as simple as possible. The LBF does not require any parameter (or hyper–parameter for the SVM models) to be optimized. Therefore, in this case the only parameter to be optimized in the SVM algorithm was the penalty factor (C).

Training the SVM models consisted in tuning the C parameter. In this respect, we tested six possible values (0.1, 0.25, 0.5, 1, 2 and 4) of C. A total of 50 bootstrap resampling iterations were used for both tuning C and assessing the predictive accuracy of the SVM models. The optimal C parameter was chosen as the one that minimized the training root mean square error (RMSE) which was calculated as follows (eq. 6):

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \quad (6)$$

where y_i is the ith observed value, $\hat{y}_i$ is the ith predicted value, and n is the number of samples. The normalized RMSE (nRMSE) was also calculated for the calibrations. The nRMSE is computed as follows (eq. 7):

$$\text{nRMSE} = \frac{\text{RMSE}}{y_{max.} - y_{min.}} \quad (7)$$

where $y_{max.}$ and $y_{min.}$ are the maximum and the minimum values of the observed soil attribute in the calibration samples selected by each sampling algorithm.

The models obtained were applied on the validation sets and the RMSEs of these predictions were also computed.

As both FCMS and cLHS are stochastic methods, to reduce noisy effects in the results, we repeated the calibration sampling procedure 10 times for each calibration set size with its correspondent SVM. The averages of the RMSEs as well as the averages of the nRMSEs of the 10 repetitions are the final RMSEs and nRMSEs reported here.

3.5 Assessment of the sampling representativeness in the PC space

In order to further evaluate the performance of the sampling algorithms we performed an analysis of the representativeness of the selected samples in the PC space. As explained in the section 2.2 the cLHS algorithm ensures a good representation of the original variability. In this respect, we also wanted to investigate whether the KSS and FCMS (in addition to covering the predictor space) can also guarantee a good representation of the original statistical distribution of the vis–NIR data in the PC space.

For each sampling algorithm and for each calibration set size, the sample mean ($\overline{x}$) and the sample variance (s^2) of the PC variables were compared to the original mean (μ) and the original variance (σ^2) of the PCs. Note that σ^2 and μ, are equivalent to 1 and 0 respectively since the PC variables are standardized to zero mean and unit variance. Both the absolute difference between variances ($|s^2 - \sigma^2|$) and the absolute difference between means ($|\overline{x} - \mu|$) were computed as follows (eqs. 8 and 9):

$$|s^2 - \sigma^2| = \quad |s^2 - 1| = \left|\frac{1}{k}\left(\sum_{j=1}^{k} s^2_{\mathrm{pc}\,j}\right) - 1\right|; \tag{8}$$

$$|\overline{x} - \mu| = \quad \left|\frac{1}{k}\left(\sum_{j=1}^{k} \overline{x}_{\mathrm{pc}\,j}\right) - 0\right|, \tag{9}$$

where $s^2_{\mathrm{pc}\,j}$ and $\overline{x}_{\mathrm{pc}\,j}$ are the sample variance and the sample mean of the jth PC, and k is the total number of PCs retained.

4. Results and discussion

4.1 Soil attributes and vis–NIR characteristics

Soil attributes varied widely in both datasets (Table 1). Both, clay content and Ca^{++} showed a positive skewness, which indicates that the mass of the distribution is concentrated below the median value.

Table 1. Descriptive statistics of the soil attributes of samples in both datasets.

Soil attribute	Units	S.d.[a]	Mean	Min.[a]	1st Qu.[b]	Median	3rd Qu.[b]	Max.[a]	Skewness
					Field scale dataset				
Clay	%	16.6	24.8	2.0	14.0	18.0	27.0	81.0	1.56
Ca^{++}	$cmol_c\ kg^{-1}$	14.0	14.4	0.0	5.4	10.1	18.7	99.7	2.08
					Regional scale dataset				
Clay	%	16.3	35.7	6.0	21.5	35.0	48.6	81.1	0.18
Ca^{++}	$cmol_c\ kg^{-1}$	16.2	25.9	2.0	14.0	22.0	32.0	98.1	1.50

[a] Min., Max. and S.d. correspond to the minimum, maximum and standard deviation respectively;
[b] 1st Qu, 3rd Qu. Correspond to the first and third quartiles respectively.

The vis–NIR reflectance spectra of the field dataset (Figure 2a) showed well defined absorption features near to 1455 and 1915 nm bands. These are assigned to hygroscopic water in clay minerals (Ben-Dor *et al.*, 2008). The spectra of all samples showed the influence of iron oxides with central absorption bands at 435, 550 and 850 nm, which is characteristic of the presence of goethite and hematite (Demattê and Garcia, 1999; Fernandes *et al.*, 2004). In most of the samples we observed absorption features in the 2207 nm and 2160 nm which are related to the kaolinite content (Demattê *et al.*, 2004; Viscarra Rossel and Behrens, 2010). Mean values of soil vis–NIR reflectance showed a significant inverse correlation with clay content (r = -0 .72, p< 0.05) and with Ca^{++}(r = -0 .61, p< 0.05). This is probably related to the fact that soils with high clay content present high energy absorption, while soils with high sand content present higher albedo due to higher amounts of quartz (White *et al.*, 1997). Concerning the inverse correlation observed between Ca^{++} and mean reflectance, it is possible that this is a consequence of a secondary correlation between Ca^{++} and clay content (r = 0 .62, p< 0.05) and not to a direct influence of the Ca^{++} on the albedo. In general, the vis–NIR reflectance spectra of soil samples showed similar characteristics in the

shapes of the absorption features. This means that mineralogical differences among samples are relatively small taking into account that soil mineralogy has strong influence on the shape and position of absorption features (Ben-Dor and Banin, 1995). In this sense, the main differences between samples are associated to textural variations.

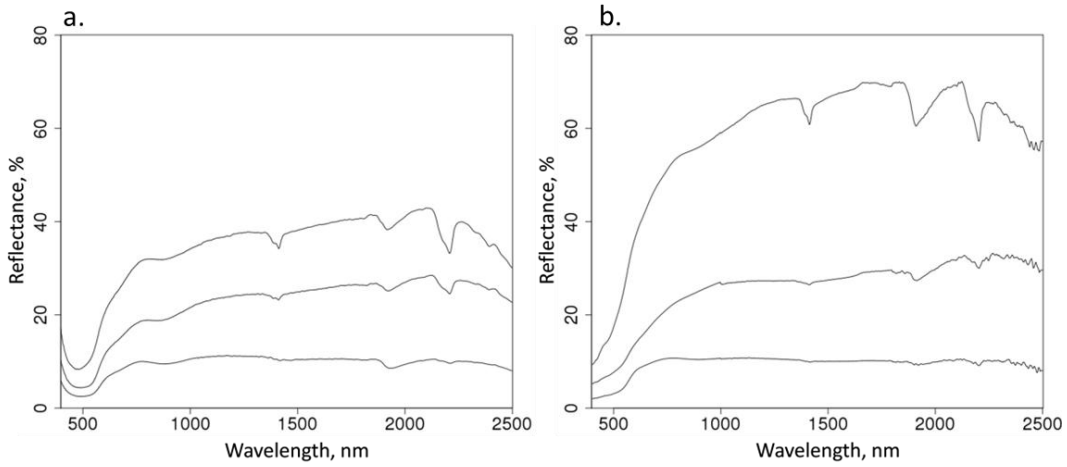

Figure 2. Reflectance spectra corresponding to the samples with: the lowest mean reflectance, the highest mean reflectance and the closest sample to the median of the mean reflectance values. a. Filed scale dataset; b. Regional scale dataset.

The vis–NIR spectra of the regional dataset (Figure 2b) presented larger variation in comparison to the field dataset (Figure 2a). In the regional dataset we also observed that most of the samples showed an absorption feature in the 2207 nm related to the presence of kaolinite. Furthermore most of the samples presented typical characteristics of energy absorption at 1455 nm and 1915 nm assigned to soil hygroscopic water. Other features attributed to soil attributes such as pedogenic oxides showed contrasting influence on the soil spectra. A significant inverse correlation (r = -0 .64, p< 0.05) between mean reflectance and clay content was observed, however in this case the correlations between Ca^{++} and mean reflectance (r = -0 .05, p> 0.05) and Ca^{++} and clay content (r = 0 .09, p> 0.05) were not significant.

In the PC analysis, 7 and 8 PCs were retained for the field dataset and for the regional dataset respectively. In both cases, these PCs accounted for 99.9% of the total vis–NIR variation. The descriptive statistics corresponding to the retained PCs are presented in the Table 2. As the PCs were standardized the standard

deviation of all them was 1. The mean of all PCs were 0 and the skeweness indicated that the mass of the distributions is concentrated around the mean.

Table 2. Descriptive statistics of the retained PCs in both datasets.

Variable	S.d.[a]	Mean	Min.[a]	1st Qu.[b]	Median	3rd Qu.[b]	Max.[a]	Skewness
Field scale dataset								
PC 1	1.00	0.00	-3.18	-0.49	0.25	0.69	2.80	-0.76
PC 2	1.00	0.00	-5.95	-0.65	0.10	0.65	3.08	-0.61
PC 3	1.00	0.00	-2.11	-0.81	-0.01	0.71	4.23	0.46
PC 4	1.00	0.00	-8.85	-0.60	0.01	0.61	3.03	-1.46
PC 5	1.00	0.00	-5.75	-0.61	0.00	0.63	4.35	-0.09
PC 6	1.00	0.00	-4.55	-0.57	0.09	0.65	6.59	-0.06
PC 7	1.00	0.00	-4.06	-0.64	0.00	0.56	6.55	0.44
Regional scale dataset								
PC 1	1.00	0.00	-2.02	-0.83	-0.08	0.76	2.28	0.26
PC 2	1.00	0.00	-3.37	-0.68	0.08	0.69	3.73	-0.16
PC 3	1.00	0.00	-2.93	-0.72	0.07	0.73	2.51	-0.21
PC 4	1.00	0.00	-2.87	-0.67	-0.09	0.53	3.35	0.59
PC 5	1.00	0.00	-4.85	-0.55	0.10	0.63	4.79	-0.49
PC 6	1.00	0.00	-4.35	-0.56	-0.03	0.63	3.79	-0.34
PC 7	1.00	0.00	-4.74	-0.52	0.06	0.59	3.92	-0.41
PC 8	1.00	0.00	-3.04	-0.70	-0.09	0.66	4.38	0.41

[a] Min., Max. and S.d. correspond to the minimum, maximum and standard deviation respectively;
[b] 1st Qu, 3rd Qu. Correspond to the first and third quartiles respectively.

4.2 Sampling algorithms and the effect of the calibration set size

The results of the effects of the calibration set size as well as the sampling algorithm on the accuracy of the models are presented in the Figure 3. For all the sampling algorithms in both datasets and for calibration set sizes < 200 samples, we observed a general trend in which the training RMSE, the nRMSE and the prediction RMSE decreased considerably as the calibration set size increased. In most of the cases at calibration set sizes ≥ 200 samples, the errors remained relatively stable, however in the case of the KSS in the field dataset, the training RMSEs of clay content and Ca^{++} showed a slightly decreasing tendency. Similar trends have been reported in other works in which the effect of the calibration set size is critical when small calibration set sizes are used (eg. Shepherd and Walsh, 2002; Brown *et al.*, 2005; Grinand *et al.*, 2012; Kuang and Mouazen, 2012).

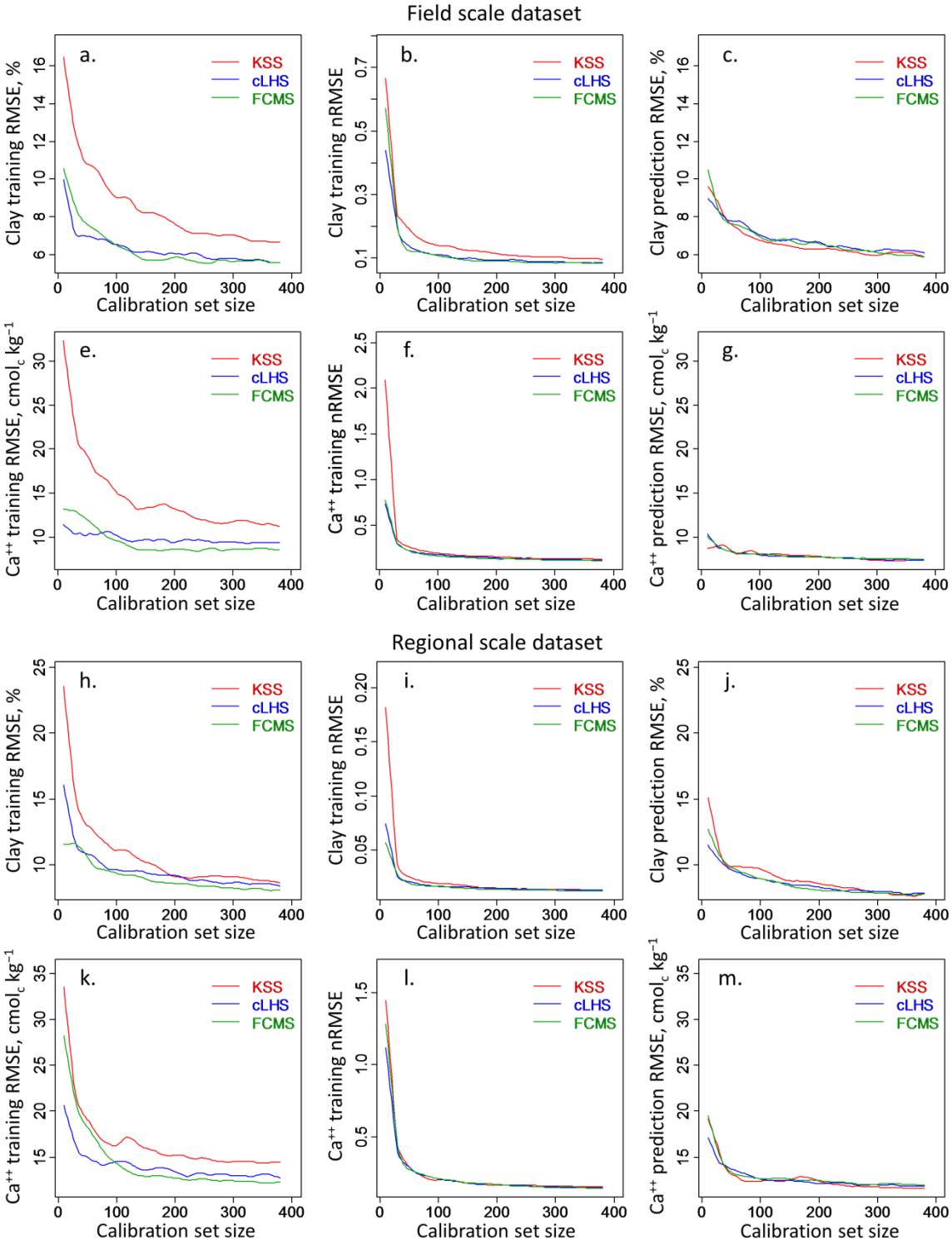

Figure 3. Training RMSE, nRMSE and prediction RMSE of CC and Ca^{++} against calibration set size in both datasets.

The highest training RMSEs for the field dataset were returned by the KSS algorithm (Figure 3a,e). In terms of the training RMSE the differences between KSS and both cLHS and FCMS were markedly high. In the case of clay content, for

calibration set sizes ≤ 90 samples the cLHS returned better results than the FCMS (Figure 3a). For example, the models of clay content calibrated with 50 samples produced training RMSEs of 10.8 % with the KSS, 7.0 % with the cLHS and 7.6 % with the FCMS. Similarly for the Ca^{++} models calibrated with 50 samples, the KSS produced the highest training RMSEs of 19.3 $cmol_c\ kg^{-1}$, while with the cLHS the training RMSE was 10.2 $cmol_c\ kg^{-1}$ and 12.0 $cmol_c\ kg^{-1}$ for the FCMS. For calibration set sizes > 90 samples the FCMS produced lower training RMSEs in comparison to the cLHS.

The highest training RMSEs in the regional dataset were also produced by the samples selected with the KSS algorithm (Figure 3h,k). In the case of clay content the differences between the KSS and both the cLHS and the FCMS in terms of the $RMSE_{tr}$ were markedly wider for calibration set sizes < 200 samples. In this case the cLHS presented lower performance than the FCMS. However, for the models of Ca^{++} the cLHS presented the lowest training RMSEs for calibration set sizes < 100 samples.

Concerning the nRMSEs in the field scale dataset, the FCMS and the cLHS returned very similar results, while the models corresponding to the KSS produced higher nRMSEs (Figure 3b,f), especially for calibration set sizes < 200 samples in the case of clay content. For the models of Ca^{++}, at calibration set size ≤ 30 the nRMSEs produced by the KSS samples were dramatically higher than those produced by the cLHS and the FCMS samples. For example, for the Ca^{++} models calibrated with 10 samples, the KSS returned a nRMSE of 2.09 while the cLHS produced an nRMSEs of 0.73 and the FCMS a value of 0.78. For calibration set sizes > 100 samples, the nRMSE of the models of Ca^{++} were very similar for the three sampling algorithms.

In the regional dataset for clay content the three sampling algorithms produced very similar nRMSEs for calibration set sizes > 230 (Figure 3i). Nevertheless, for calibration set sizes ≤ 20 the KSS samples produced much higher nRMSEs in comparison to the cLHS and the FCMS samples. On the other hand, in the case of the models of Ca^{++}, the three sampling algorithms produced comparable results in terms of nRMSE (Figure 3l).

The reason why the differences between the sampling algorithms were much wider for the $RMSEs_{tr}$ than for the nRMSEs, relies on the range of the soil attribute values of the samples selected by the algorithms. For example the ranges of the clay content values for a calibration set size = 150 samples selected with the KSS, cLHS and FCMS were 6–81%, 7–75% and 7–74 % respectively. Similarly for rest of the attributes and the rest of the calibration set sizes we observed that the KSS tends to select a wider range of values in comparison to the FCMS and the cLHS. This is due to the fact that the KSS algorithm selects extreme samples while the FCMS and the cLHS algorithms do not. Extreme samples can be advantageous for calibration in some cases, especially when the relationship between the predictors and the soil attribute is known (Minasny and McBratney, 2010). However these relationships are usually unknown. On the other hand, in the case of field vis–NIR measurements where many outlier samples can arise (due to uncontrolled conditions) the KSS would probably not be a good option since the outlier samples are included in the calibration set.

Based on the prediction RMSEs we found mixed results for the three sampling algorithms. At calibration set sizes ≤ 40, the cLHS returned the lowest prediction RMSE for clay content in the field dataset. However at calibration set sizes > 40 the KSS presented slightly lower results than both cLHS and FCMS. For Ca^{++} the three sampling algorithms produced comparable results, nevertheless at calibration set sizes ≤ 20 the KSS produced slightly lower results in comparison to the other algorithms. For the predictions of clay content in the regional dataset, the KSS was outperformed by both the cLHS and the FCMS. For calibration set sizes ≤ 90 the cLHS produced slightly lower prediction RMSEs than the FCMS, and for calibration set sizes >90 these algorithms produced comparable results. For the Ca^{++} predictions in the regional dataset we can divide the calibration set sizes in the three following regions: 10–30, 40–120, 130–380 samples. In the first calibration set size the KSS and the FCMS returned very similar results while the cLHS produced the lowest prediction RMSEs; in the second calibration set size region the KSS produced lower results than the other algorithms; and in the third calibration set size region the three algorithms produced comparable results.

Overall, the results obtained for both datasets are similar. The differences between the errors produced by samples selected by the different algorithms are larger at low calibration set sizes. Apparently a large calibration set size ensures a good coverage of the PC space and therefore the differences between sampling strategies in terms of the uncertainty of the models error should be lower.

4.3 Sample representativeness in the PC space

For all the three algorithms the absolute difference between the sample variance and the original variance ($|s^2 - \sigma^2|$) as well as the difference between the sample mean and the original mean ($|\overline{x} - \mu|$) decreased as the calibration set size increased. In other words, the original distribution of the PCs can be better replicated by increasing the calibration set size.

The s^2 and the $\overline{x}$ of the calibration sets selected by the cLHS showed the highest similarity to their population equivalents (σ^2 and μ) (Figure 4). These results were expected since the cLHS algorithm is a stratified sampling based on the distribution of the variables, while the KSS and the FCMS are distance–based algorithms. Nevertheless, we consider that a good sampling strategy must ensure both: a good coverage of the predictor space and a good replication of the original distribution which can result advantageous for models calibration.

In general, for calibration set sizes ≥ 130 samples the cLHS and the FCMS produced comparable results; however at calibration set sizes < 130 samples the cLHS returned much lower differences. Comparing the KSS to both the cLHS and the FCMS in terms of the differences between s^2 and σ^2 and also between $\overline{x}$ and μ, the KSS was largely outperformed by the other algorithms.

Based on these results our expectation is that the analyses between the sample distribution and the original distribution at different calibration set sizes can be very useful for identifying an adequate calibration set size. For example, we found that the $|\overline{x} - \mu|$ of the cLHS in the field dataset becomes relatively stable at a calibration set size of 120 samples (Figure 4b). In this sense this information could be used as criteria to set 120 as the adequate calibration set size taking in-

to account also that no considerable variations in $|s^2 - \sigma^2|$ were observed. Similarly this kind of analysis can be used for both FCMS and the KSS.

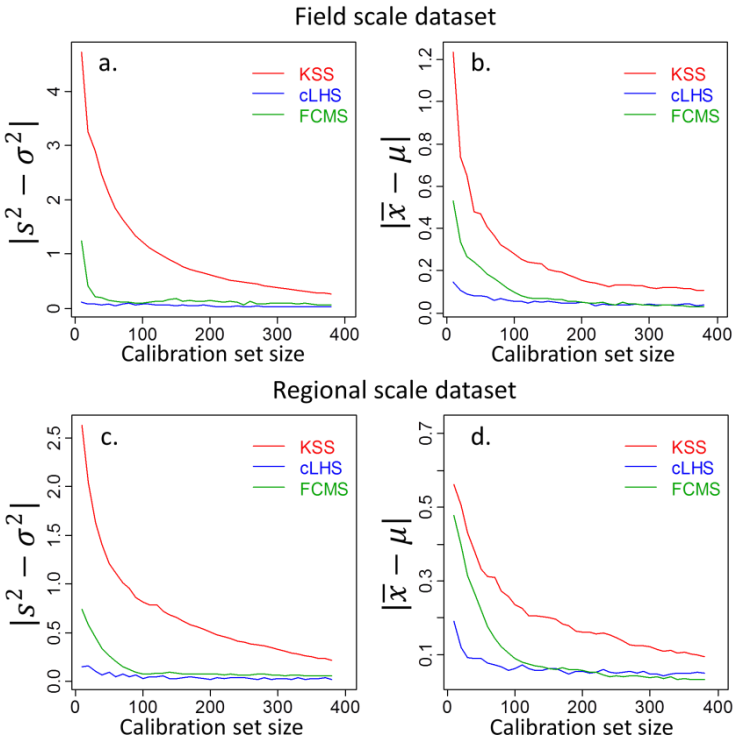

Figure 4. Calibration set size against the absolute difference between the sample variance (s^2) and the original variance (σ^2); and absolute difference between the sample mean ($\overline{x}$) and the original mean (μ).

5. Conclusions

In this work we investigated the effect of the calibration set size and three different calibration sampling strategies (Kennard-Stone, KSS; fuzzy c-means, FCMS; and conditioned Latin hypercube, cLHS) on the error of vis–NIR models calibrated for clay content and Ca^{++}. We also analyzed the sample representativeness on the basis of the sampling strategies and we proposed a method for identifying the optimal calibration set size based only on the analysis of the vis–NIR data (i.e. without prior knowledge of the soil attributes to be predicted).

We found that the error of the soil vis–NIR models depends on the calibration set size. Particularly for low calibration set sizes the errors are higher probably due to insufficient coverage of the predictor space. In this respect, when the number

of calibration samples is relatively low the sampling algorithm plays a critical role on the accuracy of the vis–NIR models.

The highest training errors were returned by the KSS. However this algorithm tends to select samples with a wider range of soil attribute values in comparison to the cLHS and the FCMS algorithms. This is due to the fact the KSS selects extreme samples. In this sense we believe that the inclusion of extreme samples in the calibration set can be beneficial when the dataset does not contain outlier samples. In terms of the prediction errors, the three sampling algorithms returned comparable results.

Concerning the sample representativeness in the PC space, for all the algorithms we found that the original distribution of the vis–NIR data in the principal component (PC) space can be better replicated by increasing the calibration set size. Our results showed that the samples selected by the cLHS and the FCMS algorithms better replicate the original distribution of the PCs in comparison to those selected by the KSS algorithm. For low calibration set sizes the cLHS better replicated the original distribution of the PCs in comparison to the FCMS. However at calibration set sizes ≥ 130 the cLHS and the FCMS produced comparable results. We consider that the comparison between the distribution of the calibration set and the original distribution of the population of samples is an adequate strategy for identifying an optimal calibration set size without any explicit knowledge of the soil attributes to be predicted. Furthermore, for the calibration of models it can be beneficial to select a calibration sample set whose distribution is close or equal to the distribution of the population.

Acknowledgements

We thank the State of São Paulo Research Foundation (FAPESP) for funding part of this research (process 07/58656–8).

References

Aizerman, M., Braverman, E., Rozonoer, L. 1964. Theoretical foundations of the potential function method in pattern recognition learning. Automation and Remote Control 25, 821–837.

Ben–Dor, E. Taylor, G.R. Hill, J. Demattê, J.A.M. Whiting, M.L. Chabrillat, S. and Sommer, S. 2008. Imaging spectroscopy for soil applications. Advances in Agronomy 97, 321–392.

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

Bezdek, J.C. 1981. Pattern Recognition with Fuzzy Objective Function Algorithms, Plenum, NY.

Bramley, R. G. V. E., Janik L.J. 2005. Precision agriculture demands a new approach to soil and plant sampling and Analysis – Examples from Australia. Communications in Soil Science and Plant Analysis 36, 9–22.

Brown, D.J., Bricklemyer, R.S., Miller, P.R. 2005. Validation requirements for diffuse reflectance soil characterization models with a case study of VNIR soil C prediction in Montana. Geoderma, 129, 251–267.

Camargo, M. N. Klant, E. and Kauffman, J. H. 1987. Classificação de solos utilizada em levantamentos pedológicos no Brasil. Campinas. Boletim Informativo da Sociedade Brasileira de Ciência do Solo, Viçosa, 12. n.1, 11–13.

Daszykowski, M., Walczak, B., Massart, D.L. 2002. Representative subset selection, Analytica Chimica Acta 468, 91–103.

de Gruijter, J.J .and McBratney, A. 2010. Sampling for High-Resolution Soil Mapping. In: Proximal Soil Sensing, Progress in Soil Science, edited by R. A. Viscarra Rossel, A. B. McBratney, B. Minasny, Springer Netherlands, Netherlands, p. 3–14.

De Maesschalck, R., Jouan–Rimbaud, D., Massart, D.L. 2000. The Mahalanobis distance. Chemometrics and Intelligent Laboratory Systems 50, 1–18.

Demattê, J.A.M, Campos, R.C. Alves, M.C. Fiorio, P.R. and Nanni, M.R. 2004. Visible–NIR reflectance: a new approach on soil evaluation. Geoderma 21, 95–112.

Demattê, J.A.M. and Garcia, G.J. 1999. Alteration of soil properties through a weathering sequence as evaluated by spectral reflectance. Soil Science Society of America Journal 63, 327–342.

Drucker, H., C. J. C. Burges, L. Kaufman, A. Smola, and V. Vapnik. 1996. Support vector regression machines. In: Advances in Neural Information Processing Systems (NIPS), Vol. 10. pp. 155–161.

Dunn, J.C. 1973. A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well-Separated Clusters. Journal of Cybernetics, 3, p.32–57.

Fernandes, R. B. A. Barrón, V. Torrent, J. and Fontes, M. P. F. 2004. Quantificação de óxidos de ferro de Latossolos brasileiros por espectroscopia de reflectância difusa. Revista Brasileira de Ciencia do Solo 28, 245–257.

Fu, X. , Ying, Y. , Yang, D. 2011. A comparative study of representative subset selection for NIR model updating. American Society of Agricultural and Biological Engineers Annual International Meeting 2011 4, 3411-3421

Gogé, F., Joffre, R., Jolivet, C., Ross, I., Ranjard, L. 2012. Optimization criteria in sample selection step of local regression for quantitative analysis of large soil NIRS database, Chemometrics and Intelligent Laboratory Systems 110, 168–176.

Grinand, C., Barthes, B.G., Brunet, D., Kouakoua, E., Arrouays, D., Jolivet, C., Caria, G., Bernoux, M. 2012. Prediction of soil organic and inorganic carbon contents at a national scale (France) using mid-infrared reflectance spectroscopy (MIRS). European Journal of Soil Science 63, 141–151.

IUSS Working Group WRB 2006. World Reference Base for Soil Resources 2006. 2nd edition. Rome FAO, World Soil Resources Reports No. 103.

Kennard, R.W., Stone, L. 1969. Computer aided design of experiments. Technometrics 11, 137–148.

Kim, H. J., Sudduth, K. A., and Hummel, J. W. 2009. Soil macronutrient sensing for precision agriculture. Jounal of environmental monitoring 11, 1810–1824.

Kuang, B., Mouazen, A.M. Influence of the number of samples on prediction error of visible and near infrared spectroscopy of selected soil properties at the farm scale. European Journal of Soil Science 63, 421-429.

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

McKay M.D., Conover W. J. and Beckman, R. J. 1979. A comparison of three methods for selecting values of input variables in the analysis of output from a computer code. Technometrics, v. 21, p. 239-245.

Minasny,B.and McBratney, A. 2010. Conditioned Latin hypercube sampling for calibrating soil sensor data to soil properties. In: Proximal Soil Sensing, Progress

in Soil Science, edited by R. A. Viscarra Rossel, A. B. McBratney, B. Minasny, Springer Netherlands, Netherlands, p. 15–28.

Minasny. B. and Mcbratney. A. B. 2006. A conditioned Latin hypercube method for sampling in the presence of ancillary information. Computers and Geosciences 32, 1378–1388.

Odeh, I.O.A., McBratney, A.B., Chittleborough, D.J. 1992. Soil pattern recognition with fuzzy-c-means: application to classification and soil-landform interrelationships. Soil Science Society of American Journal, v. 56, p. 505–516.

R Development Core Team. 2011. R: A language and enviroment for statistical computing. Vienna: R Foundation for Statistical Computing.

Raij, B., Van. Quaggio, J.A., Cantarella, H., Ferreira, M. E., Lopes A. S., Bataglia. C.O. 1987. Análise química do solo para fins de fertilidade. Campinas: Fundação Cargill, 170.

Rodionova, O.Y. , Pomerantsev, A.L. 2008. Subset selection strategy. Journal of Chemometrics 22, 674-685.

Shepherd, K.D. and Walsh, M.G. 2002. Development of reflectance spectral libraries for characterization of soil properties. Soil Science Society of America Journal 66, 988–998.

Siano, G.G., Goicoechea, H.C. 2007. Representative subset selection and standardization techniques. A comparative study using NIR and a simulated fermentative process UV data. Chemometrics and Intelligent Laboratory Systems 88, 204–212.

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, 1792–1799.

Viscarra Rossel, R., and Behrens, T. 2010. Using data mining to model and interpret soil diffuse reflectance spectra. Geoderma 158, 46–54.

Viscarra Rossel, R.A., Jeon, Y.S., Odeh, I.O.A., McBratney, A.B. Using a legacy soil sample to develop a mid-IR spectral library. Australian Journal of Soil Research 46, 1-16.

Wetterlind, J., Stenberg, B. and Soderstrom, M. 2010. Increased sample point density in farm soil mapping by local calibration of visible and near infrared prediction models. Geoderma 156, 152–160.

White, K., Walden, J., Drake, N., Eckardt, F., Settle, J. 1997. Mapping the iron oxide content of dune sands, Namib sand sea, Namibia, using landsat thematic mapper data. Remote Sensing of Environment 62, 30–39.

Wu, W., Walczak, B., Massart, D.L., Heuerding, S., Erni, F., Last, I.R., Prebble, K.A. 1996. Artificial neural networks in classification of NIR spectra data: design of the training set. Chemometrics and intelligent laboratory systems 33, 35–46.

Zhu, X., Shan, Y., Li, G., Huang, A., Zhang, Z. 2009. Prediction of wood property in Chinese Fir based on visible/near-infrared spectroscopy and least square-support vector machine. Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy 74, 344–348.

Manuscipt 2: A comparison of calibration sampling schemes at the field scale

Geoderma, submitted on June 2012

Karsten Schmidt[a], Thorsten Behrens[a], Jonas Dauman[a],
Leonardo Ramirez-Lopez[a,b], Ulrike Werban[c], Peter Dietrich[c]
and Thomas. Scholten[a]

[a]Institute of Geography, Physical Geography and Soil Science, University of Tübingen,
Rümelinstraße 19–23, 72070, Tübingen, Germany.

[b]Georges Lemaître Centre for Earth and Climate Research, Earth and Life Institute,
Université Catholique de Louvain,
3 Place Louis Pasteur – 1348, Louvain la Neuve, Belgium.

[c]Department of Monitoring and Exploration Technologies, UFZ, Helmholtz Centre for environmental Research, Leipzig, Germany

Abstract

High-resolution digital soil sensing and mapping is an important and emerging field to meet the strong and growing global demand for high-resolution soil property data. Yet, the combination of geophysical sensing and pedometrical techniques to produce soil property maps is complex and requires a well structured design from data collection to model validation. We compare different sampling design strategies – an extension of conditioned Latin Hypercube Sampling, Fuzzy k-means Sampling and Response Surface Sampling – as a basis for predicting soil texture, soil organic carbon and soil pH-value at two soil depth intervals using electromagnetic induction (EM38DD and EM31) and gamma spectroscopy (U, K, Th) data. Two different sample set sizes, two different regression approaches (Multiple Linear Least Squares and Random Forests), as well as several resampling and independent validation approaches are compared. The results show that a combination of Latin Hypercube Sampling and Random Forests re-

gression should be recommended, since Latin Hypercube Sampling shows the best spread within the state space of the sensors and the best independent validation results for both sample set sizes. The comparison between the validation approaches reveals a complex picture and points to the necessity of adequate independent validation approaches. Yet, leave-group-out cross-validation and .632 bootstrapping can be recommended as the best estimate in this study. Future work should focus on detailed analysis of Latin Hypercube Sampling and why it outperformed the other approaches. Therefore, comparisons with other sampling approaches should be conducted as well as specific sampling-for-validation approaches. Concluding, this study shows that there are complex interactions between sampling design, regression approaches and validation approaches, which can have a high influence on the final soil property maps and their accuracy estimates.

Keywords: weighted Latin Hypercube sampling, fuzzy k-means sampling, response surface sampling, regression, validation, soil sensing, iSOIL.

1. Introduction

High resolution digital soil sensing and mapping is an emerging research topic in soil science and environmental research (Viscarra-Rossel et al., 2010). The aim is to use on-the-go geophysical measurements from different sensors to get detailed information about soil spatial distribution, which then can be used for example to optimize fertilization. This additional financial and technical expenditure requires a careful planning of the whole process, from sampling to producing digital soil maps, to make it efficient and feasible. As in any scientific discipline data collection is most crucial since all subsequent analysis steps rely on the calibration samples selected. Generally, a sample set drawn from a population should reflect the entire population as well as possible. And, again with regard to efficiency, it should be small.

Concerning the location of samples two different concepts can be chosen for digital soil mapping studies on the field scale: the coverage of the feature space or the

coverage of the geographical space. If existing information is available in terms of relevant environmental covariates, as is the case in high resolution digital soil sensing and mapping studies, a coverage of the feature space should be chosen and regression approaches should be applied to spatially map measured soil properties.

Several approaches are documented to select samples based on available covariates and to cover the feature space (Brus and Heuvelink, 2007; de Gruijter et al., 2010; Lesch, 2005; Minasny and McBratney, 2006). Yet, comparisons of different sampling approaches are largely missing most probably to the restricted budged for field work and lab analysis. Comparisons on sample size in relation to the calibration sampling strategy are required especially for small sample sizes (Brown et al., 2005). Additionally, Minasny and McBratney (2010) point out to the importance to investigate the connection between the calibration sampling method and the prediction accuracy of models.

Within this study we compare three different sampling schemes, two different sampling sizes and two different regression approaches against several cross-validation and independent validation approaches to select calibration samples for a test site and to build soil property models. The Rosslau field site (0.36 km^2) is located close to the Elbe River in Saxony-Anhalt, Germany. It is a grassland site with a high variation in soil texture and is therefore a good test bed for this study. The variation stems from different sediments of the Elbe river.

We compare (i) a weighted version of conditioned Latin hypercube sampling (cLHS; Minasny and McBratney, 2006) which also samples the extremes, (ii) fuzzy k-means sampling (FKMS; de Gruijter et al., 2010; Minasny and McBratney, 2002) and response surface sampling (RSS; Lesch et al., 1995; Lesch, 2005).

This setup allows for various comparisons and analysis to answer questions such as:

- Are there differences in prediction accuracy between the sampling schemes?
- Are there differences in prediction accuracies between the regression approaches?

- Is the influence of the sampling schemes higher (or lower) compared to the regression approaches?
- Is there an optimum combination of sampling scheme and regression approach?
- Are the predictions of the different sampling and regression approaches comparable?

2. Material and methods

2.1 Sampling design

2.1.1 Weighted conditioned Latin hypercube sampling with extremes

The idea behind conditioned Latin hypercube sampling (cLHS; Minasny and McBratney, 2006) is to cover the state space of all covariates by maximally stratifying the marginal distribution while also preserving the correlation between the covariates in the sample set. The Latin hypercube is constructed by random sampling from the cumulative distributions of the covariate data using a simulated annealing optimization approach, which additionally focuses on preserving the correlation between the covariates in the selected sample set.

Based on the implementation of Minasny (2004) we extended cLHS in two ways. First we applied a weighting scheme to account for the fact that different sensors provide signals of different accuracy and noise. The assumption is that the higher the Kriging (cf. section 2.4) cross-validation accuracy, in terms of the correlation coefficient, the lower the noise of the signal and the higher the likelihood for providing a stable and reliable indicator for calibration. Hence, we used the 10-fold cross-validation correlation coefficients as weights. The higher the weight the higher the priority in the simulated annealing optimization approach, i.e. to optimize towards an equal sampling of n = 1 across the strata of the quantile distribution of the corresponding covariate. Second, to ensure a full coverage of the sensor state space we set the extremes (min and max) of all sensors as fixed sampling locations (wecLHS).

2.1.2 Fuzzy k-means sampling

Fuzzy k–means sampling (FKMS) has been proposed as a calibration sampling method by de Gruijter et al. (2010) and was conducted using the FuzME package (version 3.5c; Minasny and McBratney, 2002). It is based on the well known k-means clustering algorithm extended by a fuzzy membership function (Bezdek, 1981; Dunn, 1973). It is expected that samples belonging to each cluster share similar characteristics while the dissimilarity between samples in different clusters is maximized. Hence, FKMS should show similar characteristics in terms of state space coverage as expected from cLHS.

The sampling locations are the centroids of fuzzy k-means clusters of discretization cells of the sensor data. The number of clusters determines the number of calibration samples to be selected.

2.1.3 Response surface sampling

Following a principle components analysis (PCA) of the sensor data, RSS applies a response surface design aiming to optimize the estimation of the regression model parameters when using ordinary least squares estimation as well as to minimize the effects of the spatially dependent autocorrelations (Lesch et al., 1995; Lesch, 2005). We used the ESAP software package (Lesch et al., 2000) in this study. The major restriction of ESAP is the limited influence the user has on the number of samples, since only 6, 12, or 20 samples can be selected (Lesch et al., 2000).

2.2 Sample set sizes

Generally, the required sample set size depends on the soil spatial variability. Yet, the biggest constrain is the expense of data collection and lab analysis. Within the frame of the EU FP7 iSOIL project we agreed to collect 30 samples for each sampling scheme and each testing site. This was conducted for wecLHS and FKMS in this study. RSS was considered as an additional reasonable sampling scheme but due to the restriction described above only 20 samples were collected in this study. To make the different sample set sizes comparable and also to ana-

lyze the effect of different sample set sizes, we produced subsets of 20 samples for wecLHS and FKMS. For both approaches we tested repeated (10 times) random subsets and averaged the validation accuracies ($wecLHS_{rand}$ and $FKMS_{rand}$). For wecLHS we also sampled a wecLHS subset ($wecLHS_{sub}$), which thus should show the same effects as the original wecLHS sample set in terms of state space coverage, correlation between the sensor signals and the coverage of the extreme values of the sensors in a reduced set of 20 samples. As a result we compare 6 different approaches with two different sample set sizes in this study. A further reduction of sample sizes via sampling or subsampling was not considered, since the differences in the validation accuracies between the sampling designs were already large for some soil properties.

2.3 Field sampling

Samples were taken as composite samples of 5 subsamples (Figure 1) within an area of 1 by 1 m according to the iSOIL Sampling protocol (Behrens et al., 2009). For each location soil pH, soil organic carbon (SOC) and soil texture (sand silt and clay) was analysed for the depth intervals of 0-10 cm and 10-30 cm. These depth intervals had been chosen in the broader context of the iSOIL project aiming in analysing the effect and importance of different sensors in predicting soil properties at different depth across different field sites. Therefore, the results might also be an indicator for choosing a soil depth in comparable studies.

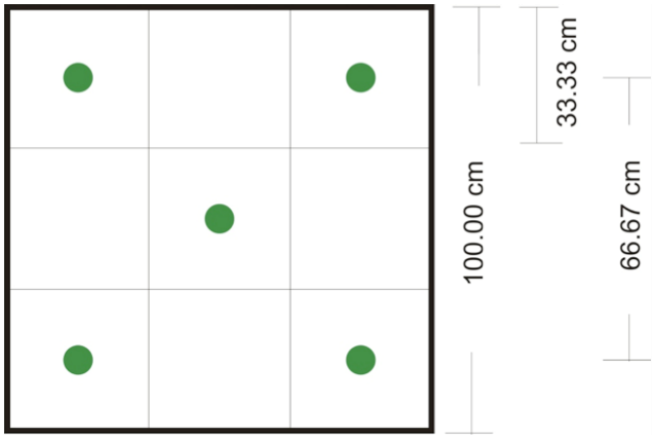

Figure 1. Relative location of the sub-samples for the composite sample collection.

2.4 Soil sensing

Mobile geophysical platforms have been equipped with frequency-domain electromagnetic-induction instruments and a portable gamma-ray spectrometer. The devices were placed on a sledge including Differential Global Positioning System (DGPS) dragged by a tractor.

Measurements of electrical conductivity (ECa in mS/m) were made using an EM38DD and an EM31 (Geonics Ltd.). Both are widely spread instruments for near surface applications that measure the vertical and horizontal dipole orientation, respectively, resulting in different effective depths of exploration of 0.75 m and 1.5 m for the EM38DD signals and 6 m for EM31, respectively (Callegary, 2007).

A portable gamma-ray spectrometer (4l NaI(Tl) – crystal, automatic peak-stabilization, GF instruments) with 512 channels at an energy range between 100 keV and 3 MeV was used for the field measurements in Roßlau. We use a 5s sampling interval for gamma-ray measurements. The measured counts per second were converted into the concentration of K (in %), U and Th (in ppm) and the dose rate (in nGy/h). Investigation depth of the sensor is limited to about 30cm soil depth.

Even though the selection of the sensors might have been chosen according to the soil depth sampled and analysed, we extended the range of the sensors to deeper depth or larger volumes to account for the influence of potentially relevant contextual information and thus the potential interaction of the sensor signals in the regression models. In addition larger volumes might have an averaging effect in terms of filtering out noise, which might also have a positive effect in guiding the selection of samples.

Ordinary Kriging interpolation of the sensor data was conducted using the gstat package (Pebesma, 2004) in R (R Development Core Team, 2012). Finally, all sensor datasets were transformed linearly to the range of [0,1] for regression analysis.

2.5 Calibration

To calibrate the soil properties we compared multiple linear regression (MLR), and random forest regression (RF) using R (R Development Core Team, 2012) and the corresponding lm model (R Development Core Team, 2012) based on Wilkinson and Rogers (1973) and the randomForest package (Liaw and Wiener, 2009). RF is a machine learning method consisting of an ensemble of randomized classification and regression trees (Breiman et al., 1984; Breiman, 2001) where numerous trees are generated and finally aggregated to give one single prediction. In regression problems the prediction is the average of the individual tree outputs. Each tree of a forest is based on a bootstrap sample of the original training data. In addition to this bagging function (Breiman, 1996), the best split at each node of the tree is searched only among a randomly selected subset of predictors. All trees are grown to maximum size without pruning.

In contrast to RF, which is a non-parametric and non-linear approach, the multiple linear regression is used for examining linear correlations between multiple independent variables and the dependent variable. We apply the least square criterion for calibrating the model (Rao et al., 1999).

2.6 Validation

2.6.1 Resampling

Resampling (Good, 2006) can be used to estimate the generalization error of a calibration model when no independent validation set is available. Therefore, the dataset is repeatedly splitted into a calibration set and a validation set (Molinaro et al., 2005). The validation accuracies are then averaged over all validation results.

Within this study we compare several resampling approaches to relate the estimated generalization error to the generalization error of an independent sample set (section 2.6.2).

We tested:

i. 10-fold cross validation (10cv)

ii. Leave-group-out cross-validation (lgocv)

iii. bootstrapping (boot)

iv. .632 bootstrapping (.632boot)

In *10cv* 10 randomly non-overlapping subsets are generated. A model is build on 9 subsets and is validated against the remaining subset. This process is repeated 10 times until every subset was once used for validation.

For *lgocv*, *boot* and *.632boot* random subsamples are chosen as validation sets. For *lgocv* (or *repeated random sub-sampling validation* or *Monte Carlo cross-validation*; Molinaro et al., 2005) we randomly splitted the sample set into 75 % of the samples for calibration and 25 % without replacement for validation for 25 times.

A bootstrap sample is a randomly chosen (with replacement) set of n samples from a data set of n instances (Efron and Tibshirani, 1993). The probability of a given sample to appear in the validation set is *0.632n*. The *boot* validation accuracy is obtained from the samples that were not selected for the bootstrap (apparently *0.368n*).

The *.632boot* approach merges the overestimated prediction error and the underestimated resubstitution error (Efron, 1983; Efron and Tibshirani, 1993; Efron and Tibshirani, 1995; Molinaro et al., 2005). It is defined as (Kohavi, 1995):

$$\mathrm{acc}_{\mathrm{boot}} = \frac{1}{b}\sum_{i=1}^{b}(0.632 \cdot \epsilon 0_i + 0.368 \cdot \mathrm{acc}_{\mathrm{S}}) \tag{1}$$

where:

- $\mathrm{acc}_{\mathrm{boot}}$ = estimated .632 bootstrap accuracy
- b = number of bootstrap sample sets
- $\varepsilon 0_i$ = accuracy estimate of bootstrap sample set i
- acc_s = resubstitution accuracy estimate on the full dataset

Compared to the *boot* approach, the *.632boot* method is expected to return higher accuracies, since the resubstitution accuracy is included in the estimates. The

lgocv estimates are expected to be higher compared to *boot* because the model building is based on more samples in average.

We repeated all resampling validation approaches 10 times and report the averages for each method.

2.6.2 Independent validation

Based on the three different sampling approaches and thus sample sets compared in this study two of these sample sets can be used as independent validation sets for estimating the generalization error; e.g. the sample sets for FKMS and wecLHS can be used to validate the RSS calibration model.

To provide a comparable measure regarding the differences between the *boot* and the *.632 boot* methods we tested the following two approaches:

(i) fully independent, i.e. validation against the sample sets of the other two sampling approaches and

(ii) partially independent, i.e. validation against all sample points from all three sampling approaches.

This offers the option to analyse over- and underfitting effects of single sampling scheme vs. calibration model settings, when there are unexpected differences between the fully independent and partially independent as well as the *boot* and the *.632boot* validation approaches compared to the other sampling schemes and regression models.

3 Results and Discussion

3.1 Sampling

To avoid negative effects of multicollinearity only sensor data with cross-correlation values below 0.8 were selected for wecLHS and FKMS. Table 1 shows the results of the cross-correlation analysis. The selection between two sensors was based on the Kriging cross-validation results. The sensor with the highest R^2 cross-validation value of two highly correlated signals was selected while the oth-

er sensor was removed. Hence, the EM 31, K and Th sensor signals were used as input data to derive the sampling locations. Due to the principal component analysis in RSS this approach is not required here.

Table 1: Cross-correlation table of the geophysical datasets.

	EM 31	EM 38v	EM 38h	K	Th	U
EM 31 [Ec_a mS * m^{-1}]	1	0.91	0.85	-0.47	0.09	0.16
EM 38v [Ec_a mS * m^{-1}]		1	0.92	-0.58	0.05	0.18
EM 38h [Ec_a mS * m^{-1}]			1	-0.72	-0.11	-0.09
K [%]				1	0.50	0.48
Th [ppm]					1	0.86
U [ppm]						1

According to the Kriging cross-validation results the weights applied for the sensor data in the wecLHS approach are: 1 for EM31, 0.81 for Th, and 0.5 for K, which showed the lowest signal-to-noise ratio (Table 2).

Table 2: Kriging 10-fold cross-validation results of the geophysical measurements.

	R^2	**RMSE**	**Range**
EM 31 [Ec_a mS * m^{-1}]	0.9991	0.49	1.3 – 123.8
EM 38v [Ec_a mS * m^{-1}]	0.9972	0.78	2.2 – 102.4
EM 38h [Ec_a mS * m^{-1}]	0.9985	0.66	0 – 87.94
K [%]	0.50	0.13	0.18 – 1.06
Th [ppm]	0.81	1.13	1.78 – 10.1
U [ppm]	0.57	0.66	1.17 – 3.98

The location of the sampling sites is shown in Figure 2 for all initial sampling schemes.

Figure 2. Location of the sampling sites draped over an areal image (ESRI, 2012). wecLHS: Weighted conditioned Latin Hypercube Sampling, FKMS: Fuzzy k-means Sampling, RSS: Response Surface Sampling.

Not all samples could have been collected due to partially flooded sites. Hence, instead of the 30 sample locations for both wecLHS and FKMS only 29, and 28 respectively, were finally sampled and analysed.

The location of the samples in the state space is shown in Figure 3. All approaches almost cover the entire state space and thus should be well suited for calibrating a soil property model on the sensor data. Yet, there are slight differences: RSS and wecLHS show the widest range in covering the sensor data. wecLHS shows a distribution, which is less clustered compared to FKMS and RSS. The fact that wecLHS as applied in this study does not cover the full range of EM31 is due to the one location which was not sampled.

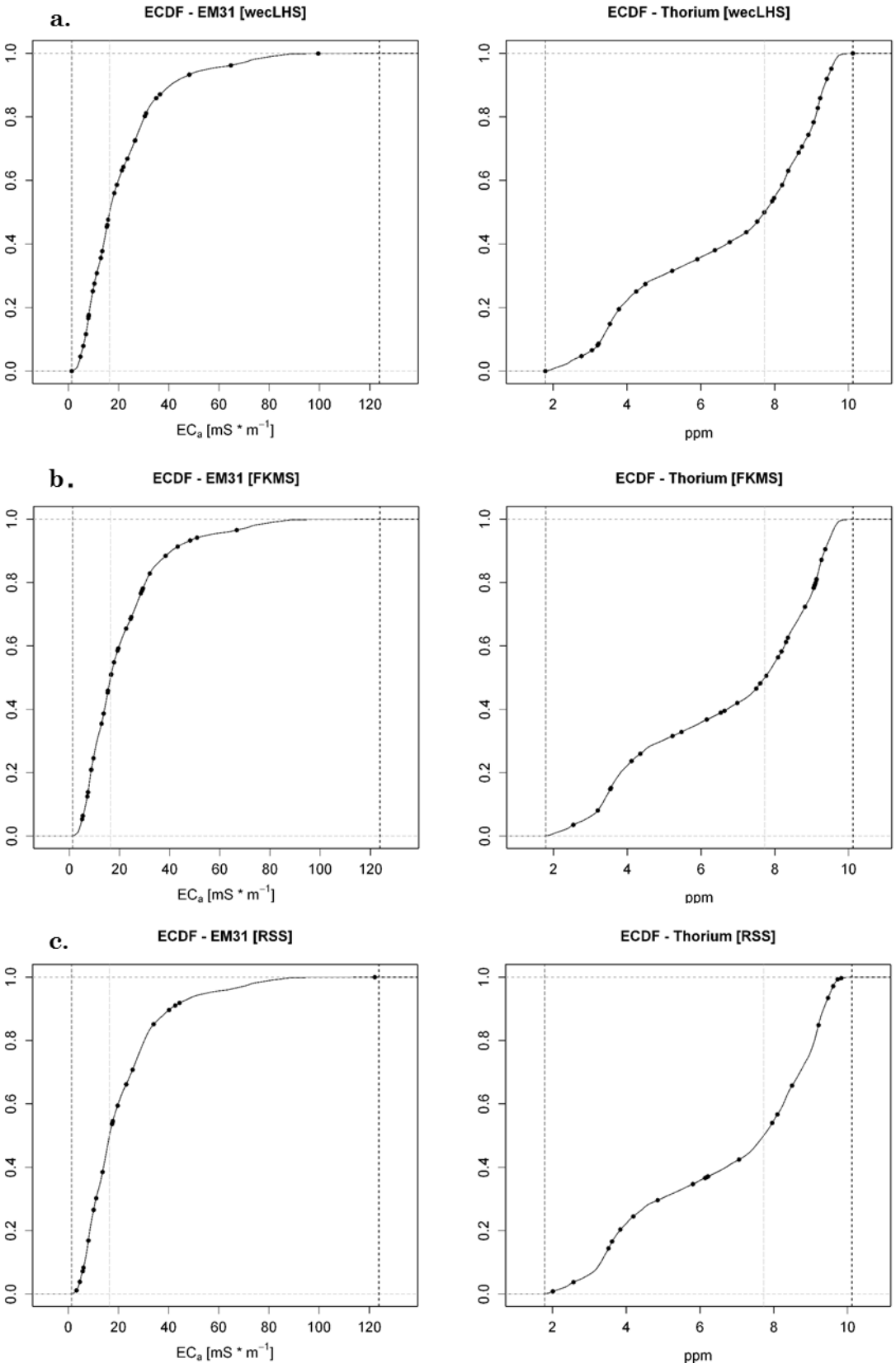

Figure 3. Location of the sampling sites in the state space shown for EM31 and Thorium signals. wecLHS: Weighted conditioned Latin Hypercube sampling, FKMS: Fuzzy k-means Sampling, RSS: Response Surface Sampling. Minimum, Median and Maximum are shown with vertical lines.

3.2 Soil properties

The lab analysis results are given in Table 3. SOC as well as the soil texture classes show a relatively wide range in the distribution, which is comparable between the two different soil depth layers.

Table 3. Lab analysis data for all sampling schemes.

Property	0-10 cm				10-30 cm			
	Min.	Median	Mean	Max.	Min.	Median	Mean	Max.
				All				
SOC	0.9	2.3	2.5	6.4	0.7	1.7	1.7	3.7
pH	4.5	5.3	5.4	7.4	4.9	5.9	5.9	7.6
Sand	5	18	32	87	4	18	31	87
Silt	8	34	32	49	8	34	32	75
Clay	4	40	36	60	1	40	36	59
				wecLHS				
SOC	1.4	2.3	2.6	6.4	0.7	1.6	1.7	3.7
pH	4.8	5.4	5.5	7.4	4.9	6.1	6.0	7.6
Sand	5	17	33	86	4	18	31	76
Silt	9	34	32	49	15	35	34	75
Clay	5	40	35	60	1	39	35	59
				FKMS				
SOC	1.0	2.3	2.5	6.4	0.8	1.7	1.7	3.5
pH	4.5	5.3	5.3	5.8	4.9	5.9	5.9	6.6
Sand	5	17	28	85	4	16	28	84
Silt	8	35	33	46	9	34	32	46
Clay	7	46	39	58	7	45	39	58
				RSS				
SOC	0.9	2.2	2.3	3.5	0.7	1.8	1.8	2.5
pH	4.6	5.3	5.4	7.0	5.0	5.9	5.8	7.4
Sand	10	27	37	87	11	25	36	87
Silt	8	30	30	48	8	31	31	48
Clay	4	38	32	53	4	39	33	52

Considering the sampling schemes separately, it can be seen that single soil properties of single sampling schemes show a reduced range in properties. This is for example the case for the combination of $SOC_{0\text{-}10cm}$/FKMS, $Silt_{10\text{-}30cm}$/wecLHS, and $Clay_{10\text{-}30cm}$/FKMS. Hence, no general assessment of the sampling schemes can be conduced based on the lab data.

3.3 Validation results

3.3.1 Independent validation

Table 4 summarizes the validation results as averages over all soil properties to reveal the general performance of the sampling schemes as well as the two cali-

bration approaches. The independent validation result is most important. The best combination for the initial sample set sizes (wecLHS = 29, FKMS = 28, and RSS = 20) is wecLHS and RF outperforming RSS/RF as well as FKMS/RF with about 10 % in variance explained. Fully and partially independent validation show comparable results, where the partially independent results suggest a performance that is in average about 4 % higher in terms of variance explained. This is as expected since the calibration data of the models were included. In the following we are focusing on the fully independent validation.

The comparison of the reduced $FKMS_{rand}$ and $wecLHS_{rand/sub}$ sampling sizes with RSS shows that LHS, with a decrease of about 4 % of variance explained compared to wecLHS with all samples, still outperforms the other approaches. The $wecLHS_{sub}$ approach outperforms FKMS and RSS with the original samples sizes in most cases.

Table 4. Average validation accuracies (**R^2**) based on cross-validation (*leave-one-out bootstrapping* [boot], *leave-group-out cross-validation* [lgocv], *.632 bootstrapping* [632boot], *and 10-fold cross validation* [10cv]), fully independent [f.ind] and partially independent [p.int] validation for different soil depth and for Random Forests [RF] and multiple linear regression models [MLR].

	wecLHS [obs. 29]		FKMS [obs. 28]		RSS [obs. 20]		$wecLHS_{rand}$ [obs. 20]		$wecLHS_{sub}$ [obs. 20]		$FKMS_{rand}$ [obs. 20]	
	MLR	RF	MLR	RF	MLR	RF	MLR	RF	MLR	RF	MLR	RF
						0-10 cm						
Boot	0.63	0.64	0.66	0.71	0.55	0.63	0.60	0.61	0.57	0.62	0.62	0.66
Lgocv	0.74	0.76	0.78	0.81	0.70	0.75	0.73	0.73	0.70	0.76	0.76	0.79
632boot	0.71	0.76	0.74	0.8	0.66	0.76	0.70	0.73	0.68	0.74	0.72	0.77
10cv	0.78	0.78	0.78	0.78								
f.ind	**0.75**	**0.80**	**0.63**	**0.68**	**0.67**	**0.66**	**0.70**	**0.74**	**0.68**	**0.78**	**0.62**	**0.65**
p.ind	0.78	0.86	0.68	0.75	0.71	0.72	0.71	0.78	0.72	0.82	0.66	0.70
						10-30 cm						
Boot	0.52	0.60	0.61	0.59	0.64	0.64	0.52	0.60	0.42	0.54	0.58	0.56
Lgocv	0.65	0.74	0.73	0.72	0.75	0.75	0.70	0.72	0.61	0.68	0.71	0.69
632boot	0.60	0.72	0.69	0.72	0.75	0.76	0.64	0.72	0.53	0.67	0.69	0.69
10cv	0.77	0.75	0.78	0.76								
f.ind	**0.70**	**0.71**	**0.56**	**0.63**	**0.56**	**0.59**	**0.63**	**0.65**	**0.67**	**0.65**	**0.54**	**0.59**
p.ind	0.73	0.80	0.63	0.74	0.66	0.69	0.67	0.71	0.66	0.69	0.59	0.69
						0-30 cm						
Boot	0.57	0.62	0.64	0.65	0.59	0.63	0.56	0.60	0.49	0.58	0.56	0.61
Lgocv	0.69	0.75	0.75	0.77	0.72	0.75	0.71	0.73	0.66	0.72	0.71	0.74
632boot	0.66	0.74	0.71	0.76	0.71	0.76	0.67	0.73	0.61	0.71	0.67	0.73
10cv	0.78	0.76	0.78	0.77								
f.ind	**0.73**	**0.75**	**0.59**	**0.66**	**0.61**	**0.62**	**0.66**	**0.70**	**0.68**	**0.72**	**0.66**	**0.62**
p.ind	0.75	0.83	0.65	0.75	0.69	0.70	0.68	0.74	0.70	0.76	0.68	0.69

In average RF performs 5 % better in variance explained compared to MLR. Yet, in terms of the variance explained the variation between the sampling schemes is higher compared to the regression approaches.

A closer look at specific soil properties reveals a more detailed picture. It can be seen that FKMS is only slightly outperformed by wecLHS, regarding the full sample sizes. Based on the reduced set however, $wecLHS_{sub}$ is the clear winner. RSS performs worst. $wecLHS_{rand}$ also outperforms RSS and $FKMS_{rand}$. Yet, compared to $wecLHS_{sub}$ it performs significantly worse in numbers of best performances across all soil properties (Table 5).

For the large sample sets MLR and RF are comparable. For the reduced set RF seems to have an advantage.

Table 5. Fully independent prediction accuracies (R2) of all soil properties measured based on predictions by Random Forests (RF) and multiple linear regressions (MLR). **Bold** and *italic* figures indicate the best and the worst prediction accuracy achieved by each sampling design (wecLHS, FKMS). In terms of a ranking we finally summed up all **best** and *worst* performances.

Property	wecLHS [obs. 29]		FKMS [obs. 28]		RSS [obs. 20]		$wecLHS_{rand}$ [obs. 20]		$wecLHS_{sub}$ [obs. 20]		$FKMS_{rand}$ [obs. 20]	
	MLR	RF	MLR	RF	MLR	RF	MLR	RF	MLR	RF	MLR	RF
						0-10 cm						
SOC	0.58	**0.65**	*0.57*	0.64	0.55	0.60	0.49	0.62	0.52	**0.67**	*0.46*	0.58
pH	0.52	**0.70**	0.15	*0.13*	0.27	0.30	0.45	0.55	0.41	**0.66**	0.16	*0.11*
Sand	**0.92**	0.91	*0.88*	0.91	**0.89**	*0.80*	**0.89**	0.87	0.86	**0.89**	**0.89**	0.88
Silt	0.87	0.87	*0.64*	**0.88**	0.75	0.75	0.80	0.84	0.76	**0.85**	*0.70*	0.82
Clay	0.88	0.87	**0.89**	*0.86*	**0.90**	*0.82*	0.85	0.83	0.86	0.84	0.87	0.83
						10-30 cm						
SOC	0.54	**0.63**	*0.44*	0.58	0.48	0.48	0.45	0.49	**0.53**	0.47	*0.40*	0.52
pH	**0.44**	0.38	0.09	*0.04*	0.23	0.23	0.36	0.32	**0.44**	0.33	0.11	*0.04*
Sand	**0.85**	0.83	*0.75*	0.79	*0.70*	*0.70*	0.78	0.83	0.77	**0.82**	0.74	0.72
Silt	0.82	0.83	*0.62*	**0.88**	*0.57*	0.74	0.70	0.81	0.78	**0.83**	0.61	0.82
Clay	0.86	*0.85*	**0.88**	**0.88**	*0.80*	*0.80*	0.84	0.81	0.85	0.81	0.84	**0.86**
						0-30 cm						
SOC	0.56	**0.64**	*0.50*	0.61	0.52	0.54	0.47	0.56	0.53	**0.56**	*0.43*	0.55
pH	0.48	**0.54**	0.12	*0.09*	0.25	0.27	0.40	0.43	**0.43**	**0.43**	0.14	*0.08*
Sand	**0.88**	0.87	*0.81*	0.85	0.79	*0.75*	0.83	**0.85**	0.81	**0.85**	0.81	0.80
Silt	0.84	0.85	*0.63*	**0.88**	*0.66*	0.75	0.75	**0.83**	0.77	**0.83**	*0.66*	0.82
Clay	0.87	*0.86*	**0.89**	0.87	0.85	*0.81*	0.85	0.82	**0.86**	0.82	0.85	0.85
Count **best**/*worst* performance	**4**/*0*	**5**/*2*	**3**/*6*	**4**/*4*	**2**/*4*	**0**/*5*	**1**/*0*	**2**/*0*	**4**/*0*	**10**/*0*	**1**/*5*	**1**/*3*

The calibration works best for soil texture followed by SOC and pH. RSS and especially FKMS fail to reasonably calibrate pH models. Generally, the upper layer returns slightly better prediction results (Figure 4).

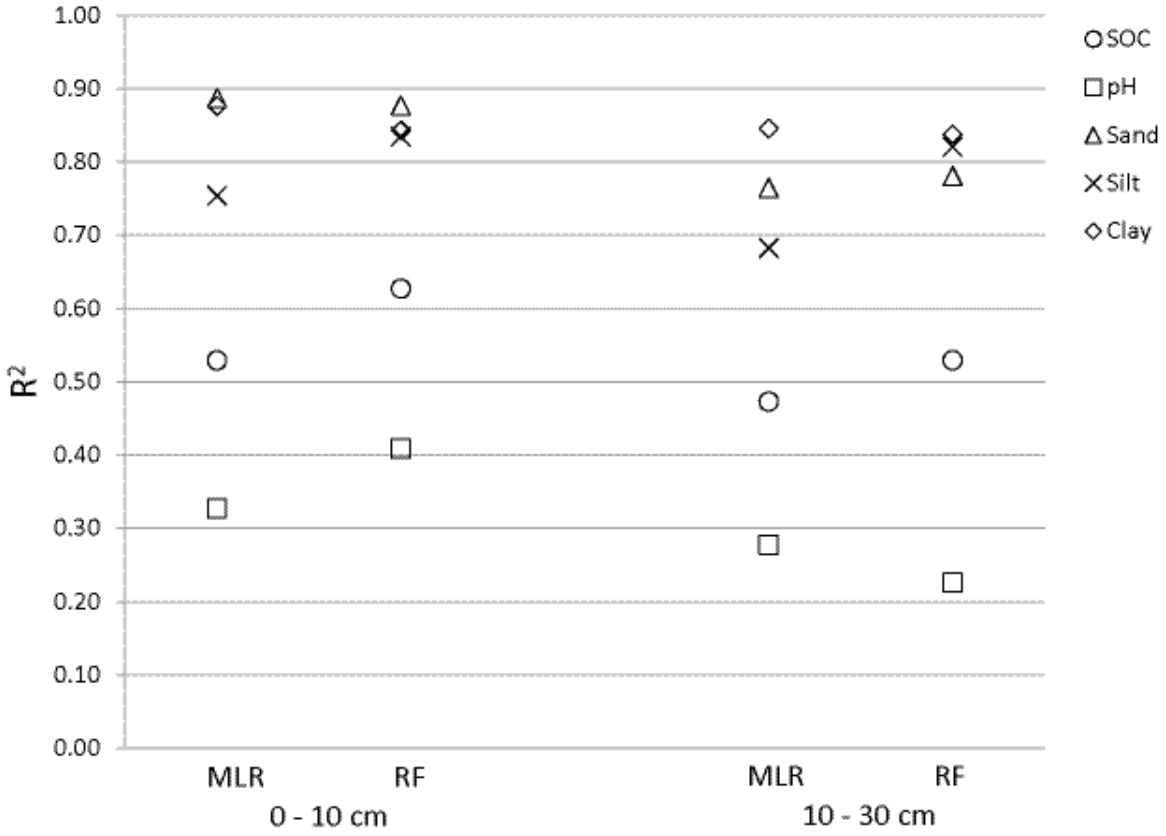

Figure 4. Average prediction accuracies (R^2) of all sampling approaches for all soil properties measured and fully independent validation based on predictions by Random Forests (RF) and multiple linear regression (MLR)

3.3.2 Cross-validation

In general it can be expected that cross-validation approaches suggest better performances compared to independent validation. This is the case for RSS and the FKMS approaches. Here, cross-validation in average performs about 3 % better with respect to the variance explained for MLR and RF respectively, compared to the fully independent validation. In contrast for wecLHS and $wecLHS_{sub}$ (Table 4) the fully independent validation approach returns better results (6 % in average) compared to the resampling validation approaches.

The order of the cross-validation approaches in terms of decreasing estimated variance explained is *cv* > *lgocv* > *.632boot* > *boot*. This is basically related to the number of samples left out for validation for each subset: 10% < 20 % < 36,8 % (including weighted average of the samples used for calibrating the model), and < 36.8 % respectively. Hence, the more samples excluded from the model the more complicated to calibrate the model. For small sample sizes this effect becomes more important since all samples are required to calibrate the model.

The *.632boot* approach is intended to account for this fact. In average it is the best estimate for the independent validation set. Yet in some cases it over- or underestimates up to 19 %. In average it is an underestimation of 5 %. Compared to the partially independent validation the *.632boot* approach shows the highest similarity in terms of calibration accuracy.

Ten times *10cv* does not seen to be an appropriate estimator for small sample sets since the validation results are too optimistic and at least 3 samples in a subset are required for reasonable results and interpretation.

3.4 Digital soil maps

To analyse the difference between the sampling and the regression approaches in terms of the final digital soil maps, we use the sand content (0-10 cm), since this soil property returned the highest independent validation accuracies (R^2 = 0.80 to 0.92). For this specific case over all three initial sampling schemes, MLR provided slightly better results (2.3 %) in average. For the two larger sample sets (wecLHS and FKMS) RF produced slightly better results of about 1 %.

The digital soil maps of sand content [%] for 0-10 cm are shown in Figure 5 for RSS, FKMS, wecLHS and $wecLHS_{sub}$. It can be seen that the maps for FKMS, wecLHS and $wecLHS_{sub}$ are very similar for each regression approach, whereas there are differences between the maps produced by RF and MLR. For both regression approaches the RSS maps show comparable differences to the other sampling approaches, especially in the leftmost site. This area is frequently flooded so the water regime is different at this site and thus the relation between soil properties and sensor signals. Since this effect is visible across all sampling

schemes it is also very likely that the differences result from non-linearity in the spatial soil property – sensor signal relation, which should be accounted for better in RF, and which is confirmed by the results (cf. section 3.3).

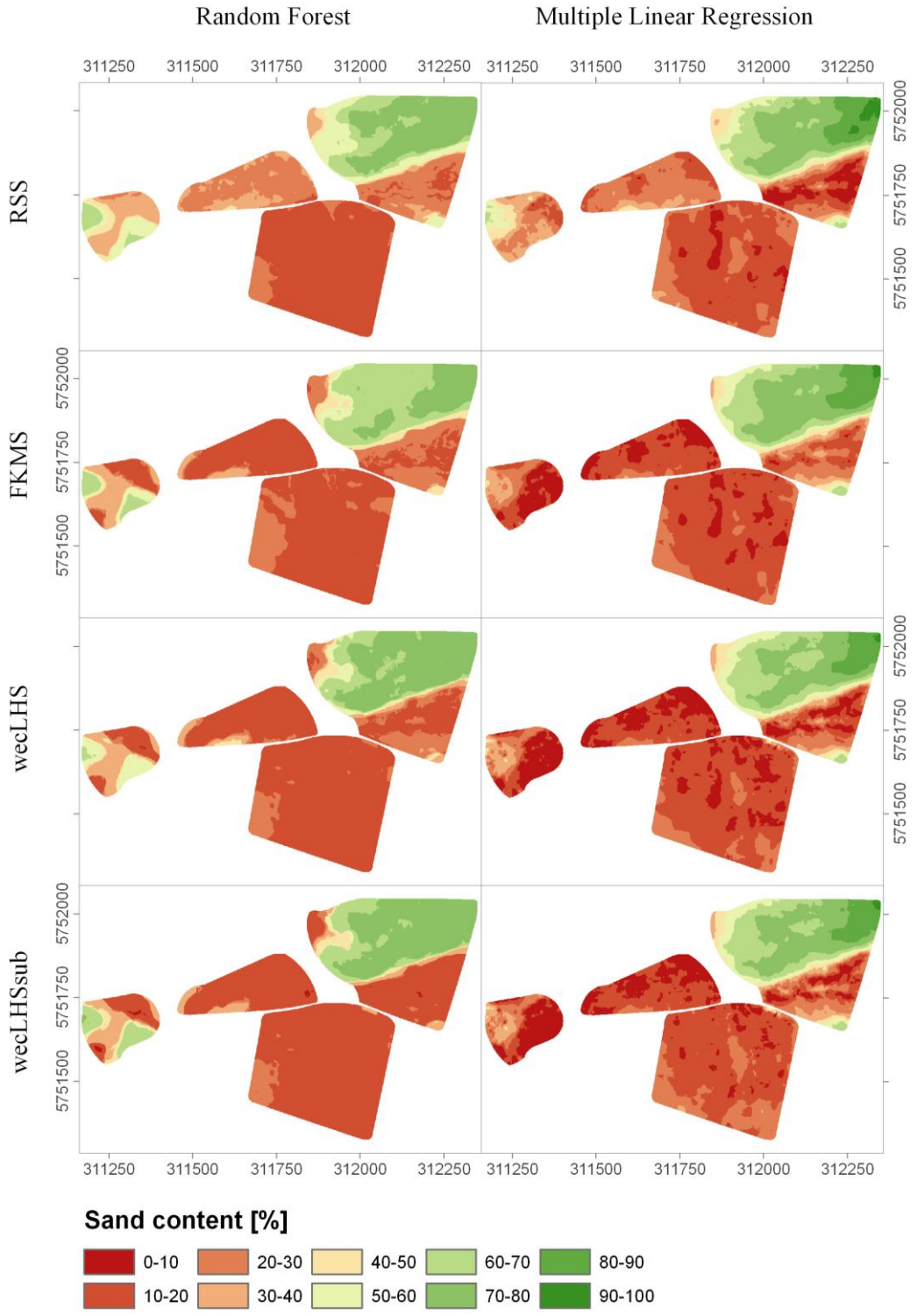

Fig. 5: Digital soil maps of sand content (%; 0-10 cm) produced with Random Forests (RF) and multiple linear regression (MLR) for the sampling approaches RSS, FKMS, wecLHS and wecLHS$_{sub}$.

Interestingly the linear model (MLR) shows more details and a wider range of sand content compared to the non-linear RF model. The wider range is a fact of the underlying computational approach – binary partition tree vs. a regression formula – where the predicted output of the tree is restricted to the range of the calibration data. Since the accuracy of both approaches is the same the finer details shown in the MLR maps compared to the RF maps must be interpreted as spurious accuracy.

Table 6 shows the correlations between the maps produced. For sand content (0-10cm) the differences between the regression approaches are higher compared to the sampling approaches.

Table 6. Correlation between the digital soil maps of sand content (%; 0-10 cm) produced with Random Forests (RF) and Multiple Linear Regression (MLR) for the sampling approaches RSS, FKMS, wecLHS and $wecLHS_{sub}$.

Sand 0-10 cm		RF				MLR			
		FKMS	wecLHS	RSS	$wecLHS_{Sub}$	FKMS	wecLHS	RSS	$wecLHS_{Sub}$
RF	FKMS	1	0.99	0.92	0.98	0.82	0.81	0.89	0.78
	wecLHS		1	0.92	0.98	0.84	0.83	0.90	0.80
	RSS			1	0.89	0.80	0.78	0.91	0.74
	$wecLHS_{Sub}$				1	0.77	0.77	0.85	0.74
MLR	FKMS					1	0.99	0.92	0.98
	wecLHS						1	0.92	0.99
	RSS							1	0.89
	$wecLHS_{Sub}$								1

3.5 Discussion

The most interesting results of this comparison are that:

- validation accuracy strongly depends on the sampling scheme
- 20 samples produce comparable results to ~30 samples
- wecLHS outperforms the other sampling approaches
- even with high validation accuracies the resulting digital soil maps show pronounced differences between the regression approaches
- independent validation should be recommended
- the *.632boot* resampling-validation approach seems to be the best estimator of the independent validation

Concerning sampling the advantages of the proposed wecLHS approach compared to FKMS and RSS are (i) the evenly covered state spaces of the sensor data, (ii) the preservation of the correlation between the sensor signals in the sample set, (iii) the inclusion of the extreme sensor values and (iv) the weighing scheme, which helps to focus on the strongest soil response. However, based on the results of this study it is relatively hard to differentiate between the effects of these three components. Even the result that the $wecLHS_{sub}$ approach with only 20 samples performs nearly as good as the original wecLHS based on 29 samples does not help to reveal the influence of a single factor.

The effect of the extremes might apparently be of major influence. However, since it is the extremes of the sensor data which are covered and not the extremes of the soil properties (Table 3) there is no influence of the range of the soil property distribution. The fact that wecLHS estimates the RSS and FKMS samples best indicates that the coverage of the feature space is a key element. Here, comparisons with other approaches such as FKMS with extragrades (deGuijter et al., 2010) and other promising approaches (e.g., Brus and Heuvelink, 2007; Kennard and Stone, 1969) should be conducted in future works. The fact that the fully and partially independent validations show comparable ranges between the sampling approaches ensures that there is no overfitting in the independent validation data.

Since, at least one of the advantages of wecLHS must be a key component a detailed study based on different wecLHS settings is recommended to reveal this component. This might help to design a new optimized sampling approach. It is also recommended (but expensive) to apply validation strategies, which are specifically designed for calibration sampling (e.g., Brus et al., 2011) to minimize all possible effects of sampling bias or autocorrelation.

The most remarkable fact in terms of validation is that the independent validation of the wecLHS approach provided better estimates compared to the resampling validation approaches, which was not the case for RSS and FKMS. This again shows that wecLHS provided a more evenly distribution in the state space. As a consequence the independent validation was better for wecLHS but

worse – as commonly expected – for the other sampling approaches. This is pronounced by the fact that the $wecLHS_{sub}$ approach outperformed the FKMS approach with 28 samples.

The fact that cross-validation can sometimes lead to much too optimistic assumptions (Brus et al., 2011) can be seen at several cases where the best resampling validations results correspond to the worst independent validation results. Bootstrapping turned out as the most conservative estimate. Yet it is too pessimistic for wecLHS, whereas the *.boot632* estimates is too optimistic in general. *lgocv* is the best estimate for the independent validation for wecLHS. Hence, there is not "the" best resampling validation measure.

The comparison between the regression approaches shows that in terms of the number of best performances RF must be recommended over MLR. However, the both approaches are comparable in most cases and there are no differences in the interpretation of the validation results of the different sampling schemes.

The analysis of the digital soil maps reveals that the LHS approaches and FKMS produce more similar results compared to RSS, which shows distinct differences. $WecLHS_{sub}$ results in almost the same map as the ones produced by wecLHS and FKMS indicating that 20 samples are sufficient for soil property calibrations in this study. In contrast to the average validation results the differences between the maps are higher between the regression approaches compared to the sampling approaches. Since, the MLR results show finer spatial variations, which do not seem to be relevant, RF must also be recommended from this point of view.

4. Conclusion

Three interacting methodological components of Digital Soil Mapping were jointly analyzed in this study on field scale sensor integration for soil property calibration: 3 sampling, 6 validation, and 2 regression approaches. Since the sampling approaches comprised different sample set sizes we sub-sampled two of them in different ways so that in total 6 sampling approaches were compared.

Beyond the general recommendation to apply the proposed wecLHS sampling approach in combination with Random Forests several remarkable results were found. Most interesting is that in average the difference in validation accuracy is largest between the sampling schemes followed by the differences between regression models. Even though, the regression models are based on totally different concepts the influence of the sampling schemes is higher. This indicates relatively linear relationships between the sensor and the soil data in this study.

The results show that based on a good sampling strategy only 20 samples are required for integrating multiple sensor data and to provide high prediction accuracies – in average over 70 % of variance explained for pH, SOC, and soil texture in two depth.

A final recommendation concerning the resampling validation approach cannot really be given. The *.632boot* and the *lgocv* approach (for wecLHS) provided best estimates for the independent validation. Yet, independent validation is recommended since resampling validation can result in either under- or over estimation of the independent accuracy as shown in this study.

Following studies should focus on wecLHS and determine which component makes it superior compared to the other approaches tested. In this respect other approaches such as FKMS or LHS with extragrades should be compared. Therefore independent validation is required as well as detailed analysis of feature importance (Behrens et al., 2010a,b).

Acknowledgements

We acknowledge funding from *iSOIL*, which is a Collaborative Project (Grant Agreement number 211386) co-funded by the Research DG of the European Commission within the RTD activities of the FP7 Thematic Priority Environment; iSOIL is one member of the SOIL TECHNOLOGY CLUSTER of Research Projects funded by the EC.

We thank Dick Brus for providing the RSS sample set and Claudia Dierke, Anne-Kathrin Nüsch, Helko Kotas, Andreas Schoßland and Marco Pohle for assistance with the field measurements.

References

Behrens, T., Mayr, T., Brus, D., Werban, U., Bellamy, P., Dietrich, P., 2009. iSOIL sampling protocol. Unpublished.

Behrens, T., Zhu, A. X., Schmidt, K., Scholten, T., 2010a. Multi-scale digital terrain analysis and feature selection in digital soil mapping. Geoderma, 155, 175-185.

Behrens, T., Schmidt, K., Zhu, A. X., Scholten, T., 2010b. The ConMap approach for terrain-based digital soil mapping. European Journal of Soil Science, 61, 133-143.

Bezdek, J.C., 1981. Pattern Recognition with Fuzzy Objective Function Algoritms. Plenum Press, New York.

Breiman, L., 1996. Bagging predictors. Machine Learning, 24, 123-140.

Breiman, L., 2001. Random forests. Machine Learning, 45, 5-32.

Breiman, L., Friedman, J., Olshen, R., Stone, C., 1984. Classification and Regression Trees. Wadsworth, Belmont, CA.

Brown, D.J., Bricklemyer, R.S., Miller, P.R., 2005. Validation requirements for diffuse reflectance soil characterization models with a case study of VNIR soil C prediction in Montana. Geoderma, 129, 251–267.

Brus, D.J., Heuvelink, G.B.M., 2007. Optimization of sample patterns for universal kriging of environmental variables. Geoderma 138, 86–95.

Brus, D.J., Kempen, B., Heuvelink, G.B.M., 2011. Sampling for validation of digital soil maps. European Journal of Soil Science 62, 394–407.

Callegary J. B., Ferre T. P. A., Groom R. W., 2007. Vertical spatial sensitivity and exploration depth of low-induction-number electromagnetic-induction instruments, Vadose Zone Journal, 6, 158-167.

de Gruijter, J.J., McBratney, A.B., Taylor, J., 2010. Sampling for high resolution soil mapping. In: Viscarra-Rossel, R.A., McBratney, A.B., Minasny, B., (Eds.), Proximal Soil Sensing. Springer,

Dunn, J.C., 1973. A Fuzzy Relative of the ISODATA Process and its Use in Detecting Compact, Well-Separated Clusters. J. Cybernetics, 3, 32-57.

Efron, B., 1983. Estimating the error rate of a prediction rule: improvement on cross-validation. J. Am. Stat. Assoc. 78, 316–331.

Efron, B., Tibshirani, R.J., 1993. An Introduction to the Bootstrap. NY Monographs on Statistics and Applied Probability, Chapman and Hall 57, London.

Efron, B., Tibshirani, R.J., 1995. Cross-validation and the bootstrap: Estimating the error rate of a prediction rule. Technical Report 176, Stanford University, Stanford.

ESRI, 2012. ArcGIS Desktop: Release 10. Redlands, CA: Environmental Systems Research Institute.

Good, P., 2006. Resampling Methods. 3rd ed. Birkhauser, Boston.

Kennard, R.W., Stone, L., 1969. Computer aided design of experiments. Technometrics 11, 137–148.

Kohavi, R., 1995. A study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. Proc. of the 14th Int. Joint Conf. on A.I. 2, Canada, 1137–1143.

Lesch, S.M., Strauss, D.J., Rhoades, J.D., 1995. Spatial prediction of soil salinity using electromagnetic induction techniques: 2. An efficient spatial sampling algorithm suitable for multiple linear regression model identification and estimation. Water Resour. Res. 31, 387-398.

Lesch, S.M., 2005. Sensor-directed response surface sampling designs for characterizing spatial variation in soil properties. Comput. Electron. Agric., 46:153-180.

Lesch, S.M., Rhoades, J.D., Corwin, D.L., 2000. The ESAP Version 2.01r user manual and tutorial guide. Research Report #146. George E. Brown Jr., Salinity Laboratory, Riverside, CA, 153pp.

Liaw, A., Wiener, M., 2009. Random Forest: Breiman and Cutler's Random Forests for Classification and Regression. http://cran.r-project.org/web/packages/randomForest/index.html.

Minasny, B., 2004. Latin Hypercube Sampling. http://www.mathworks.com/matlabcentral/fileexchange/4352-Latin-hypercube-sampling.

Minasny, B., McBratney, A.B., 2002. FuzME version 3.0, Australian Centre for Precision Agriculture, The University of Sydney, Australia.

Minasny, B., McBratney, A.B., 2006. A conditioned Latin hypercube method for sampling in the presence of ancillary information. Computer & Geosciences, 32, 1378-1388.

Minasny, B., McBratney, A.B., 2010. Conditioned Latin hypercube sampling for calibrating soil sensor data to soil properties. In: Viscarra-Rossel, R.A., McBratney, A.B., Minasny, B., (Eds.), Proximal Soil Sensing. Springer, 15–28.

Molinaro, A.M., Simon, R., Pfeiffer, R.M., 2005. Prediction error estimation: a comparison of resampling methods. Bioinformatics, 21, 3301-3307.

Pebesma, E.J., 2004. Multivariable geostatistics in S: the gstat package. Computers & Geosciences, 30, 683-691.

R Development Core Team, 2012. R: A language and environment for statistical computing.

R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/.

Rao, C.R., Toutenburg, H., Fieger, A., Heumann, C., Nittner, T., Scheid, S., 1999. Linear Models: Least Squares and Alternatives. Springer Series in Statistics, 427 pp.

Viscarra-Rossel, R.A., McBratney, A.B., Minasny, B., 2010. Proximal Soil Sensing. Springer, 486 pp.

Wilkinson, G.N., Rogers, C.E., 1973. Symbolic descriptions of factorial models for analysis of variance. Applied Statistics, 22, 392-399.

Manuscipt 3: Distance and similarity-search metrics for use with soil vis–NIR spectra

Geoderma, accepted in August 2012, doi: 10.1016/j.geoderma.2012.08.035

Leonardo Ramirez-Lopez[a,b], Thorsten Behrens[a], Karsten Schmidt[a], Raphael Viscarra Rossel[c], Jose A.M. Demattê[d], Thomas Scholten[a]

[a]Institute of Geography, Physical Geography and Soil Science, University of Tübingen,
Rümelinstraße 19–23, 72070, Tübingen, Germany.

[b]Georges Lemaître Centre for Earth and Climate Research, Earth and Life Institute,
Université Catholique de Louvain,
3 Place Louis Pasteur–1348, Louvain la Neuve, Belgium.

[c]CSIRO Land and Water, Bruce E. Butler Laboratory,
GPO Box 1666 Canberra ACT 2601, Australia

[d]Soil Science Department, Escola Superior de Agricultura "Luiz de Queiroz",
University of São Paulo
Av.: Pádua Dias, 11 CP 9. Piracicaba – SP 13418–900. Brazil.

Abstract

Many techniques used in visible and near infrared (vis–NIR) soil sensing are based on the measurement of the similarity or distance between samples. The question that frequently arises when two samples are very close in the vis–NIR space is whether they are also close (or similar) in terms of soil compositional characteristics. A good soil vis–NIR similarity metric must be also able to reflect the soil compositional similarity. In this respect, the main aims of this work were as follows: *i.* investigate the relationship between soil vis–NIR similarity and soil compositional similarity and *ii.* evaluate different distance metric algorithms for soil vis–NIR similarity search.

We evaluated the following distance metrics: Euclidean (ED), Mahalanobis (MD), spectral angle mapper (SAM), surface difference spectrum (SDS), spectral infor-

mation divergence (SID), principal component distance (PC-M), optimized PC distance (oPC-M), locally linear embedding distance (LLE-M) and σ-locally linear embedding (σLLE-M). The first five methods mentioned previously correspond to methods that operate in the spectral space while the remaining ones work by projecting the vis–NIR data onto a low dimensional space.

We used a global soil vis–NIR spectral library (GSSL) to test the different distance metric algorithms. The GSSL was divided into a reference set (Xr) and an unknown set (Xu). The distance algorithms were used to find in Xr the most spectrally similar samples of Xu. In order to evaluate the compositional similarity, the clay content and pH values of the Xu were compared to the clay content and pH values of the samples found in Xr by each algorithm.

The experimental results showed that the vis–NIR similarity measures that better reflect the soil compositional similarity are those corresponding to the oPC-M, LLE-M and σLLE-M methods. We also show that the SDS approach is a suitable method for computing distances in the spectral space. Finally, in this paper we discuss how these methods can also be used in proximal soil vis–NIR sensing applications.

Keywords: soil spectroscopy, locally linear embedding, vis–NIR similarity, compositional similarity, proximal soil sensing, global soil spectral library.

1. Introduction

Soil infrared spectroscopy and soil spectral libraries (SSL) have become powerful tools in soil science helping to analyze and store large amounts of soil information efficiently. Hence, the size of these libraries has been increasing and some initiatives to create regional, national and global SSL have emerged (e.g. Viscarra Rossel, 2009; Wetterlind and Stenberg, 2010; Terhoeven–Urselmans *et al.*, 2010). Several authors have suggested that regional and/or global SSL can be used for improving field scale vis–NIR models of soil attributes (e.g. Brown, 2007; Sankey *et al.*, 2008; Wetterling and Stenerg, 2010). This improvement would be reached by using the samples in the SSL that share similar soil characteristics to the

samples in the target field. Ge *et al.* (2011) indicates that poor performance of soil models would result from discrepancies between library samples and field samples not only in terms of vis–NIR characteristics but in terms of compositional characteristics. In this respect, the similarity metric used is of fundamental importance for selecting from the SSL the most suitable samples for within field soil modeling.

The measurement of the similarity between samples in soil vis–NIR datasets is a very complex task since the vis–NIR variation is highly multivariate and influenced by several soil compositional attributes. The vis–NIR features result from a high and variable mixture of several clay minerals and organic compounds present in the soil which are expressed as highly overlapped and non–specific vis–NIR absorptions. Currently, there is lack of research about quantitative approaches for soil vis–NIR similarity analysis. In addition, there are no strategies to evaluate the accuracy of these similarity measures. In most of the current approaches, distance metric algorithms are used for measuring the vis–NIR similarity between soil samples. The question that frequently arises when two samples are very close in the vis–NIR spectral space is whether they are also close (or similar) in terms of soil composition. In this respect, good soil vis–NIR similarity metric must also be able to reflect the soil compositional similarity. If the vis–NIR distance metric does not fulfill this requirement, then it is highly probable that the tasks in which the distance measures are involved will present poor performance.

The most commonly used distances in soil spectroscopy (and soil sensing in general) are the Euclidean distance (ED), the Mahalanobis distance (MD) and the spectral angle mapper (SAM) distance. The ED and MD distances can be measured directly on the spectral space or in a projected space such as the principal component (PC) space. The computation of the MD in the PC space has become the standard procedure for vis–NIR distance measurements. Recently, another spectral similarity measure called spectral information divergence (SID) (Chang, 2000) has been successfully used in remote sensing and imaging spectroscopy (e.g. van der Meer, 2006; He *et al.*, 2011).

Reliable distance estimations based on projection methods (such as PC analysis) depend on a good representation of the high–dimensional data in a low dimensional space. In this sense, there are still some questions that need to be answered regarding this method. For instance, the number of PC features to retain is based on the percentage of explained variance, so that the PC features with low significance are ignored. In some cases, using PC analysis, it is possible to retain the 99% of the total variance in the first two PC features but it does not necessarily mean that those features are enough for representing well the soil compositional variability.

In this context, the objectives of this work were as follows: *i.* investigate the relationship between soil compositional similarity and soil vis–NIR similarity *ii.* explore and develop suitable approaches for measuring similarities among samples in soil vis–NIR datasets, *iii.* provide a method to evaluate soil vis–NIR distance measurements and *iv.* introduce new methods for measuring the soil vis–NIR distances. For this latter objective we present the spectral difference surface (SDS) distance, optimized principal component Mahalanobis (oPC–M) distance, the locally linear embedding Mahalanobis (LLE–M) distance and the sigma locally linear embedding Mahalanobis (oLLE–M) distance.

2. Materials and methods

2.1 Algorithms

The algorithms and methodologies in this study focus on the analysis of the soil similarity between a reference data set of n reference samples $(\mathrm{Xr}, \mathrm{Yr}) = \{\mathrm{xr_i}, \mathrm{yr_i}\}_{i=1}^{n}$ and an unknown set of m samples $(\mathrm{Xu}, \mathrm{Yu}) = \{\mathrm{xu}_j, \mathrm{yu}_j\}_{j=1}^{m}$ where the Yu values are "unknown". Here Xr and Xu represent the spectra of each set and Yr and Yu represent a given soil attribute. In other words, distance metric algorithms are used to find in the reference set the subset $(\mathrm{Xru}, \mathrm{Yru}) = \{\mathrm{xru}_j, \mathrm{yru}_j\}_{j=1}^{m}$ of m samples which are the most similar ones to the unknown samples.

2.1.1 Distance metrics for similarity search in the vis–NIR space

2.1.1.1 Euclidean and Mahalanobis distances

Both, the Euclidean and the Mahalanobis distances (ED and MD respectively) are widely used in many research fields for measuring similarities among samples. In the computation of the ED, each feature has equal significance. Hence, correlated variables, which may represent irrelevant features, can have a disproportional influence on the analysis (Brereton, 2003).

For computing the distance (d) between samples in Xr and samples in Xu the following equation is used:

$$d(\mathrm{xr}_i, \mathrm{xu}_j) = \sqrt{(\mathrm{xr}_i - \mathrm{xu}_j) M^{-1} (\mathrm{xr}_i - \mathrm{xu}_j)^T} \quad (1)$$

where M is the identity matrix in the case of the ED while for MD M is the covariance matrix of Xr ∪ Xu. In contrast to the ED, the MD accounts for the correlation between features by using the covariance matrix. Another characteristic of the MD is that it is scale–invariant which means that the result will not change if all the dimensions are scaled equally (Laskar *et al.*, 2011).

The MD can also be viewed as the ED of the feature space after applying a linear transformation. Such linear transformation is done by using a factorization of the inverse covariance matrix as $M^{-1} = W^{\mathrm{T}}W$, where W is merely the square root of M^{-1}. Therefore, the MD between xr_i and xu_j is equivalent to the ED between $W\mathrm{xr}_i$ and $W\mathrm{xu}_j$ which is calculated as follows:

$$d(\mathrm{xr}_i, \mathrm{xu}_j) = \sqrt{(W\mathrm{xr}_i - W\mathrm{xu}_j)(W\mathrm{xr}_i - W\mathrm{xu}_j)^T} \quad (2)$$

The computation of the MD in the original vis–NIR space can involve some problems. For instance, M can result in a singular matrix which cannot be inverted. The reason for this is that the spectral features are usually highly correlated (De Maesschalck *et al.*, 2000). In addition, for MD computations the number of samples in the dataset must be larger than the number of spectral features.

2.1.1.2 Spectral Angle Mapper (SAM)

The SAM or dot–product cosine was introduced by Yuhas *et al.* (1992). It has been extensively used in remote sensing for unsupervised classification of multi–spectral and hyper–spectral images. In soil spectroscopy SAM has been used for soil similarity analysis (e.g. Farifteh *et al.*, 2007; Lugassi *et al.*, 2010). It is also used as scale invariant similarity measure. The similarity is measured by calcu-Lating the angle between samples in the spectral space. The SAM is calculated as:

$$\mathrm{SAM}(\mathrm{xr}_i, \mathrm{xu}_j) = \cos^{-1} \frac{\sum_{u=1}^{nb} \mathrm{xr}_{i,u}\, \mathrm{xu}_{j,u}}{\left(\sum_{u=1}^{nb} (\mathrm{xr}_{i,u})^2\right)^{1/2} \left(\sum_{u=1}^{nb} (\mathrm{xu}_{j,u})^2\right)^{1/2}} \quad (3)$$

where nb is the number of spectral bands or features corresponding to both xr_i and xu_j.

The SAM is insensitive to illumination and albedo magnitude effects because the angle between two spectrums is invariant with respect to the lengths of them (Park *et al.*, 2007). It is also easy to implement and its computational cost is low.

2.1.1.3 Spectral Information Divergence (SID)

The SID was introduced by Chang (2000), it uses the Kullback–Leibler divergence (*KL*) or relative entropy (Kullback and Leibler, 1951) to account for the vis–NIR information provided by each spectrum. The *KL* measures the difference between two probability distributions. However, it cannot be considered as one distance metric, since $KL\left(\mathrm{xr}_i \| \mathrm{xu}_j\right)$ is not equal to $KL(\mathrm{xu}_j \| \mathrm{xr}_i)$. Nevertheless, a distance metric can be obtained by using the SID approach. The SID similarity is given by the sum of the estimated divergence between xr_i and xu_j and the estimated divergence between xu_j and xr_i. The SID is computed as follows:

$$\mathrm{SID}\left(\mathrm{xr}_i, \mathrm{xu}_j\right) = KL\left(\mathrm{xr}_i \| \mathrm{xu}_j\right) + KL(\mathrm{xu}_j \| \mathrm{xr}_i) \quad (4)$$

$$= \sum_{l=1}^{b} p_l \log\left(\frac{p_l}{q_l}\right) + \sum_{l=1}^{b} q_l \log\left(\frac{q_l}{p_l}\right) \quad (5)$$

where b represents the number of variables or spectral features, p and q are the probability vectors of xr_i and xu_j respectively which are calculated as:

$$p = \frac{\mathrm{xr}_i}{\sum_{l=1}^{b} \mathrm{xr}_{il}} \quad (6)$$

$$q = \frac{\mathrm{xu}_j}{\sum_{k=1}^{b} \mathrm{xu}_{jl}} \quad (7)$$

2.1.1.4 Surface Difference Spectrum (SDS)

The SDS is a new distance metric approach. This method involves a multi-scale analysis of the differences (e) between two soil spectra xr_i and xu_j. A derivative function (a) is applied on e as a function of the frequency (or wavelength) delay returning a 3D spectrum of differences. The SDS is calculated in the spectral space as:

$$\mathrm{SDS}\ (\sigma) \begin{cases} e(\mathrm{xr}_i, \mathrm{xu}_j) = (\mathrm{xr}_i, \mathrm{xu}_j) & \text{if } \sigma = 0 \\ a(e,e) = \frac{1}{\sigma}\ (e_i - e_{i+\sigma}) & \text{otherwise} \end{cases} \quad (8)$$

from this spectral surface the distance (d) between xr_i and xu_j is formulated as:

$$d(\mathrm{xr}_i, \mathrm{xu}_j) = \frac{1}{2}[g + g\ w(\mathrm{xr}_i, \mathrm{xu}_j)] \quad (9)$$

where

$$g = \sqrt{\sum_{\sigma=0}^{s} \sum_{i=1}^{nb-\sigma} \mathrm{SDS}^2} \quad (10)$$

and

- σ is the frequency delay which can vary from 0 to nb - 1,
- nb is the number of bands or features of the spectra.
- nb–σ represents the number of features of generated at each frequency delay computation.
- s is the optimum value of σ.

- $w(xr_i, xu_j)$ is given by the inverse of the squared correlation coefficient between xr_i and xu_j. This weight is obtained from a robust correlation analysis (Chu *et al.*, 2011). This correlation coefficient is calculated on the basis of a window that moves over the whole spectral region. The final correlation coefficient is the average of the correlations found with this moving window. In this case we use a window width of 11 bands (110 nm) as suggested in Chu *et al.* (2011). This correlation coefficient is used to further improve the performance of the SDS method.

After performing the computation of d, the values of e and a must be normalized.

In SDS the parameter σ needs to be optimized. Each spectrum generated at each frequency delay iteration can be interpreted as the derivative energy spectrum of the spectral difference between xr_i and xu_j.

Figure 1a provides an example of the *SDS* between two very similar vis–NIR spectra. Because the shapes of both spectral samples are similar the e between them is almost constant, therefore its corresponding surface (Figure 1b) is nearly flat. Figure 1c shows two slightly different spectra, their main differences are found in the absorption bands around 2200 nm which indicate the presence of kaolinite (Demattê *et al.*, 2004) and the energy absorption to the Al–OH in general (Viscarra Rossel and Behrens, 2010). These differences are reflected in the Euclidean distance spectrum and in the derived surface (Figure 1d) where some irregular patterns can be found. Figures 1e and 1f show the SDS analysis for two spectrums with very different vis–NIR patterns where e and its correspondent surface are very irregular.

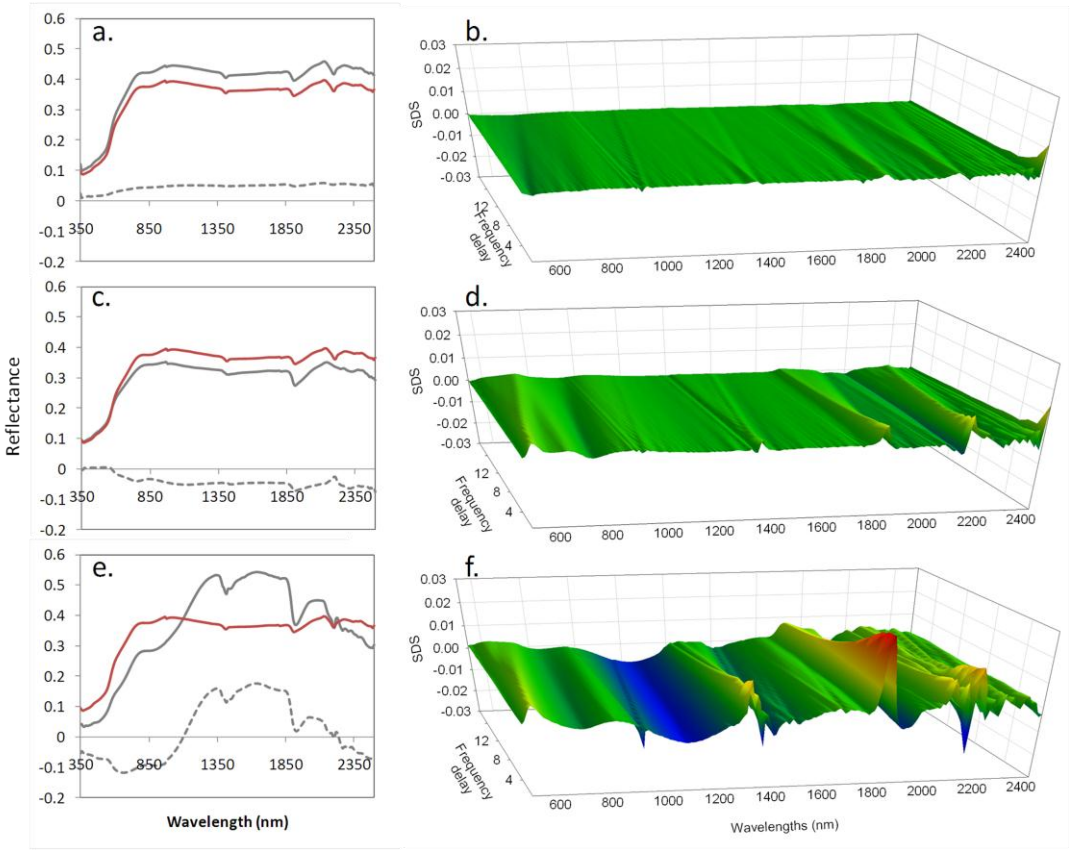

Figure 1. Example of the SDS computation. Dash lines in a, c and e represent the Euclidean distance spectrums (*E*) or *SDS* at frequency delay of *0* (*σ=0*). The 3–D representations of the SDSs (with *σ= 1,.., 15*) of a, c, and e are showed in b, d and f respectively.

In this approach additional information about the difference between xr_i and xu_j is generated at each frequency delay iteration. Note that when $e(xr_i, xu_j)$ is constant, then its $a(e,e)$ spectrum and w will be 0.

The optimization framework to identify the adequate number of frequency delays is described as follows: first, the SDS distance is computed for the frequency delays corresponding to a sequence from 0 to a predefined threshold (t). In this sense a set of t SDS distance matrices is obtained. Then, by using these matrices the spectral nearest neighbor (NN) of each xr_i is selected from Xr. Each xr_i sample and its NN are compared in terms of soil compositional attributes (e.g. clay content and pH) i.e. a comparison between yr_i and the corresponding soil attribute of its spectral NN ($ÿr_i$) is carried out. For compositional comparison the root mean square of compositional differences ($RMSD_c$) is used. The $RMSD_c$ is calculated as:

$$\mathrm{RMSD_c} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\mathrm{yr}_i - \ddot{\mathrm{y}}\mathrm{r}_i)^2}, \quad (11)$$

Note that $\ddot{\mathrm{y}}\mathrm{r}_i \in \mathrm{Yr}$. Finally, the adequate number of frequency delays (s) to be used is the one corresponding to the SDS distance matrix that minimizes the $\mathrm{RMSD_c}$.

2.1.2 Distance metrics for similarity search in projected spaces

2.1.2.1 The principal components–Mahalanobis (PC–M) distance

The principal component (PC) analysis is a projection method which is the standard approach used to reduce the dimensionality of soil vis–NIR spectra. An eigen-decomposition of covariance matrix or a singular value decomposition (SVD) of the data matrix can be used for the projection. Other PC methods such the non-linear iterative partial least squares (NIPALS, Wold, 1966) can also be used for PC projections. Here we used the SVD for the PC analysis. This method frequently requires preprocessing of the spectral data (usually centering and scaling) prior the PC projection, nevertheless, this depends on the data to be used and the purpose of the analysis.

For computing the distances between reference and unknown samples, first a PC analysis of Xr ∪ Xu is carried out, then a set of PC features are retained based on the cumulative amount of variance explained and the rest of them are ignored assuming they do not contain useful information. Finally, the Mahalanobis distance is computed on the retained PC features.

2.1.2.2 The optimized PC Mahalanobis (oPC–M) distance

Here we propose the oPC–M distance method which only differs from the standard PC distance in the way in which the number of PC features to retain is calculated. The goal in the oPC–M approach is to identify the optimal number of PC features (or level of compression) representing the soil compositional variability. The rationale behind this approach is based on the assumption that soil vis–NIR

variability should also reflect the soil compositional variability, at least in terms of those attributes that have strong influence on the vis–NIR spectra.

In order to find the optimal number of PC features to retain we propose a very simple framework which is similar to the one used in the SDS method and it is based on the minimization of the $RMSD_c$. A PC analysis of Xr ∪ Xu is carried out, then the PC features are retained one at a time (according to its explained variance), in each iteration a Mahalanobis distance matrix is computed. By using each MD matrix, the nearest neighbor (NN) of each xr_i is also selected from Xr as a function of the number of PC retained. The Xr samples and their spectral NNs are compared in terms of soil compositional attributes. The optimal number of PCs is the one corresponding to the distance matrix that minimizes $RMSD_c$.

2.1.2.3 Locally Linear Embedding–Mahalanobis (LLE–M) distance

The Locally Linear Embedding (LLE) algorithm was introduced by Roweis and Saul (2000) and extended in Saul and Roweis (2003). This approach has not been applied in soil vis–NIR spectroscopy so far. However it has been successfully used for dimensionality reduction of hyperspectral remote sensing images (e.g. Chen and Qian, 2007; Ma *et al.*, 2010)

The LLE is basically a non–linear dimensionality reduction method. It can be viewed as an unsupervised metric learning algorithm which learns the global manifold structure from local neighborhoods. In other words, LLE is able to reconstruct complex structures from locally Euclidean structures. The locally linear embedding of X (where X = Xr ∪ Xu) is carried out in three main steps (Figure 2):

1. Select neighbors: In this step the Euclidean distance is used to find the *k*–nearest neighbors of each data point x_o. These samples are used as description of local patches of the manifold.
2. Compute a weight matrix: Here a weight matrix (*W*) is computed in order to optimally reconstruct x_o from its neighbors. A weight is assigned to each neighbor of x_o. The neighbors must be able to represent or reconstruct x_o. The more similar the neighborhood samples are to x_o the more accurate the

reconstruction of each x_o will be. High weight values are assigned to those neighbors which contribute more to the reconstruction of x_o, and low weight values to those neighbors that have a small contribution. A weight of 0 is assigned for the cases in the data set which do not belong to the set of neighbors of each x_o.

The reconstruction weights of each x_o are calculated as follows: *i*. create a *G* matrix by subtracting x_o from the matrix of its neighbors, *ii*. compute the local covariance (*C*) as $C=G^TG$ and *iii*. solve the linear system $C\ w = 1$, where w are the weights of each neighbor.

3. Compute the low–dimensional coordinates. Finally x_o is embedded onto a low dimensional space (*V*) by using the reconstruction weights. This step is performed by choosing the low dimensional coordinates of each v_o which minimizes the following cost function:

$$\phi(V) = \sum_{i=1}^{N}\left(v_o - \sum_{k=1}^{N} w_{ok}v_k\right)^2 \qquad (12)$$

The goal here is to find low dimensional outputs v_o that are reconstructed by the same weights w_{ok} as the high dimensional inputs x_o (Saul and Roweis, 2003).

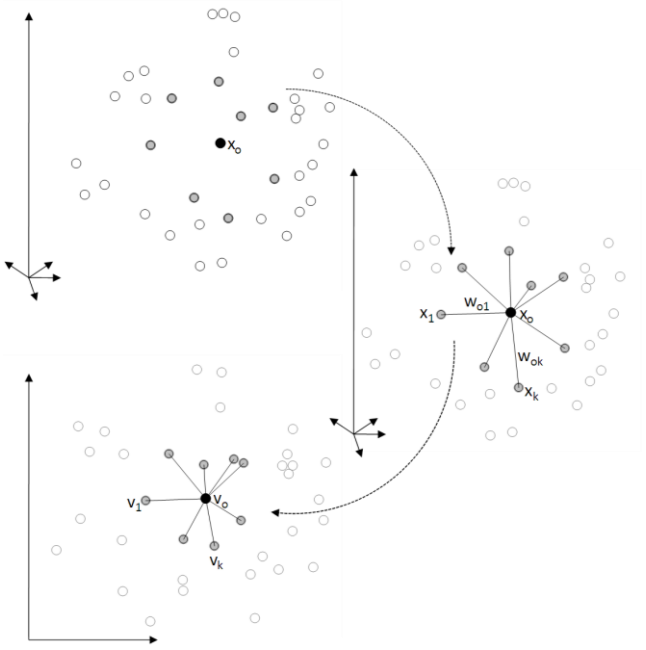

Figure 2. Summary of the LLE algorithm, mapping high dimensional inputs x_o to low dimensional outputs v_o (adapted from Roweis and Saul, 2000).

The maximum dimensions LLE projection is determined by the number of nearest neighbors (k). The maximum number of dimensions (d) that LLE can be expected to recover is strictly less than the number of neighbors used (Roweis and Saul, 2000). Here we fixed the number of dimensions to $d = k$-1. This means that we used all the dimensions that can be recovered with the k nearest neighbors retained, in order to capture all the information extracted by the LLE algorithm.

The only free parameter which needs to be optimized is k.

The LLE–M distance corresponds to the Mahalanobis distance measured in the projected LLE space of X. Prior the computation of the LLE–M distances between samples of Xr and Xu, is necessary to identify the optimal number of neighbors (k) to use in the LLE projection. The procedure for optimizing the number of LLE dimensions is basically the same as in the oPC–M method for optimizing the number of PCs, i.e. it is based on the minimization of the $RMSD_c$. We tested from 3 to 55 neighbors in steps of 1.

2.1.2.4 The surface difference spectrum for locally linear embedding–Mahalanobis distance (σLLE–M)

Taking into account that the neighborhood selection, which is the only non–linear step of the LLE, plays a key role in the performance of LLE algorithm (Chen and Liu, 2011), we propose an extension of the LLE–M called σLLE–M. The only difference between LLE–M and σLLE–M lies on the neighbor selection (step 1). While LLE–M uses the Euclidean distance for neighbor search, σLLE–M uses the distance metric function derived from SDS. The SDS algorithm is used to find appropriate neighbors better to represent and reconstruct the query point. Some works have demonstrated that by improving the neighbor search by using different distance metrics, the performance of locally linear embedding algorithm can be improved (e.g. Varini *et al.*, 2006; Pan *et al.*, 2009).

As in the standard LLE–M , the calculation of the optimal number of nearest neighbors is carried out in the same way as in the oPC–M method.

2.2 The vis–NIR soil spectral library

We used a global soil spectral library (GSSL) developed by the World Agroforestry Centre (ICRAF) and the ISRIC - World Soil Information (2010). The GSSL comprises 4438 soil samples of which 3643 have both, soil chemical attribute and texture information. These 3643 samples originate from 55 countries spanning America (27%), Africa (24%), Asia (23%), Europe (23%) and Oceania (3%). These soils have large texture variability and are represented by all 32 soil groups in the world reference base for soil resources (IUSS Working Group WRB, 2006)

Briefly, the reflectance spectra were recorded using a FieldSpec® FR vis–NIR spectrometer (Analytical Spectral Devices, Bolder, Colorado, USA) which collects the spectral measurements in a range of 350 to 2500 nm and is characterized by a Full Width Half Maximum of 3 nm for the 350-1000 nm region and 10 nm for the 1000-2500 nm region. The GSSL was resampled to 10 nm spectral resolution, with a total of 216 spectral features. Further details of the GSSL such as optical setup and sample preparation for spectral measurements are given in Shepherd *et al*. (2003). Chemical and texture analysis of soil samples were performed by ISRIC according to the procedures for soil analysis given in Van Reeuwijk (2002).

2.3 Transformation of the vis–NIR reflectance spectra and pre–processing

The spectra were transformed to absorbance units (log 1/Reflectance) and then the first derivative was computed. By using this procedure the centers of the peaks are converted to zero and it is a good way of accurately pinpointing the position of a broad peak. In addition the first derivative can remove the effect of baseline offsets. We applied this pre–processing technique since it reveals relevant spectral variability and it is useful for finding differences between samples (Gemperline and Kalivas, 2006).

2.4 Soil compositional similarity search based on the vis–NIR spectra

The algorithms described in section 2.1 were implemented in R (R Development Core Team, 2011).

For each sample in the unknown set xu_j, we searched for its most spectrally similar (closest) sample (xru_j) in Xr. Assuming that soils with similar vis–NIR spectra share similar compositional characteristics we evaluated the performance of the distance metrics presented in this study by their ability to identify samples with similar clay content and pH values. We used these two soil attributes because they have very different effects on the soil vis–NIR spectra. Clay content has a strong effect on the vis–NIR reflectance intensity (Demattê *et al*., 2004) affecting the whole spectrum. On the other hand, pH has localized and weak effects on the spectra. Soil pH does not have a direct spectral response (Stenberg *et al*., 2010). This is probably due to the fact that the hydrogen ions (measured in pH) are held by the exchange complex of soil. So that spectral responses associated with pH should occur at some specific parts of the wavelengths of clay minerals and organic compounds responsible for cation exchange capacity.

The distance metric algorithms were tested on the GSSL. We randomly selected 700 samples as unknown set (Xu,Yu) and the remaining samples (2943) were used as reference set (Xr,Yr).

The root mean square of differences (RMSD) between Yu and Yru was used as parameter to evaluate the performance of the algorithms. In this case the RMSD was calculated as follows:

$$\mathrm{RMSD} = \sqrt{\frac{1}{n}\sum_{i=1}^{m}(\mathrm{yu}_j - \mathrm{yru}_j)^2}, \tag{13}$$

where yu_j is the soil attribute value of each sample in the unknown set and yru_j is the soil attribute value of its corresponding most similar sample found in Xr. We also compared the results using the coefficient of determination (R^2).

The methodological framework of the compositional similarity search is summarized in the Figure 3.

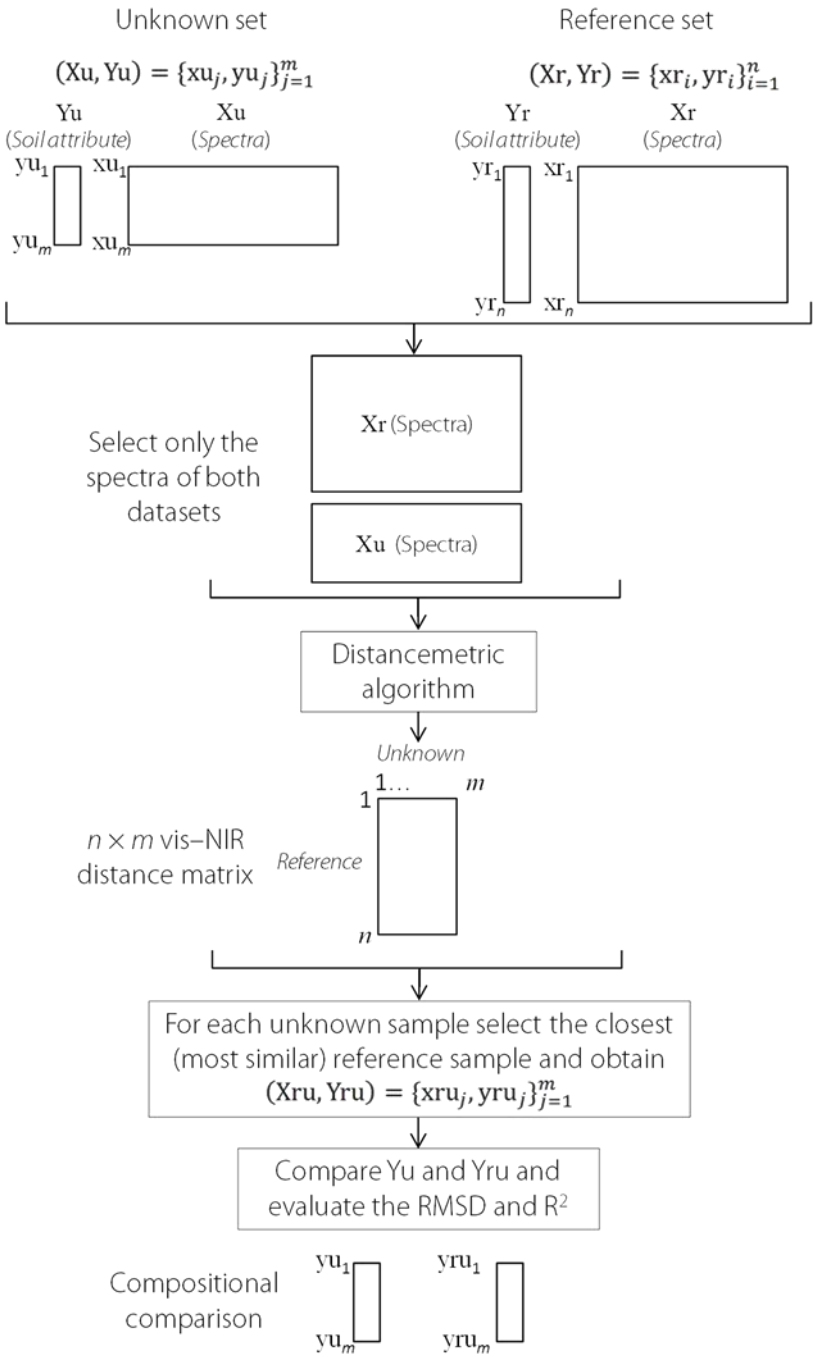

Figure 3. Methodological framework for soil compositional similarity search using soil vis–NIR distances.

3. Results

3.1 Soil attributes and vis–NIR spectral characteristics

As expected, a large variation of clay content and pH was observed among the samples in the library (Figure 4). Clay content ranged between 0 and 96.8% with

a mean of 33.1 %. Soil acidity ranged between 3.0 and 10.5 pH units with a mean of 6.1. No significant correlation (r=0.01, p> 0.05) between clay content and pH was observed. For both attributes the mass of the distribution is concentrated below the median (30.5% for clay content and 5.9 for pH).

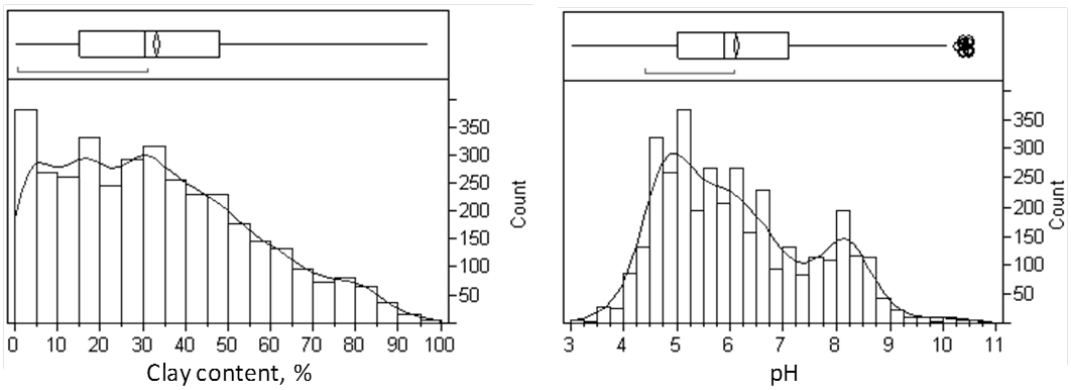

Figure 4. Histograms and outlier box-plots of clay content and pH values.

In order to illustrate the variability of the GSSL we selected 10 soil spectra (Figure 5a) using a conditioned Latin hypercube sampling (Minasny and McBratney, 2006) on a projected principal component space. To facilitate the analysis of absorption features we also used the continuum removed spectra (Clark and Roush, 1984) of these selected samples (Figure 5b). The large variation of soil attributes is also reflected in the soil spectral variation. The continuum removed spectra of soil samples indicates a large soil mineralogical variation in the GSSL as consequence of the diversity of soil formation environments where samples were collected. In the NIR region, the main spectral variations among samples were observed at wavelengths from 2160 nm to 2230 nm which seem due to the energy absorption of soil minerals such as kaolinite and smectite (Demattê and Garcia, 1999; Demattê *et al.*, 2004; Viscarra Rossel and Behrens, 2010). In a lesser extent, the spectra showed considerable variability at 1395 nm which is related to the presence of kaolinite (Viscarra Rossel and Behrens, 2010). Bands associated with the hygroscopic moisture content corresponding to 1380 nm and 1455 nm (Ben–Dor *et al.*, 2008) presented also high variation among samples. In the vis region, large spectral variations were observed at wavelengths of 435, 550 and 850 nm which are associated to the content of iron oxides (Demattê and Garcia, 1999). Some samples with weak absorption features show also typical spectral characteristics of soils with high organic matter content (Stoner and Baumgard-

ner, 1981; Ben–Dor *et al.*, 1999) and/or low levels of crystalline iron and amorphous iron (Ben–Dor *et al.*, 2008).

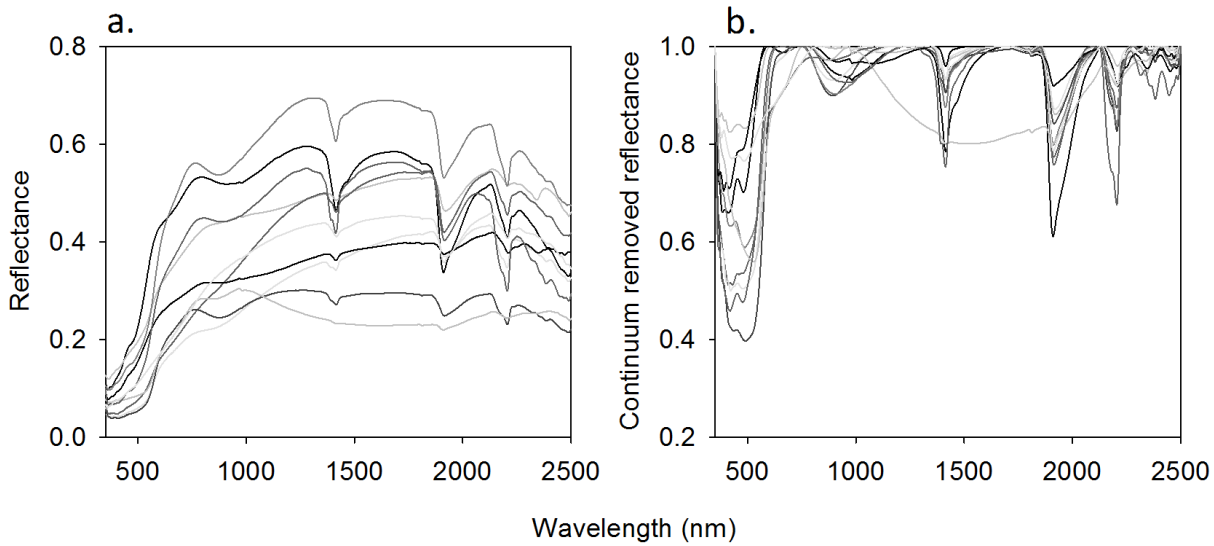

Figure 5. Reflectance (a) and continuum removed spectrums (b) of ten soil samples selected to illustrate the vis–NIR variation in the GSSL.

3.2 Optimizations

4.2.1 Frequency delays for surface difference spectrum (SDS)

Using the SDS method we found that the distance metric can be improved gradually by increasing the number of wavelengths delays (σ) involved in the distance computation. In the case of pH this gradual improvement is followed by the stabilization of the $RMSD_c$ in which no significant improvement is observed. For clay content, the sensibility of the $RMSD_c$ to frequency delay variations is low. Despite this, there is a clear tendency of the reduction of the $RMSD_c$ in the first 11 frequency delays, and after this point the $RMSD_c$ increases (Figure 6).

Figure 6 shows that around 9 and 5 frequency delays were necessary to reduce the $RMSD_c$ for clay content and pH respectively. This also suggests that new information about the spectral similarity between samples emerges when the neighborhood wavelengths are taken into account. For computing the final SDS distance matrix we used 5 frequency delays (σ=5) taking into account that the $RMSD_c$ for pH comparisons is more affected by σ than the $RMSD_c$ for clay content comparisons.

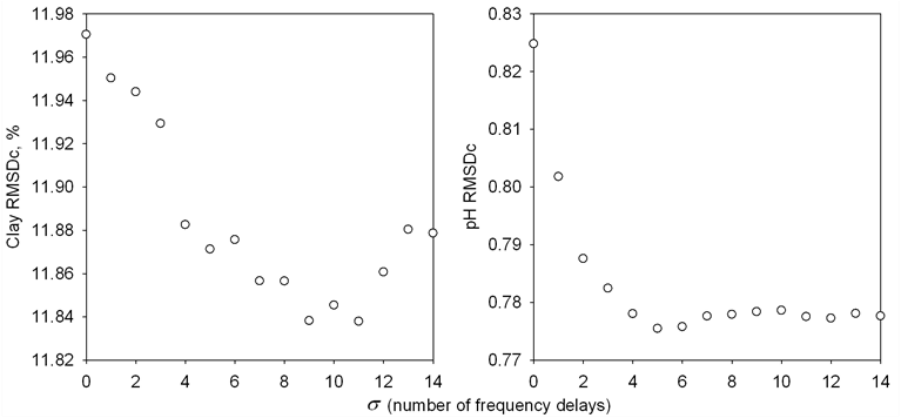

Figure 6. Number of frequency delays (σ) against the root mean square of compositional differences in the reference set ($RMSD_c$) of the SDS distance.

3.2.2 Number of PC features for the standard PC–M method

In the standard PC–M distance, the selection of the number of PC features to retain does not take into account the compositional similarity between samples, instead it is based only in the proportion of the explained variance of each component. Here, the first 9 PC features accounted for 43.9, 20.9, 12.3, 10.1, 3.9, 1.8, and 0.9 % respectively (Figure 7). This means that by using these features the 95.9% of the soil vis–NIR variation can be retained. For this reason, only the first 9 PC features were used for computing the PC–M distances. The rest of the PC features were excluded from the analysis since their individual explained variances were lower than 0.5%.

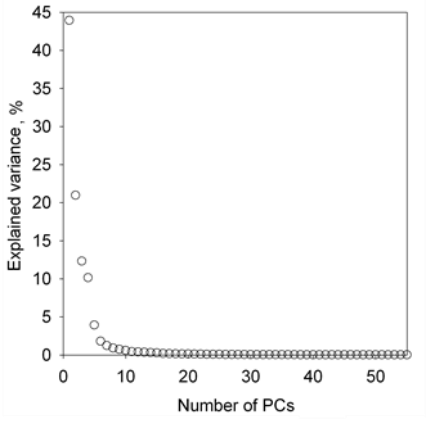

Figure 7. Percentage of explained variance by each PC feature.

3.2.3 PC features for oPC–M

Under the conventional assumption 9 PCs would be enough to represent the soil spectral variation. However, by using more than 9 PCs better results were ob-

tained in terms of soil compositional similarity search. Figure 8 shows that the $RMSD_c$ decreased steeply as function of the number of PC features used to compute the distances between samples, and then a gradual increasing of the $RMSD_c$ was observed.

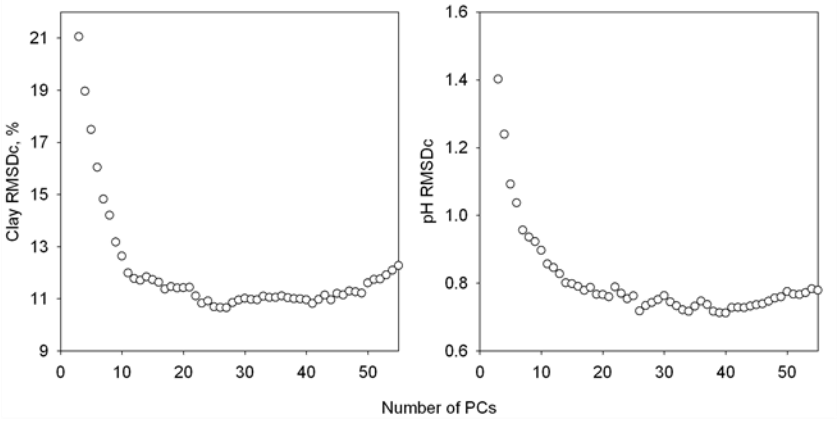

Figure 8. Number of PC features against the $RMSD_c$ of similarity search of clay content and pH.

Our results show the necessity of parameter optimization in PC analysis. By using the conventional selection of PC features, important soil compositional information contained in the vis–NIR spectra can be lost. In both cases 26 PC features returned the minimums $RMSD_c$ for clay content and pH, so that for the computation of the final oPC–M distance matrix we use these 26 PCs.

3.2.4 Euclidean nearest neighbors for LLE–M distances

The LLE–M method applies the *k* nearest neighbors search based on Euclidean distances to reconstruct each sample (x_o) form local patches. We found that by increasing the number of neighbors used in the reconstruction of each x_o the $RMSD_c$ can be reduced steeply until reaching a stability point (Figure 9). The lowest $RMSD_c$s for clay content and pH comparisons were returned by using 48 and 49 nearest neighbors respectively. These number of neighbors returned very similar results in both cases. For the final LLE–M distance matrix computation we used 49 nearest neighbors.

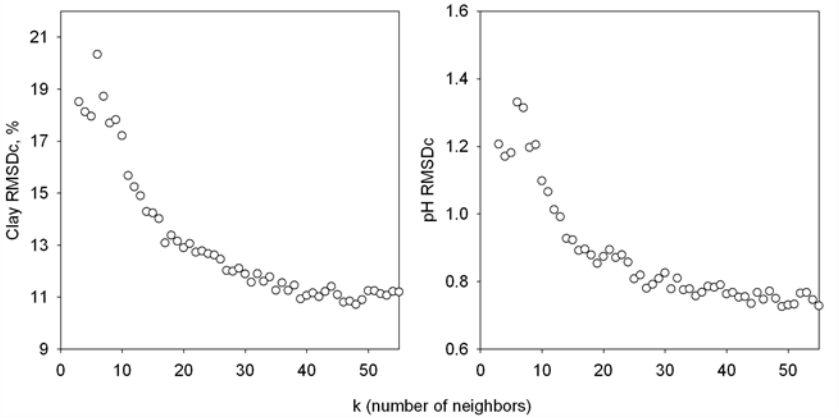

Figure 9. Number of nearest neighbors (k) used in the LLE–M method against the $RMSD_c$ of similarity search of clay content and pH. Note that the number of locally linear embedding dimensions (d) is $d = k–1$.

3.2.5 SDS nearest neighbors for σLLE–M

By using the σLLE–M method we observed that the $RMSD_c$ can be reduced steeply by increasing the number of nearest neighbors selected by using the SDS distance matrix. The lowest $RMSD_c$s for clay and pH comparisons were obtained by using 50 and 52 nearest neighbors in the LLE–M algorithm. For the final σLLE–M distance matrix computation we used 52 nearest neighbors.

In this optimization step, similar results between LLE–M and σLLE–M in terms of the $RMSD_c$s obtained for clay content and pH were found (Figure 9; Figure 10). For example, by using the selected distance matrices, the LLE–M and the σLLE–M methods returned the same $RMSD_c$s which was of 10.25% while for pH the $RMSD_c$s were 0.73 and 0.72 for the LLE–M and σLLE–M methods respectively.

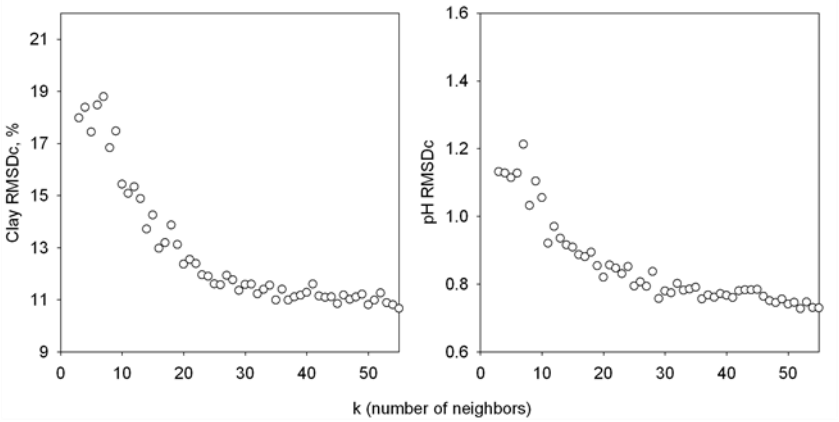

Figure 10. Number of SDS–nearest neighbors (k) used in the σLLE–M method against the $RMSD_c$ of similarity search of clay content and pH. As in LLE–M the number of locally linear embedding dimensions (d) is $d = k–1$.

3.3 Comparison of the different distance metric algorithms for compositional similarity search

Once the parameters of the SDS, oPC–M, LLE–M and σLLE–M algorithms were optimized we searched for the most similar samples to the unknown set (Xu) in the reference set (Xr).

In the spectral space the MD returned the poorest results (Table 1). The samples found in Xr by using the MD showed a low degree of compositional similarity to the samples in unknown set in terms of clay content and pH. However the use of the Mahalanobis distance in the PC spaces and in the LLE spaces (PC–M, oPC–M, LLE–M and σLLE–M methods) returned good results (Table 1). The best results in the spectral space were obtained by the SDS method with RMSD values of 12.34% and 0.74 for clay content and pH, respectively. The SDS and the SAM methods returned better results than those obtained with the standard PC–M method.

Table 1. Results of the comparisons of clay content and pH between the samples in the unknown set and their correspondent most similar samples found in the reference set for each vis–NIR similarity search method.

Method	Parameter value*	Clay content		pH	
		R^2	RMSD	R^2	RMSD
Similarity search in the spectral space					
ED	–	0.70	12.67	0.67	0.85
MD	–	0.30	21.37	0.32	1.30
SAM	–	0.72	12.42	0.70	0.80
SID	–	0.68	13.40	0.71	0.79
SDS	5	0.72	12.34	0.75	0.74
Similarity search in the projected spaces					
PC–M	9	0.72	12.59	0.68	0.83
oPC–M	26	0.76	11.39	0.80	0.64
LLE–M	49	0.80	10.42	0.75	0.74
σLLE–M	52	0.80	10.49	0.80	0.65

*Parameters for each method are: σ for SDS, PC features for PC–M and oPC–M, and number of nearest neighbors (*k*) for LLE–M and σLLE–M.

In the projected PC space the standard PC–M distance method can be improved by using our proposed parameter optimization (Table 1). For both attributes, the oPC–M returned lower RMSDs in comparison to the standard PC method. This confirms that important soil information useful for distance measurements and compositional similarity search is also contained in PC features with explained variance lower than 0.5%.

For the locally linear embedding approaches (LLE–M and oLLE–M), we found that they are reliable methods for similarity search. The samples found in Xr by the LLE–M and the oLLE–M methods showed similar degrees of compositional similarity to the samples of Xu set in terms of clay content with RMSDs of 10.42% and 10.49% respectively. Although this, for pH the oLLE–M method (RMSD=0.65) outperformed the LLE–M method (RMSD=0.74). A probable explanation for this relies on the distance used for neighbor selection step carried out in both methods. While the LLE–M uses the ED the oLLE–M uses the SDS distance. In this sense, the SDS distance seem to be better than ED for neighbor selection since it returned better similarity results for pH and in a lower extend for clay content.

Comparing the oLLE–M and the oPC–M methods the samples found by these methods returned similar results for pH being the RMSD of the oPC–M method slightly lower than the one presented by the oLLE–M method (Table 1). However the samples found in Xr by the oLLE–M method retuned a lower RMSD for clay content than those found by the oPC–M method.

4. Discussion

Concerning the poor performance of the MD, we can deduce that it was caused by the use of the covariance matrix computed in the original spectral space. As we mentioned in section 2.1.1.1 the MD is equivalent to the ED after a linear transformation of the data by using the square root of the inverse covariance matrix. This means that after that transformation the vis–NIR distances or similarities no longer represent the soil compositional similarity. For this reasons the classi-

cal estimates of the covariance matrix in the original high dimensional space of vis–NIR data sets should be avoided.

We showed that the conventional methods (ED, MD, SAM and SID) used in remote sensing do not have a satisfactory performance when used in soil vis–NIR spectroscopy. One probable reason for that relies on the fact that in high dimensional spaces the notion of similarity becomes less accurate (Abou–Moustafa and Ferrie, 2008).

The SDS method mitigates very well the complexity problem by extracting additional features from the Euclidean distance spectra computed in the original vis–NIR space. Unlike the ED, MD, and SAM, the SDS method takes into account the sequence of the spectra. This characteristic seems to be very useful for soil spectral similarity analysis. Nevertheless, the SDS method is computationally expensive.

One important characteristic of the SDS, oPC–M, LLE–M and oLLE–M methods is that they take into account the soil compositional similarity for the optimization of their parameters. In this sense, one can choose several soil attributes to check if the vis–NIR similarity reflects the compositional similarity properly. Our results showed that in each method the optimal parameters to maximize the similarity of clay content and pH between the unknown samples and the samples found in the reference set do not differ very much. In order to solve this problem, we suggest using only those soil attributes that have strong influence on the soil vis–NIR spectra.

For the projection approaches, we demonstrate that both, the LLE–M and the oLLE–M can improve the distance metrics that they use for the projection. In the case of oLLE–M, it returns similar results of clay content comparisons and better results for pH than LLE–M. This demonstrates the fact that reliable distance metrics in the original predictor space (such as the SDS distance) are very important for distance metric learning approaches. On the other hand, we observed that LLE–M returns similar results to the oPC–M, however the LLE–M has a higher computational cost. In this sense, the computational cost of the oLLE–M

approach is much higher than the LLE–M since it involves the computation of SDS distances.

5. Proximal soil vis–NIR sensing applications

Similarity search methods would be very useful for integrating soil spectral libraries into proximal soil sensing for field predictions of soil attributes.. For instance, given a soil spectral library (Xr) and a set of soil vis–NIR spectra measured in the field (Xu), it is possible to use a distance metric algorithm for searching the samples in Xr which are most similar to the Xu samples. Once the most similar samples have been found, specific soil models representing the field data can be calibrated. By using this procedure, redundant information as well as noisy or non–informative samples (regarding the field spectral variability) in the soil spectral library can be removed in order to infer the target soil attribute in the field. In this sense, reliable distance metrics such as oPC–M, LLE–M and σLLE–M are necessary to select from the soil spectral library samples which are actually similar to the samples collected in the field in terms of both soil vis–NIR spectra and soil composition. Furthermore, this could have implications on the generalization capacity of the calibrated models.

For calibration of soil models based on proximal soil sensing data collected at high spatial resolution is of fundamental importance to select adequately the subset of calibration samples to be used and analyzed (Christy, 2008). This implies that the data space must be efficiently covered by the calibration samples in order to ensure good prediction results. In this respect, the oPC–M, LLE–M and σLLE–M distances are potentially useful for selecting the subset of representative calibration samples. We consider that calibration sampling algorithms based on distances such as Kennard Stone (Kennard and Stone, 1969) and fuzzy c–means sampling (de Gruijter *et al.*, 2010) could be improved by using either by using oPC–M or LLE–M or σLLE–M distances.

Furthermore, in the case of proximal vis–NIR sensing where many outlier samples can arise (due to uncontrolled field conditions) a correct identification of such outliers is required since they can degrade the prediction performance of soil vis–

NIR models. The standard approach for outlier identification in proximal vis–NIR sensing is to calculate the Mahalanobis distance on the principal components retained by using the explained variance criteria (e.g. Stevens *et al.*, 2010; Knadel *et al.*, 2011). However, here we showed that the distances estimated by using this strategy are not accurate enough for describing the compositional similarity between samples. Therefore the implementation of oPC–M, LLE–M and σLLE–M for outlier detection in proximal vis–NIR sensing data might result better than the conventional approach.

6. Conclusions

In this work we showed that the soil vis–NIR similarity is directly related to the soil compositional similarity. In order to assess the similarity between soil vis–NIR spectra we used the following distance metrics: ED, MD, SAM, SID, SDS, PC–M, oPC–M, LLE–M and σLLE–M. Given a set of soil vis–NIR samples (unknown set) we searched in a reference set the most similar vis–NIR spectrum of each unknown sample. We compared the resulting subset of most similar samples against the unknown set in terms of clay content and pH values (compositional similarity).

We found that the information about the compositional similarity is useful for obtaining reliable distance measurements. The best distance metric approaches are those that better reflect the soil compositional similarity. In general, our results indicate that the distances computed in the spectral space have a lower performance in comparison to the ones computed in low dimensional spaces. Despite this, we found that in the projected PC space the conventional selection of the number of PC features can lead to a loss of information which is important for soil similarity analysis. In this sense, other optimization methods such the one used in the oPC–M that take into account the soil compositional similarity should be considered.

The worst results were obtained by using the MD method. This is attributed to the fact that in this method the covariance matrix is computed in the vis–NIR spectral space which does not reflect well the relationships in the spectral fea-

tures. For this reason the classical estimates of the covariance matrix in the original vis–NIR space should be avoided.

Finally we showed that the oPC–M, the LLE–M and the oLLE–M methods outperformed the current approaches used for soil vis–NIR distance measurements and they can be safely used for soil vis–NIR similarity measurements.

Acknowledgments

We wish to thank ICRAF and ISRIC for making the global soil vis–NIR library available. We also thank the anonymous reviewers for their constructive comments.

References

Abou-Moustafa, K., Ferrie, F. 2008. Regularized minimum volume ellipsoid metric for query-based learning. Proceedings - 7th International Conference on Machine Learning and Applications, ICMLA 2008 , art. no. 4724974 , pp. 188-193.

Ben-Dor, E., J.R. Irons, and G.F. Epema. 1999. Soil reflectance. p. 111–188. In N. Rencz (ed.) Remote sensing for the earth sciences: Manual of remote sensing. Vol. 3. John Wiley & Sons, New York.

Ben-Dor, E., Taylor, R.G., Hill, J., Demattê, J.A.M., Whiting, M.L., Chabrillat, S., and Sommer, S. 2008. Imaging spectrometry for soil applications. Advances in Agronomy 97, 321–392.

Brereton, R.G. Chemometrics: data analysis for the laboratory and chemical plant, Chichester: Wiley, 2003, 489 pp.

Brown, D.J. 2007. Using a global VNIR soil-spectral library for local soil characterization and landscape modeling in a 2nd-order Uganda watershed. Geoderma 140, 444–453.

Chang, C.I. 2000. An information theoretic– based approach to spectral variability, similarity and discriminability for hyperspectral image analysis. IEEE Transactions on Information Theory 46, 1927–1932.

Chen, G., Qian, S.-E. 2007. Dimensionality reduction of hyperspectral imagery using improved locally linear embedding. Journal of Applied Remote Sensing 1, 013509.

Chen, J., Liu, Y. 2011. Locally linear embedding: a survey. Artificial Intelligence Review 36, 29–48.

Christy, C.D. 2008. Real-time measurement of soil attributes using on-the-go near infrared reflectance spectroscopy. Computers and Electronics in Agriculture 61, 10–19.

Chu, X.–L., Xu,Y.–P., Tian, S,–B., Wang, Z., Lu, W.–Z. 2011. Rapid identification and assay of crude oils based on moving–window correlation coefficient and near infrared spectral library. Chemometrics and Intelligent Laboratory Systems 107, 44–49.

Clark, R.N., Roush, T.L. 1984. Reflectance Spectroscopy: Quantitative Analysis Techniques for Remote Sensing Applications. Journal of Geophysical Research 89, 6329–6340.

de Gruijter, J.J .and McBratney, A. 2010. Sampling for High-Resolution Soil Mapping. In: Proximal Soil Sensing, Progress in Soil Science, edited by R. A. Viscarra Rossel, A. B. McBratney, B. Minasny, Springer Netherlands, Netherlands, p. 3–14.

De Maesschalck, R., Jouan–Rimbaud, D. & Massart, D.L. 2000. The Mahalanobis distance. Chemometrics and Intelligent Laboratory Systems 50, 1–18.

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

Demattê, J.A.M., Garcia, G.J. 1999. Alteration of soil properties through a weathering sequence as evaluated by spectral reflectance. Soil Science Society of America Journal 63, 327–342.

Farifteh, J., Van Der Meer, F., Carranza, E. J. M. 2007. Similarity measures for spectral discrimination of salt-affected soils. International Journal of Remote Sensing 28, 5273–5293.

Ge, Y., Morgan, C.L.S., Thomasson, J.A. Using soil spectral libraries in support of proximal soil sensing. The Second Global Workshop on Proximal Soil Sensing, Montreal 2011.
http://adamchukpa.mcgill.ca/gwpss/Papers/GWPSS_2011_Morgan.pdf

Gemperline, P.J., Kalivas J.H. 2006. Sampling theory, distribution functions and the multivariate normal distribution. In: Practical Guide to Chemometrics, Second Edition, editor P.J. Gemperline, CRC Press Taylor & Francis Group, Boca Raton, Florida. pp.41–68

He, Y., Liu, D., Yi, S. 2011. Recursive spectral similarity measure-based band selection for anomaly detection in hyperspectral imagery. Journal of Optics 13, 015401–015402

IUSS Working Group WRB 2006. World Reference Base for Soil Resources 2006. 2nd edition. Rome FAO, World Soil Resources Reports No. 103.

Kennard, R.W., Stone, L. 1969. Computer aided design of experiments. Technometrics 11, 137–148.

Knadel, M., Thomsen, A., Mogens, G. 2011. Multisensor on-the-go mapping of soil organic carbon content. Soil Science Society of America Journal 75, 1799-1806.

Kullback, S., R. A. Leibler, R.A. 1951. On information and sufficiency. The Annals of Mathematical Statistics 22, 79–86.

Laskar, R.H., Talukdar, F.A., Paul, B., Chakrabarty, D. 2011. Sample reduction using recursive and segmented data structure analysis. Journal of Engineering and Computer Innovations 2, 59–67.

Lugassi, R., Ben-Dor, E., Eshel, G. 2010. A spectral-based method for reconstructing spatial distributions of soil surface temperature during simulated fire events. Remote Sensing of Environment 114 , 322–331

Ma, L., Crawford, M.M., Tian, J. 2010. Anomaly detection for hyperspectral images based on robust locally linear embedding. Journal of Infrared, Millimeter, and Terahertz Waves 31, 753–762.

Minasny, B., McBratney. A. B. 2006. A conditioned Latin hypercube method for sampling in the presence of ancillary information. Computers and Geosciences 32, 1378–1388.

Park, B., Windham, W. R., Lawrence, K. C., Smith D. P. 2007. Contaminant classification of poultry hyperspectral imagery using a spectral angle mapper algorithm, Biosystems Engineering, 96, 3, 323-333

Pan, Y., Ge, S.S., Al Mamun, A. 2009. Weighted locally linear embedding for dimension reduction. Pattern Recognition 42, 798–811.

R Development Core Team. 2011. R: A language and environment for statistical computing. Vienna: R Foundation for Statistical Computing.

Roweis, S.T., Saul, L.K. 2000. Nonlinear dimensionality reduction by locally linear embedding. Science 290, 2323–2326.

Sankey, J.B., Brown, D.J., Bernard, M.L., Lawrence, R.L. 2008. Comparing local vs. global visible and near-infrared (VisNIR) diffuse reflectance spectroscopy (DRS) calibrations for the prediction of soil clay, organic C and inorganic C. Geoderma 148, 149–158.

Saul, L.K., Roweis, S.T. 2003. Think globally, fit locally: unsupervised learning of low dimensional manifold. Journal of Machine Learning Research 4, 119–155.

Shepherd, K.D., Palm C.A., Gachengo C.N., Vanlauwe B. 2003. Rapid characterization of organic resource quality for soil and livestock management in tropical agroecosystems using near–infrared spectroscopy. Agronomy Jounal 95,1314–1322.

Stenberg, B., Viscarra Rossel, R.A., Mouazen, A.M., Wetterlind, J. 2010. Visible and near infrared spectroscopy in soil science. In: Donald L. Sparks, Ed. Advances in Agronomy, Vol. 107, Burlington: Academic Press, pp. 163–215.

Stevens, A., Udelhoven, T., Denis, A., Tychon, B., Lioy, R.,Hoffmann, L., van Wesemael, B. 2010. Measuring soil organic carbon in croplands at regional scale using airborne imaging spectroscopy. Geoderma 158, 32–45.

Stoner, E. R.; Baumgardner, M. F. 1981. Characteristics variations in reflectance of surface soils. Soil Science Society of America Journal 45, 1161–1165.

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, 1792–1799.

van der Meer, F. 2006. The effectiveness of spectral similarity measures for the analysis of hyperspectral imagery. International Journal of Applied Earth Observation and Geoinformation 8, 3–17.

Van Reeuwijk L.P. 2002. Procedures for Soil Analysis. 6th Edition. International Soil Reference and Information Centre, Wageningen, and Food and Agricultural Organization of the United Nations, Rome.

Varini, C., Degenhard, A., Nattkemper, T.W. 2006. ISOLLE: LLE with geodesic distance. Neurocomputing 69, 1768–1771.

Viscarra Rossel, R. 2009. The Soil Spectroscopy Group and the development of a global soil spectral library. NIR news 20, 14–15.

Viscarra Rossel, R., and Behrens, T. 2010. Using data mining to model and interpret soil diffuse reflectance spectra. Geoderma 158, 46–54.

Wetterlind, J., Stenberg, B. 2010. Near-infrared spectroscopy for within-field soil characterization: small local calibrations compared with national libraries spiked with local samples. European Journal of Soil Science 61, 823–843.

Wold, H. 1966. Estimation of principal components and related models by iterative least squares. In: P.R. Krishnaiah, editor, Multivariate Analysis. Academic Press, New York. pp. 391–420.

World Agroforestry Centre (ICRAF) and ISRIC – World Soil Information. 2010. ICRAF–ISRIC Soil vis–NIR spectral Library. Nairobi, Kenya: World Agroforestry Centre (ICRAF).

Yuhas, R. H., Goetz, A.F.H., Boardman, J. W. 1992. Discrimination among semi-arid landscape endmembers using spectral angle mapper (SAM) algorithm, Summaries of the 4th Annual JPL Airborne Geoscience Workshop, JPL Pub-92-14, AVIRIS Workshop. Jet Propulsion Laboratory, Pasadena, CA, pp. 147-150.

Manuscipt 4: The spectrum–based learner: a new local approach for modeling soil vis–NIR spectra of complex datasets

Geoderma, 2013, vol. 195–196, p. 268–279 doi: 10.1016/j.geoderma.2012.12.014

Leonardo Ramirez-Lopez[a,b], Thorsten Behrens[a], Karsten Schmidt[a], Antoine Stevens [b], Jose A. M. Demattê[c], Thomas Scholten[a]

[a]Institute of Geography, Physical Geography and Soil Science, University of Tübingen,
Rümelinstraße 19–23, 72074, Tübingen, Germany.

[b]Georges Lemaître Centre for Earth and Climate Research, Earth and Life Institute,
Université Catholique de Louvain,
3 Place Louis Pasteur–1348, Louvain la Neuve, Belgium.

[c]University of São Paulo, Escola Superior de Agricultura "Luiz de Queiroz"
Soil Science Department.
Av.: Pádua Dias, 11 CP 9. Piracicaba–SP 13418–900. Brazil.

Abstract

This paper shows that *memory–based* learning (MBL) is a very promising approach to deal with complex soil visible and near infrared (vis–NIR) datasets. The main goal of this work was to develop a suitable MBL approach for soil spectroscopy. Here we introduce the spectrum–based learner (SBL) which basically is equipped with an optimized principal components distance (oPC–M) and a Gaussian process regression. Furthermore, this approach combines local distance matrices and the spectral features as predictor variables. Our SBL was tested in two soil spectral libraries: a regional soil vis–NIR library of State of São Paulo (Brazil) and a global soil vis–NIR library. We calibrated models of clay content (CC), organic carbon (OC) and exchangeable Ca (Ca^{++}). In order to compare the predictive performance of our SBL with other approaches, the following algorithms were used: partial least squares (PLS) regression, support vector regression machines (SVM), locally weighted PLS regression (LWR) and LOCAL. In all

cases our SBL algorithm outperformed the accuracy of the remaining algorithms. Here we show that the SBL presents great potential for predicting soil attributes in large and diverse vis–NIR datasets. In addition we also show that soil vis–NIR distance matrices can be used to further improve the prediction performance of spectral models.

Keywords: soil similarity; machine learning; local modeling; memory–based learning; soil spectral library; nearest neighbor.

1. Introduction

It has been demonstrated that soil visible and near infrared (vis–NIR) spectroscopy can be used to predict multiple soil attributes accurately (Viscarra Rossel *et al*., 2006; Stenberg *et al*. 2010). Soil vis–NIR libraries have become powerful tools in soil science helping to analyze and store large amounts of soil information in an efficient way (McBratney *et al*., 2006). However, the accuracy of the models usually decreases when the dataset contains very diverse samples in terms of geographical origin, mineralogy, parent material, environmental conditions, etc. For instance, Stenberg *et al*. (2010) showed with a literature review that there is positive correlation between the error of the models and the standard deviation of the soil property under investigation. Similarly, Savvides *et al*. (2010) demonstrated that the spatial relationship between soil reflectance and cation exchange capacity is scale-dependent. As a consequence, modeling soil attributes using large and diverse soil vis–NIR libraries still remains a challenging task.

In contrast to pure component systems, soil is a very complex mixture of mineral and organic constituents. Soil vis–NIR spectra result from overtones and combination bands of primary absorptions in the mid infrared region of the electromagnetic spectrum. Therefore, soil constituents present weak, broad and sometimes overlapping vis–NIR spectral responses.

The relationship between vis–NIR spectra and soil properties can vary under different soil mineralogy and their content in soil organic matter since they are the

main spectrally–active components of the soil (Stoner and Baumgardner, 1981; Brown *et al.*, 2006; Ben-Dor *et al.*, 2008; Stenberg *et al.*, 2010; Viscarra Rossel *et al.*, 2011). Information about the variability of soil attributes with a direct and strong influence on the vis–NIR features can be very important for developing models of other soil attributes. For instance, the vis–NIR spectral response of exchangeable bases is associated with the clay minerals and their amount (Madejová, 2003). Hence, soils with different mineralogy will show different spectral responses of the cations held by their clay minerals. This implies that in a given vis–NIR dataset, soil attributes associated to spectrally active constituents (e.g. minerals and organic matter) cannot be expected to be globally stable (Stenberg *et al.*, 2010), however, they can be locally stable. In this sense, when spectral variations associated to mineralogy and organic matter are reduced, the spectral variation of several other soil attributes can be highlighted. This can explain why local models work usually better than the global models. Nevertheless, vis–NIR dataset partitioning for local modeling is also challenging due to the complexity problems mentioned previously.

One reasonable approach for reducing the complexity of a given soil vis–NIR dataset (X), which is very heterogeneous, is to split X into *c* partitions or clusters, so that samples in the same partition share similar soil characteristics. In this sense, the complexity in each partition must be lower than the global complexity contained in X. In general in soil science and specifically in soil spectroscopy, several studies have demonstrated that prediction models based on (either spectrally or geographically) local partitions perform better than single or global models. In many cases the use of geographical information for partitioning a spectral dataset results in reduction of the soil (spectral) variability within each partition in comparison to the global soil (spectral) variability. Stevens *et al.* (2010) observed that vis–NIR local models of soil organic carbon perform better than global models when the soil dataset is partitioned into different soil texture classes and agro–pedological regions. They also showed that the organic carbon variability within each partition is lower than the organic carbon variability of the entire area. Guerrero *et al.* (2010) used different regional calibration sets for predicting soil attributes in each region. For modeling soil attributes in different agricultur-

al fields, Wetterlind and Stenberg (2010) used models calibrated with a national soil vis–NIR library, and models calibrated only with local samples taken from those fields. They observed that the local models outperformed the national soil vis–NIR models. Janik *et al.* (2007) suggested that local calibrations of soil spectroscopic models based on the minimization of changes in soil type may be more accurate than global calibrations. Similar conclusions are reported on the analysis of soil data for digital soil mapping. When the variability patterns of a given soil attribute differs between geomorphological or pedological regions, they should be modeled separately (McBratney *et al.*, 1991; Schmidt *et al.*, 2010).

1.2 Memory–based learning

In machine learning theory, *memory–based* learning (MBL) (a.k.a. *instance–based* learning) is a data–driven technique. It can be defined as a lazy learning approach which is closely related to case based reasoning (CBR). Like CBR, MBL resembles the human reasoning process (An, 2005): remember previous situations, adapt them for solving the current problem, examine the probability to solve the problem with the new solution, and memorize the experience for knowledge improvement. The main difference between CBR and MBL is that CBR uses knowledge–based reasoning rather than statistical methods (Mitchell, 1997). In contrast to other learning methods, the main goal in MBL is not to derive general or global target function. Instead, when a solution for a new problem is required, the experience in the form of a set of similar related samples is retrieved from memory and then those samples are combined to construct the solution to the new problem. Hence, for each new problem a new target function is derived. While a global target function may be very complex, MBL can describe the target function as a collection of less complex local (or locally stable) approximations (Mitchell, 1997). In this sense, non–linear relationships can be easily resolved. In contrast to complex learning algorithms such as neural networks or support vector machines most of the MBL systems do not require a complex function fitting process (Kang and Cho, 2008). The k-nearest neighbor algorithm is

one of the most widely known algorithms which belongs to the family of MBL methods (Solomatine *et al.*, 2008, Mitchel, 2011).

In the literature MBL is sometimes referred to as local modeling, nevertheless local modeling comprises other approaches such as cluster–based modeling and geographical segmentation–based modeling, etc. Hence, MBL can be described as a type of local modeling.

Several MBL approaches have been useful for solving difficult tasks in areas such robotics, linguistics, medical diagnosis, image analysis, etc. For example, in hydrological forecasting Solomatine *et al.* (2008) reported that the performance of MBL approaches is often better than performance of complex algorithms such artificial neural networks and model trees. In general the use of CBR and MBL in soil science research is not new, several authors have already implemented it for soil erosion modeling (Meyer *et al.*, 1992), digital soil mapping (e.g. Zhu, and Liu, 2011; Shi *et al.*, 2004; Shi, *et al.*, 2009; Qui et ql., 2006) and hydrological analysis (e.g. Ostfeld and Salomons 2005; See, 2008; Solomatine *et al.*, 2008; Akbari *et al.*, 2011). Shi *et al.*, 2004, mention that similarity–based inference can be an effective approach to knowledge acquisition and knowledge representation for digital soil mapping.

MBL approaches such as locally weighted partial least squares regression (LWR, Naes *et al.*, 1990), LOCAL (Shenk *et al.*, 1997), locally biased regression (Fearn and Davies, 2003) and comparison analysis using restructured near-infrared and constituent data (CARNAC–D, Davies and Fearn, 2006) have been successfully applied in vis–NIR spectroscopy. However, MBL has received little attention in soil spectroscopy and only few studies to date have reported its use. Chang *et al.* (2001) implemented a local principal component regression approach for performing vis–NIR-based predictions of multiple soil attributes of samples collected from several regions in the United States. In their approach, for each soil sample in the prediction set, its 30 most similar samples in a reference set were used for fitting a local model. They found high prediction accuracy for all the attributes under study. Christy and Dyer (2005), compared LWR to other regression algorithms for calibrating models of soil attributes based on data from multiple on-

the-go soil sensors (including a vis–NIR spectrometer). They report that highest prediction accuracies for most of the soil attributes evaluated were produced by LWR. Igne *et al*. (2010) used the LWR algorithm for predicting soil attributes from a NIR and mid–infrared library of samples collected at the field scale. Genot *et al*. (2011) used the LOCAL algorithm for predicting soil attributes using a regional soil NIR library. Both works demonstrated that soil attribute models calibrated with the LOCAL and the LWR regressions outperformed the global models. Fernández Pierna and Dardenne (2008) found that the LOCAL algorithm has better performance than global PLS and least squares SVM for predicting soil properties. Gogé *et al*. (2012) evaluated different local PLS modeling approaches for predicting soil attributes using a vis–NIR soil spectral library of France. They obtained high prediction accuracy with similar results for all the local approaches tested.

The main drawbacks of MBL methods lie on the computational costs and the similarity measure used for recovering samples from the memory (nearest neighbor search). For example, in both LOCAL and LWR algorithms, for n samples in a given prediction set, n models need to be calibrated. This implies that n optimizations for choosing the adequate number of PLS factors must be carried out. This can be particularly problematic when the training set includes a large number of samples and features or predictor variables. Concerning the second problem, since the accuracy of MBL relies entirely on a set of similar samples, inadequate strategies for measuring the similarity might degrade the performance of the MBL approach (Lopez de Mantaras *et al*., 2006).

In this context, our main goal was to develop a high-performance MBL for modeling complex soil spectral data. For this purpose we introduced the spectrum–based learner (SBL) which is equipped with an optimized principal components distance (oPC–M) and a Gaussian process algorithm with a linear covariance function (GPL). Our SBL can be described as a locally linear Gaussian process modeling approach which combines local distance matrices and the spectral features as source of predictor variables.

2. Theory

2.1 Linear Gaussian process regression

Gaussian process (GP) regression is equivalent (in geostatistics) to widely known kriging interpolation which has been extensively used in pedometrics research. However here, instead using geographical coordinates as input data, we use multivariate vis–NIR data. Among GP (Williams and Rasmussen, 1996) algorithms, the linear Gaussian processes (GP) uses linear covariance function or linear kernel. The GP regression is a probabilistic, non-parametric Bayesian approach. A GP is commonly described as a collection of random variables which have a joint Gaussian distribution and it is characterized by both a mean and a covariance function. In general, Gaussian processes regression is a powerful algorithm for function approximation in high–dimensional spaces (Rasmussen and Williams, 2006).

In GP regression, given a dataset of N samples with predictors ($\mathrm{X} = \{\mathrm{x}_i\}_{i=1}^{N}$) and a target attribute ($\mathrm{Y} = \{\mathrm{y}\}_{i=1}^{N}$) the function y(x) can be described by a Gaussian distribution Y ~G(0, C) where C is a $N \times N$ covariance matrix which can be defined via a covariance function (a.k.a kernel function) where

$$C = \mathrm{K}(\mathrm{X},\mathrm{X}) \quad (1)$$

There are several algorithms among the large family of covariance functions (e.g. linear, squared exponential, Gaussian, polynomial, laplacian, spline, Bayesian, wavelet, etc). Covariance functions implicitly perform a mapping of the original feature space into a high dimensional space with linear or nearly linear structure. This is also referred to as "kernel trick" (Aizerman *et al*, 1964; Schölkopf and Smola, 2002). Since GP is equivalent to kriging, the geostatistcal counterparts of the covariance functions that are used in GP are the represented by variogram models (Gneiting *et al.*, 2001).

The only free parameters to be optimized in a GP regression are the hyperparameters of the covariance function to be used. Our SBL uses the following linear covariance function for computing the local regressions:

$$K(\mathrm{X},\mathrm{X}) = \mathrm{X}^{\mathrm{T}}\mathrm{X} \tag{2}$$

In general, the covariance function can be described as:

$$K(\mathrm{X},\mathrm{X}) = \begin{bmatrix} k(\mathrm{x}_1,\mathrm{x}_1) & \cdots & k(\mathrm{x}_1,\mathrm{x}_N) \\ \vdots & \ddots & \vdots \\ k(\mathrm{x}_N,\mathrm{x}_1) & \cdots & k(\mathrm{x}_N,\mathrm{x}_N) \end{bmatrix} \tag{3}$$

For predicting a new sample at x_{new}, the predictive distribution of y_{new} is also Gaussian distributed, with mean:

$$\bar{y}_{new} = k^T(\mathrm{x}_{new})\mathrm{C}^{-1}Y \tag{4}$$

and variance:

$$\sigma^2_{\bar{y}_{new}} = C(\mathrm{x}_{new},\mathrm{x}_{new}) - k^T(\mathrm{x}_{new})\mathrm{C}^{-1}k(\mathrm{x}_{new}) \tag{5}$$

where:

$$k(\mathrm{x}_{new}) = [k(\mathrm{x}_{new},\mathrm{x}_1), k(\mathrm{x}_{new},\mathrm{x}_2), \ldots, k(\mathrm{x}_{new},\mathrm{x}_N)]^T \tag{6}$$

Further details on GP regression and covariance functions can be found in Rasmussen and Williams (2006), and Chen *et al.* (2007).

2.2 The spectrum–based learner (SBL)

Here we introduce our memory–based learning approach which is called spectrum–based learner (SBL). The SBL is a three step approach which comprises: *i.* Nearest neighbor search (recovering), *ii.* training and testing, and *iii.* fitting and predicting. As other MBL methods, the SBL does not yield a global function, instead it performs local interpolations which are based on a reference set or spectral library.

2.2.1 Nearest neighbor search (recovering)

The main goal of this step is to discover which samples in a reference set "resemble" the samples to be predicted. Recovering similar samples from a set of samples stored in a "memory" (reference set) implies that similarity or dissimilarity measurements must be carried out. For these measurements a distance matrix can be used. In the SBL the nearest neighbor search process is carried out by using an optimized principal components distance (oPC–M, Ramirez-Lopez *et al.*, 2012) which indicates how similar or dissimilar the vis–NIR spectra are to the samples to predict.

Usually for computing distances between samples in a vis–NIR dataset, the spectra are compressed by using principal component analysis. This produces a set of score variables (S) on which the Mahalanobis distance is computed. Only the first S variables that accounted for a given percentage of the original spectral variability are used. This procedure has been widely used by the vis–NIR spectroscopy community (Naes *et al*. 2002, De Maesschalck *et al*., 2000). However this approach does not take into account the information related to the soil composition of the samples.

In the oPC–M distance we also use a nearest neighbor approach for selecting the number of optimal principal components that better represents the distance or similarity in terms of soil composition (e.g. clay content, organic carbon, exchangeable Ca) between samples.

Given a set of n reference samples $(\mathrm{Xr}, \mathrm{Yr}) = \{\mathrm{xr}_i, \mathrm{yr}_i\}_{i=1}^{n}$ and a set of m samples $(\mathrm{Xu}, \mathrm{Yu}) = \{\mathrm{xu}_j, \mathrm{yu}_j\}_{j=1}^{m}$ (where Yu values are unknown) the spectra of X = {Xr ∪ Xu} are compressed by using principal component (PC) analysis, obtaining a matrix of scores (S) of $N \times p$ dimensions. The singular value decomposition algorithm was used for the PC analysis. The Mahalanobis distance (MD) matrix is computed by varying the number of PC variables (p), so that $p = (1,2,3\ldots, t)$ where t is a user–defined threshold which must be lower or equal to the number of spectral features. A set of t MD matrices are obtained as a function of the number of PCs:

$$MD^2(p) = (S - s_o)M^{-1}(S - s_o)^T \tag{7}$$

where:

$$S = \begin{bmatrix} s_{1,1} & \cdots & s_{1,p} \\ \vdots & \ddots & \vdots \\ s_{N,1} & \cdots & s_{N,p} \end{bmatrix}, \tag{8}$$

$N = n+m$, p is the number of principal components to be retained, $o = 1,2,3,\ldots N$, and M^{-1} is inverse of the variance–covariance matrix of S.

By using each MD(p) matrix, the nearest neighbor (NN) of each xr_i is also selected from Xr as a function of p. i.e. for each p a set of NNs is obtained. The NN of each sample indicates its most similar sample in terms of its vis–NIR principal components. The samples and their NNs are also compared in terms of soil compositional attributes (e.g. clay content, organic carbon, exchangeable Ca, etc). The

optimal number of PCs is the one that minimizes the root mean square of compositional differences (RMSD) between $\mathrm{yr_i}$ and $\ddot{\mathrm{y}}\mathrm{r_i}(p)$:

$$\mathrm{RMSD}(p) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\mathrm{yr}_i - \ddot{\mathrm{y}}\mathrm{r}_i(p))^2} \quad (9)$$

where yr_i is the soil attribute value of each sample and $\ddot{\mathrm{y}}\mathrm{r}_i(p)$ is the soil attribute value of its corresponding nearest neighbor in vis–NIR principal component space with p dimensions (note that $\ddot{\mathrm{y}}\mathrm{r}_i(p) \in \mathrm{Yr}$). In other words, the number of vis–NIR principal components that better represents the soil compositional similarity is taken as the optimal for the oPC–M distance computation. The rationale behind this approach is based on the fact that soil vis–NIR variability should reflect the soil compositional variability as well, at least in terms of those attributes that have strong influence on the vis–NIR spectra. Note that in this approach the soil compositional information is not used in the PC analysis, it is only used for selecting the optimal number of PCs. If compositional information were included in the PC analysis, then compositional information of the samples to be predicted would be also needed. And we assume that that information is unknown.

2.2.2 Training and testing

Training and testing are carried out in the spectral space. For each xu_i a model must be fitted by using its most similar samples i.e. its k-nearest neighbors. However prior fitting is necessary to determine the optimal number of neighbor samples (k) to be used in each calibration. In this respect, k must be optimized since it can affect the fitting process. For example, if the k is too small the calibration for xu_i can be highly affected by noise and outliers. On the other hand, if k is too large the calibration for xu_i can be affected by non–linear relationships.

For each xu_i its corresponding most spectrally similar sample in the reference set (Xr) (i.e. its first nearest neighbor in Xr) or soil spectral library is selected (based on the the oPC–M distance matrix computed in the previous step), resulting in a set of most similar samples $(\mathrm{Xru}, \mathrm{Yru}) = \{\mathrm{xru}_j, \mathrm{yru}_j\}_{j=1}^{m}$ where $\mathrm{Xru} \subset \mathrm{Xr}$ and $\mathrm{Yru} \subset \mathrm{Yru}$. The subset Xru can be viewed as the subset in the reference set that better

reproduce the soil variability of the prediction samples (Xu, Yu) therefore it can be exploited for optimizing k.

From Xr a given number of neighbors (k) of each xru_j are retained and used for the calibration of a model for predicting yru_j. In the calibration process the local distance matrix between the neighbors (a $k \times k$ matrix which is extracted from the oPC–M matrix) is included as an additional set of predictors. A Gaussian process regression with a linear covariance function (GPL) is used to predict the target soil attribute (yru_j) corresponding to xru_j. These predictions are cross-validated and the root mean square error of the predictions of Yur (RMSE_{ru}) is then computed as:

$$\text{RMSE}_{ru} = \sqrt{\frac{1}{m}\sum_{j=1}^{m}(\text{yru}_j - \widehat{\text{yru}}_j)^2} \tag{11}$$

where $\widehat{\text{yru}}_j$ is each predicted value. This process is repeated for a set of different number of neighbors and the idea is to identify the best number of neighbors (k) for calibrating models for Xru (the subset of m samples in the reference set which better mimics or resemble the spectra of the samples to be predicted) and then use it for calibrating the models for Xu. The optimal k is the one that minimizes the RMSE_{ru} of the prediction of Yru.

The SBL does not use a distance-based weighting approach, i.e. for the local regressions we did not assign weights to the neighbors based on their distances to the target sample. The reason for that is twofold. First, distance-based weighting implies the modification of all the spectral variables. Therefore if a distance score does not represent properly the similarity/dissimilarity between the samples, then it will affect the entire set of predictors of the sample which was weighted with the "noisy" distance score. Secondly, in the weighting approach the information about the position of the samples within the neighborhood is missing since only the information about the distance to the target sample is employed.

In our SBL approach the use of the linear covariance for the Gaussian process regression is motivated by two main reasons: *i.* our hypothesis is that complex covariance functions are not required since the complexity of each subset of sam-

ples for each local model is low, and *ii.* the linear covarinace function has no hyperparameters to be optimized and therefore simplifies and speed up the local modeling process.

2.2.3 Fitting and predicting

Once the optimal k is found, a new local GPL model is fitted for each xu_i with its k-nearest neighbors found in Xr. The predictors are the vis–NIR reflectance and the distance values of each xu_i with respect to its k-nearest neighbors. After each calibration the prediction of the target attribute is carried out. Figure 1 shows a summary of the above steps carried out in the SBL approach.

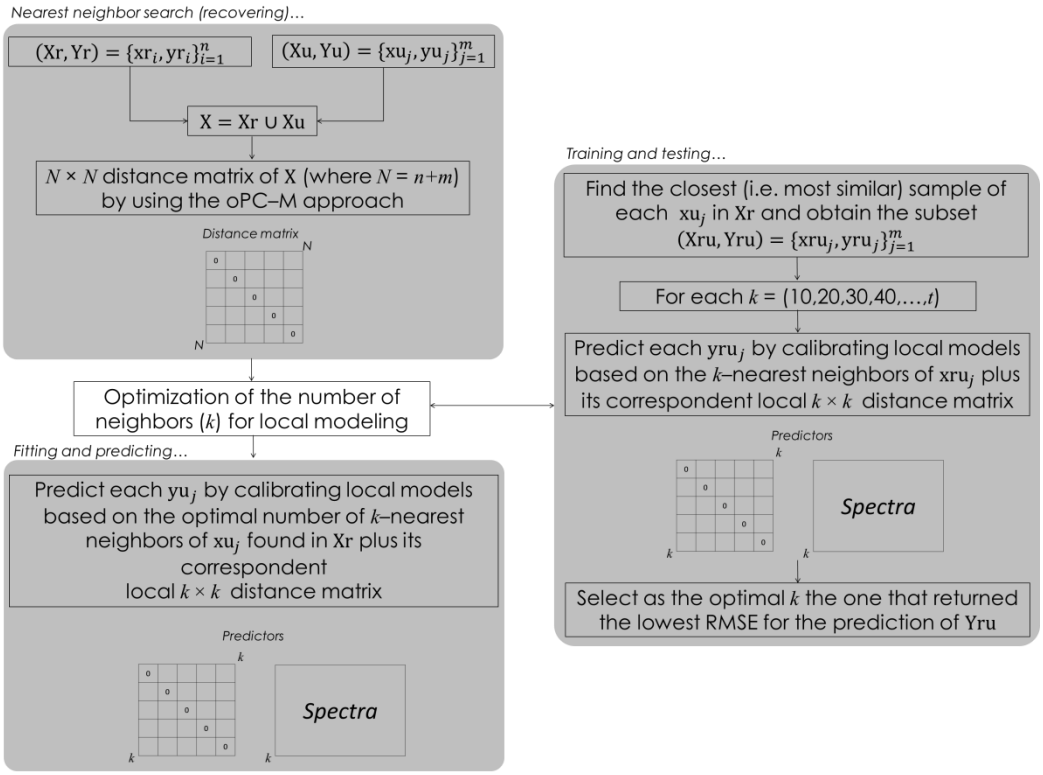

Figure 1. Description of the recovering, training and testing, and fitting and predicting steps of the SBL approach.

Overall, an important characteristic in our SBL approach is that the only parameters to be optimized are: the optimal number of principal components for the oPC–M distance matrix computation and the number of nearest neighbors for (k) for the local GPL regressions. Another characteristic is that instead using the distance information to assign weights (as in LWR) the SBL local distance matrices are used as additional predictors.

3. Materials and methods

3.1 The vis–NIR soil spectral libraries

In order to test our spectrum–based learner (SBL), we used two vis–NIR soil libraries: a regional soil spectral library (R–SSL) of the State of São Paulo (Brazil) and a global soil spectral library (G–SSL). The analyses were carried out separately for each soil spectral library.

3.1.1 The regional soil spectral library of the State of São Paulo (R–SSL)

The soil samples in the R–SSL are historical samples which were collected for soil survey purposes in different agricultural fields located in 11 sub-regions of the State of São Paulo (Brazil). The R–SSL comprises 4200 samples including 927 soil profiles (2781 samples) collected at three depths 0–20 cm, 40–60 cm, and 80–100 cm. In terms of texture, the variability of the samples in the R–SSL is large (Figure 2a). The R–SSL comprises soils of 10 soil groups (Arenosols, Leptosols, Cambisols, Ferralsols, Nitisols, Lixisols, Alisols, Acrisols, Gleysols and Planosols) according to the World Reference Base (WRB) for soil resources (IUSS Working Group WRB, 2006). Nevertheless, the predominant soil group is the Ferrasol, which is also the predominant soil group in the State of São Paulo (Oliveira *et al.*, 1999).

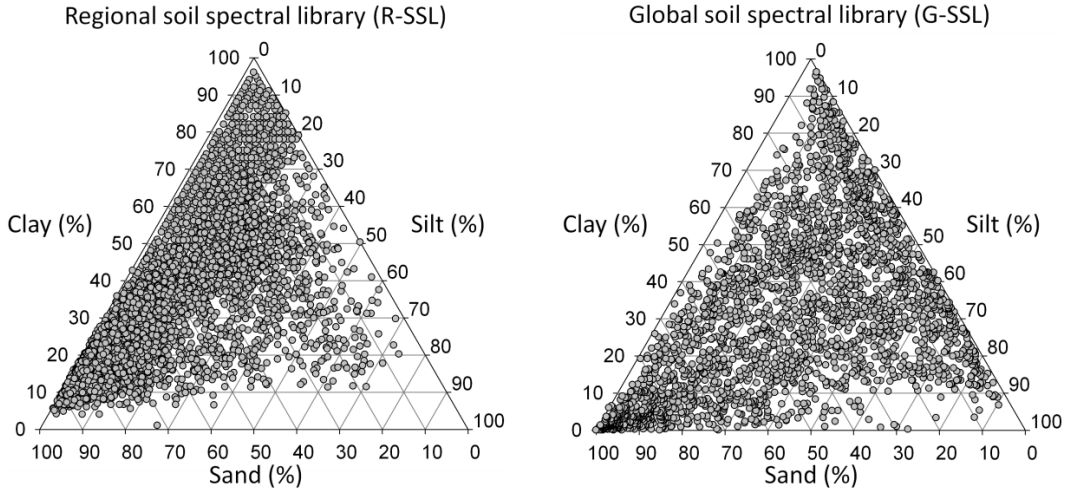

Figure 2. Texture distribution of the samples in the regional soil spectral library (left) and samples in the global soil spectral library (right).

All soil samples were dried at 45 °C for 24 h and sieved to < 2 mm.dried at 45 °C for 24 h and then sieved through a 2 mm mesh. The clay content was determined using the pipette method and exchangeable Ca was extracted from the soil with an ion exchange resin (Camargo *et al.* 2009). Soil organic carbon (OC) was measured with the Walkley–Black method (Heanes, 1984).

The vis–NIR reflectance spectra of soil samples were obtained with a FieldSpec® Pro spectrometer (ASD inc., Boulder, CO, USA) which collects the spectral measurements in a range of 350 to 2500 nm and is characterized by a Full Width Half Maximum of 3 nm for the 350-1000 nm region and 10 nm for the 1000-2500 nm region. The reflectance of each sample was calculated by taking the average of 100 scans. The spectra were resampled to a resolution of 5 nm (for a total of 431 spectral features) and then transformed to absorbance (A_λ) by:

$$A_\lambda = log_{10}\left(\frac{1}{R}\right) \tag{12}$$

where R is the reflectance.

3.1.2 The global soil spectral library (G–SSL)

The G–SSL was developed by the World Agroforestry Centre (ICRAF) and ISRIC – World Soil Information (2010). The G–SSL comprises 4438 soil samples (including 785 soil profiles), although 3643 of them have both, soil chemical attribute and texture information. For this work we only used these 3643 which include samples from 55 countries spanning America (27%), Africa (24%), Asia (23%), Europe (23%) and Oceania (3%). These soils have large texture variability (Figure 2b) and are represented by all 32 Reference Soil Groups in the WRB (IUSS Working Group WRB, 2006).

The reflectance spectra of the G–SSL samples were also recorded using a FieldSpec® Pro spectrometer. The spectra were resampled to a resolution of 10 nm for a total of 216 spectral features. See Shepherd *et al.* (2003) for more details on the optical setup and sample preparation for spectral measurements. The vis–NIR reflectance spectra were transformed to absorbance.

In this library, clay content was determined by the pipette method, the exchangeable Ca with the NH_4OAc method, and soil organic carbon (OC) with the Walkley–Black procedure (Heanes, 1984). These measurements were carried out according to the procedures for soil analysis given in Van Reeuwijk (2002).

3.2 Testing the performance of the SBL and comparison with other machine learning algorithms

Calibrations of soil attribute models were carried out separately for each soil spectral library. As validation sets, we randomly sampled 350 soil profiles (1050 samples) from the R–SSL and 125 soil profiles (900 samples) from the G–SSL. The remaining samples were used as training sets in each spectral library. The idea behind sampling entire profiles as a validation sets instead individual samples is to avoid pseudo-replication of samples (Terhoeven-Urselmans *et al.*, 2010). When samples in a validation share strong spatial correlation (e.g. belongs to the same profile) with samples in the training set, the resulting measure of model performance will be biased.

We evaluated the performance of the SBL for calibrating vis–NIR models to predict clay content (CC), soil organic carbon (OC) and exchangeable Ca (Ca^{++}) in the validation samples. These attributes in the training sets were also used for optimizing the number of PCs for the oPC–M distance matrix computation.

In addition, we use five other machine learning algorithms for predicting these attributes and for comparing their results with those obtained by the SBL approach. These algorithms were: partial least squares regression (PLS, Wold *et al.*, 1983), support vector regression machines (SVM, Drucker *et al*, 1996), locally weighted PLS (LWR, Naes *et al.*, 1990) and LOCAL (Shenk *et al.*, 1997). All the algorithms were implemented in R 2.14.1 (R Development Core Team, 2011).

The PLS algorithm has been widely used in soil spectroscopy for calibrating models of several soil attributes. The number of PLS factors is the only parameter that needs to be optimized in PLS regressions.

SVM has been widely implemented for solving complex regression and classification tasks in several fields. In soil spectroscopy, Viscarra Rossel and Behrens (2010) found that it outperforms PLS and several other non-linear algorithms. The SVM algorithm uses the so-called "kernel trick". For SVM we used a radial basis function (RBF) as covariance function. In this case a hyper-parameter of the RBF called alpha (α) and a penalty factor (C) are the parameters to be optimized.

The LWR is a non-linear version of PLS and it can be classified also as a memory–based learning algorithm. The spectra is first compressed by using principal component analysis, and then the Mahalanobis distance (MD) is computed on the first principal components which accounts for a given percentage of cumulative explained variance (set here to 99.5 %). The MD obtained is referred here to as the standard PC distance. After this procedure local PLS calibrations for each unknown sample are carried out in the spectral space using its k-nearest neighbors which are weighted according to their distance from the unknown sample. In this work, the vector of weights (W) for the k-nearest neighbors of each sample was computed by using the following tricubic function:

$$W = (1 - {d_s}^3)^3 \tag{13}$$

where d_s are the scaled distances from 0 to 1 of the k-nearest neighbors to the target unknown sample. In each local calibration the number of PLS factors must be optimized. In addition, the number of nearest neighbors (k) must be globally optimized. Only a few LWR approaches have been implemented in soil spectroscopy (e.g. Igne *et al.*, 2010), which report better mode performance compared to other algorithms such as global PLS and least squares SVM.

Similar to LWR, the LOCAL algorithm also operates by calibrating local models according to a similarity measure. There are three differences between those algorithms: *i.* in LOCAL the correlation distance between unknown samples and training samples is used for selecting the k-nearest neighbors of each unknown sample, *ii.* the LOCAL algorithm does not use any distance-based weighting function and *iii.* Each local PLS predicted value is a weighted sum of the predicted values from all the models generated between a minimum and a maximum

number of PLS factors (Fernández Pierna and Dardenne, 2008). Here we set the minimum number of PLS factors to 1. Sot that, there were two parameters to be optimized: the (maximum) number of PLS factors in each local calibration and the number of nearest neighbors (*k*).

3.2.1 *Tuning the parameters of the models*

A leave-25%-out cross validation with 10 repetitions was used for tuning the PLS factors (in global PLS, LWR and LOCAL), the *a* values of the RBF covariance and the *C* parameter in SVM. We tested 30 PLS factors for global PLS, and 56 combinations of *a* and C values for SVM calibrations. The best parameters were those that minimized the root mean squared error of cross validation (RMSE):

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\mathrm{y}_i - \hat{\mathrm{y}}_i)^2} \qquad (14)$$

where $\mathrm{y_i}$ is the predicted value, $\hat{\mathrm{y}}_\mathrm{i}$ is the observed value and *n* the number of test samples.

For local PLS models in both LWR and LOCAL we tested a maximum of 17 PLS factors. Note that the number of optimal PLS factors may vary among local models. This is due to the fact that each partition could present a different degree of complexity. For example, in cases where the variability within the neighborhood is low, then the number of optimal PLS factors will be probably low. On the other hand in cases where the variability within the neighborhood is high, then the number of optimal PLS factors will be probably high.

For selecting the appropriate number of nearest neighbors (*k*) in the LWR, LOCAL and in our SBL we tested from 30 to 400 samples in steps of 10. The optimal *k* for LWR and LOCAL were those that minimized the average of the RMSEs of local internal leave-25%-out cross validations. In the case of our SBL (as explained previously in section 2.2) the best *k* was chosen as the one that minimized the RMSE of the prediction of the most (spectrally) similar samples of the

test set (Xru) (the subset of *m* samples in the reference set which better mimics or resemble the vis–NIR spectra of the samples to be predicted).

In order to assess the global predictive capability of the SBL approach, we performed a 10-fold leave-25%-out cross validation in each local model using only the optimal number of neighbors. The global RMSE was taken as the average of the RMSE of local models.

We did not compute the training R^2 since those values are not comparable between the different approaches. The training R^2 of memory based models represents variance explained by each local model in its correspondent local portion of the data in which the variance is obviously lower than the global variance. In other words the maximum variance which can be explained by each local model is limited by the variability contained in each local portion.

After optimization of the parameters, the models were applied to the validation set. The RMSE and the coefficient of determination (R^2) were also computed for assessing accuracy of the results.

4. Results

4.1 Main characteristics of the soil spectral libraries

For the whole R–SSL the mean values of CC, OC and Ca^{++} were 23.51%, 0.64% and 15.65 $cmol_c\ kg^{-1}$ respectively (Table 1). No significant correlations ($r<0.30$, $\rho>0.05$) between the soil attributes were observed. In all cases, the distribution of the properties is positively skewed (Figure 3). A wide range of variation was observed for CC (1% to 81%). In general, the values of OC are low, due to tropical climatic conditions in the State of Sao Paulo and the fact that samples in the R–SSL were collected in croplands. A large variation in Ca^{++} was observed, with values ranging from 1 $cmol_c\ kg^{-1}$ to 170 $cmol_c\ kg^{-1}$ and a standard deviation of 14.49 $cmol_c\ kg^{-1}$.

For the whole G–SSL a large variation of the three soil attributes was observed (Table 1, Figure 3). The mean values of soil attributes were 33.08% for CC, 1.10 for OC and 10.69 for Ca^{++}. As in the R–SSL no significant correlations (r=0.01, p> 0.05) between the soil attributes were observed and their distributions were positively skewed (Figure 3). As expected, the G–SSL presented higher variability than the R–SSL in terms of soil attributes (Table 1).

Table 1. Descriptive statistics of the soil attributes of samples in the soil spectral libraries.

Soil attribute	Mean	S.D.	C.V. (%)	Min.	1st Qu.	Median	3rd Qu.	Max.
Regional soil spectral library Whole set, N = 4200								
Clay Content (%)	23.51	12.48	53.08	1.00	14.7	20.00	28.00	81.10
Organic Carbon (%)	0.64	0.39	60.94	0.06	0.35	0.58	0.81	4.00
Exchangeable Ca ($cmol_c\ kg^{-1}$)	15.65	14.49	92.59	1.00	7.00	12.00	20.00	170.00
Regional soil spectral library Training set, n = 3150								
Clay Content (%)	23.73	12.21	51.45	1.00	15.30	20.30	28.00	68.20
Organic Carbon (%)	0.65	0.39	59.76	0.06	0.41	0.58	0.87	4.00
Exchangeable Ca ($cmol_c\ kg^{-1}$)	15.22	14.63	96.12	1.00	6.00	12.00	19.00	170.00
Regional soil spectral library Prediction set, m = 1050								
Clay Content (%)	22.86	13.24	57.92	4.80	13.92	19.00	27.40	81.10
Organic Carbon (%)	0.61	0.39	64.47	0.06	0.35	0.52	0.75	2.61
Exchangeable Ca ($cmol_c\ kg^{-1}$)	16.94	13.97	82.47	1.00	8.00	14.00	21.00	129.00
Global soil spectral library Whole set, N = 3643								
Clay Content (%)	33.08	22.49	67.99	0.00	15.00	30.50	48.10	96.80
Organic Carbon (%)	1.10	1.93	175.45	0.00	0.22	0.48	1.19	45.80
Exchangeable Ca ($cmol_c\ kg^{-1}$)	10.69	15.41	144.15	0.00	0.40	3.60	15.10	168.20
Global soil spectral library Training set, n = 2743								
Clay Content (%)	32.97	21.88	66.36	0.00	15.65	30.60	47.60	96.80
Organic Carbon (%)	1.15	2.07	180.16	0.00	0.22	0.50	1.21	45.80
Exchangeable Ca ($cmol_c\ kg^{-1}$)	11.11	15.83	142.48	0.00	0.50	3.90	16.05	168.20
Global soil spectral library Prediction set, m = 900								
Clay Content (%)	33.44	24.27	72.58	0.00	12.30	30.30	50.48	95.60
Organic Carbon (%)	0.94	1.40	148.95	0.00	0.21	0.44	1.10	15.88
Exchangeable Ca ($cmol_c\ kg^{-1}$)	9.39	13.97	148.71	0.00	0.20	2.60	12.70	67.20

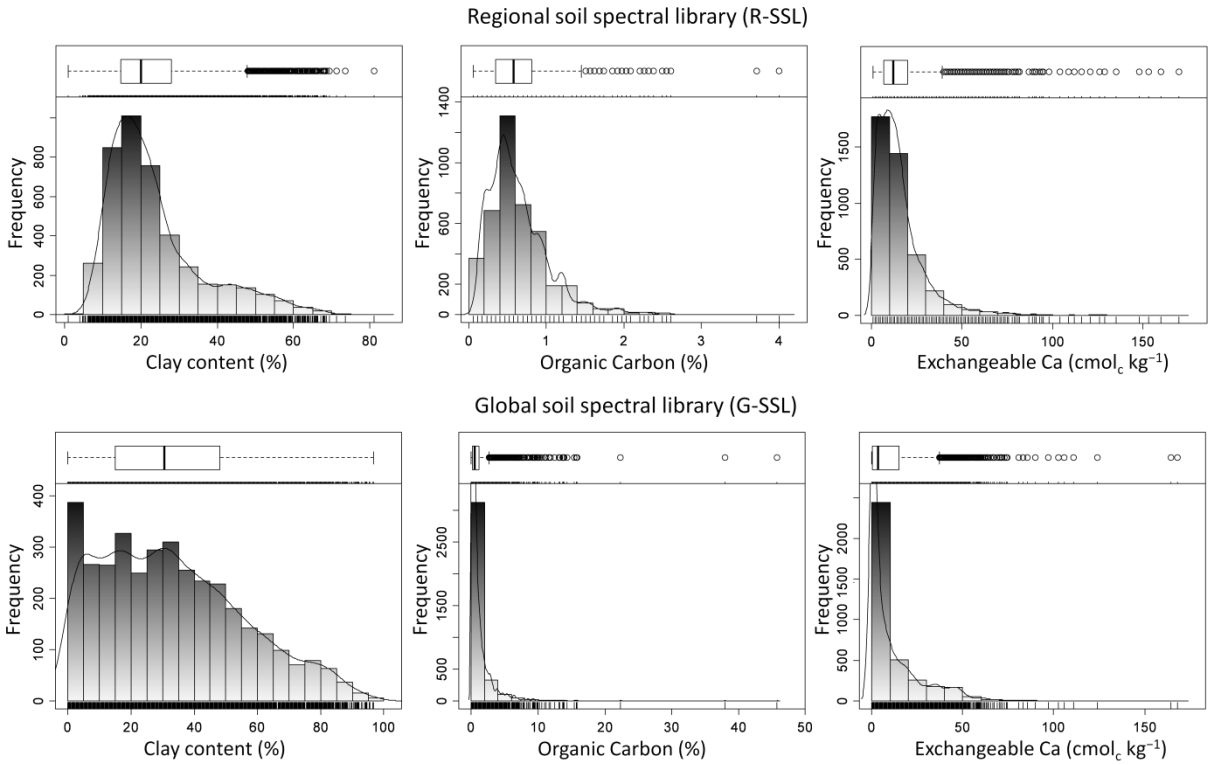

Figure 3. Histograms and box-plots of the soil attributes under study in each soil spectral library (whole sets).

In order to illustrate the soil spectral variability we selected five spectra from the R-SLL and five from the G-SLL (Figure 4). In each spectral library those spectra corresponds to: *i.* the sample with the maximum mean reflectance (or albedo), *ii.* the sample with the minimum mean reflectance, and *iii.* three samples selected by using a fuzzy k-means sampling (Bezdek, 1981).

In the R–SSL we observed a high variation of the albedo as well as of the reflectance of the mineralogical energy absorption features which in both cases is mainly due to parent material variability of the region. Overall, soils with high sand content showed higher albedo than soils with high clay content. Top layer samples (0-20 cm) showed low reflectance (especially in the 350 – 1350 nm region) due to the higher organic carbon content in comparison to the samples collected in deeper soil layers. In most of the samples we observed absorption features around 850-900 nm (related to the presence of iron oxides), 2207 nm and 2160 (both related to the kaolinite content) (Demattê *et al.*, 2004; Viscarra Rossel and Behrens, 2010). This reflects the soil composition of most of the samples in the R–SSL which are classified as Ferrasol and correspond to highly weathered

soils in general. Additional details about the R–SSL can be found in Bellinaso *et al*. (2010).

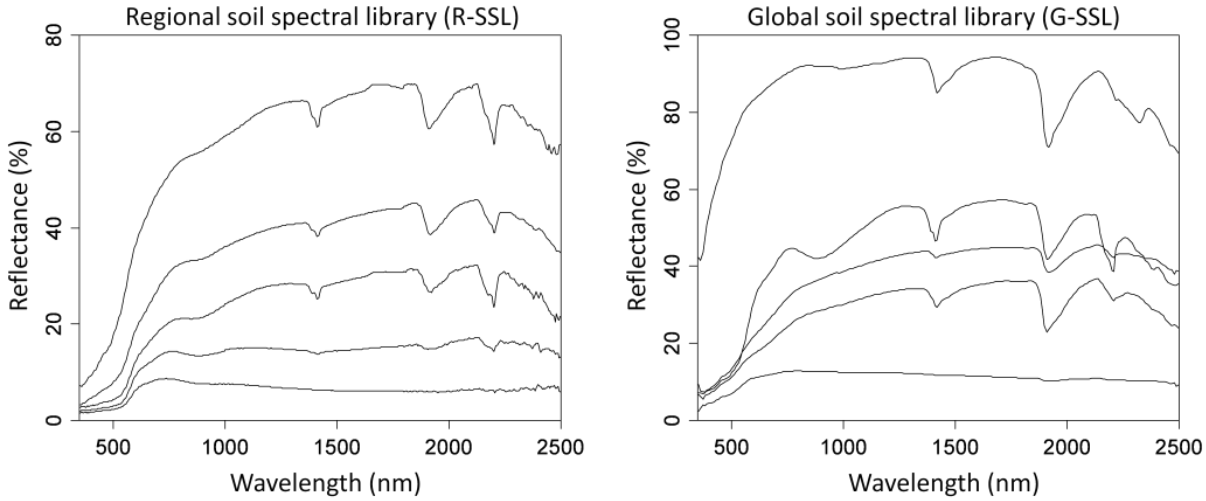

Figure 4. Five spectra from each spectral library (R–SSL on the left and G–SSL on the right). The spectra correspond to samples with the minimum and maximum mean reflectance, and three samples selected by using a fuzzy k-means sampling.

We found large spectral variation in the G–SSL as consequence of the diversity of the soil formation environments where the samples were collected (ICRAF-ISRIC, 2010). In the NIR region, the main spectral variations among samples were observed at wavelengths from 2160 nm to 2230 nm which are due to the energy absorption of minerals such kaolin and smectite (Demattê and Garcia, 1999; Demattê *et al*., 2004; Viscarra Rossel and Behrens, 2010). In general, most of the samples show well defined absorption features near to 1455 and 1915 nm, which are assigned to OH–soil hygroscopic water in clay minerals (Ben–Dor *et al*., 2008). The main spectral differences seem due to absorption bands related to iron oxides (435, 550 and 850 nm) and to kaolinite (2207 nm) (Demattê and Garcia, 1999; Demattê *et al*., 2004; Viscarra Rossel and Behrens, 2010). We observed samples with weak absorption features typical spectral of soils with high organic matter content (Stoner and Baumgardner, 1981; Ben–Dor, *et al*., 1999) and/or low levels of crystalline iron and amorphous iron (Ben–Dor *et al*., 2008).

4.2 Final parameters used for prediction

A summary of the main parameters used by the algorithms is presented in Table 2. In the R–SSL the PLS algorithm used 30, 15 and 21 PLS factors for the models of CC, OC and Ca^{++} respectively. For the calibrations carried out with the G–SSL the number of PLS factors were 28, 25 and 30 for CC, OC and Ca^{++} respectively. For SVM models in both libraries the alpha (a) hyper-parameter of the RBF was 0.01 except for the SVM model of Ca^{++} in the G–SSL which used an a value of 0.02. The C parameters of the SVM models varied from 13 (for the model of OC in the R–SSL) to 90 (for the model of CC in the G–SSL).

Table 2. Summary of the final setup of the main parameters of each algorithm in the regional soil spectral library (R–SSL) and in the global soil spectral library (G–SSL)

Algorithm	Parameter	R–SSL			G–SSL		
		CC	OC	Ca^{++}	CC	OC	Ca^{++}
PLS	Factors	30	15	21	28	25	30
SVM	a	0.01	0.01	0.01	0.01	0.01	0.02
	C	34	13	31	90	40	40
LWR [a,b]	k	380	400	370	280	320	350
LOCAL[b]	k	400	280	400	220	330	390
SBL[c]	k	360	260	140	330	310	260

[a]The number of principal components used for computing the distance matrix was 4 for the R–SSL and 10 for the G–SSL; [b]The number of PLS varied among the local models; [c]The number of principal components for computing the distance matrix was 15 for the R–SSL and 27 for the G–SSL.

For LWR models based on the R-SLL and the G–SSL, the number of principal components (PCs) used for the computations of the Mahalanobis distance (MD) matrices were 4 and 10 respectively. Figure 5 shows the number of PCs used for computing the MD matrices against the RMSD between the training samples and their correspondent most similar samples (nearest neighbors) in the training set. In the R–SSL, the lowest RMSDs was achieved with 15 PCs for CC and Ca^{++} and 14 PCs for OC. For the G–SSL the lowest RMSDs for CC, OC and Ca^{++}, were returned by distance matrices computed on 27, 28 and 27 PCs respectively. Figure 5 shows that there are only small differences in terms of RMSDs between distance matrices using 14 or 15 PCs for the R–SSL and 27 or 28 PCs for the G–

SSL. To simplify computations, we decided to use the same distance matrix computed with 15 PCs (R–SSL) and 27 PCs (G–SSL) for all the soil attributes.

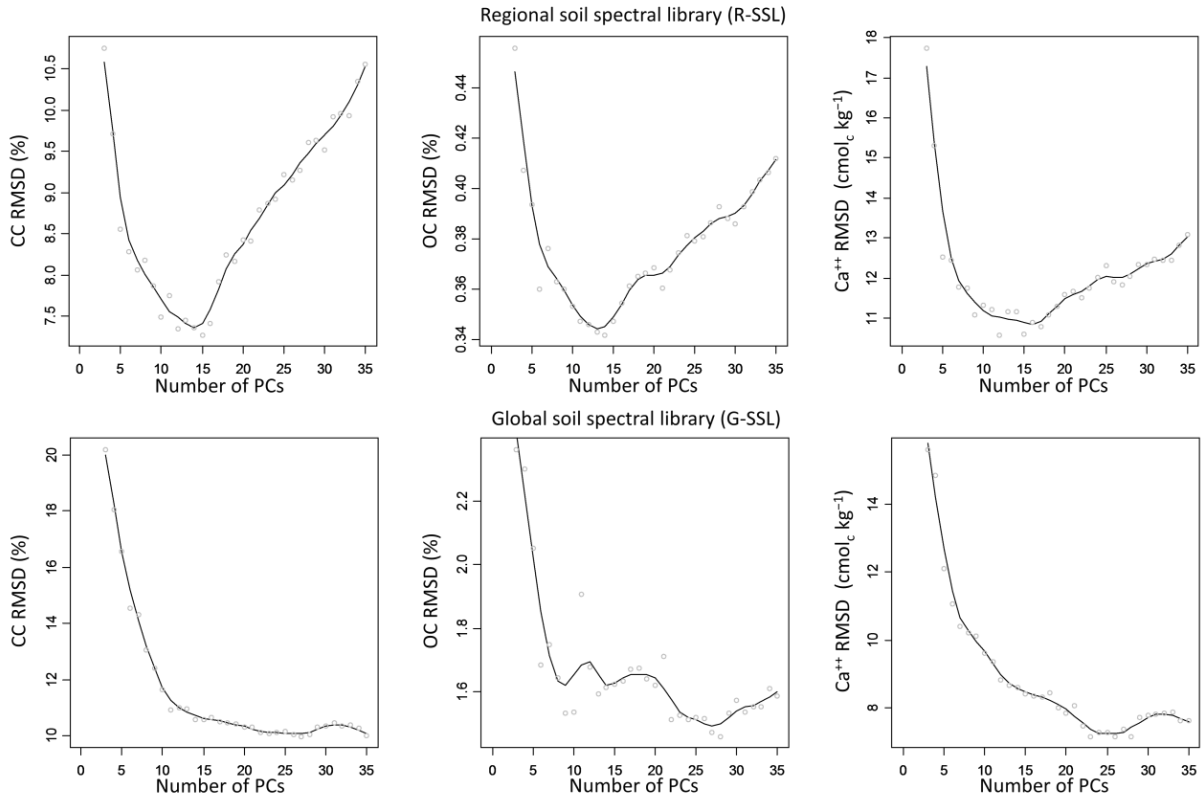

Figure 5. Number of PCs used for MD matrix computation against the root mean square of differences (RMSD) between samples in the training set and their correspondent most similar samples (nearest neighbors).

Interestingly, the oPC–M approach tends to choose a number of PCs that explain a rather large portion of the spectral variation. For instance, only the first 4 PCs for the R–SSL and the first 10 PCs for the G–SSL would have been selected if a threshold of 99.5 % of explained variance would have been used to compute the MD.

4.3 Predictive performance

Results of the predictive performance of each algorithm in both spectral libraries are presented in Table 3. Models calibrated for the R–SSL yielded lower errors than models calibrated for the G–SSL. This can be explained in terms of the di-

versity of the environmental conditions (particularly parent material) of the geographical origin of samples in the spectral libraries.

Table 3. Prediction results of the global and memory–based learning approaches.

Algorithm	CC			OC			Ca^{++}		
	RMSE $_{tr}$[a]	RMSE $_{pred}$[b]	R^2 $_{pred}$[b]	RMSE $_{tr}$[a]	RMSE $_{pred}$[b]	R^2 $_{pred}$[b]	RMSE $_{tr}$[a]	RMSE $_{pred}$[b]	R^2 $_{pred}$[b]
Regional soil spectral library (R–SSL)									
PLS*	6.47	6.10	0.78	0.28	0.28	0.48	9.60	9.13	0.59
SVM*	6.33	6.40	0.78	0.28	0.27	0.54	9.57	9.84	0.54
LWR**	5.64	6.15	0.79	0.27	0.28	0.48	8.53	9.11	0.62
LOCAL**	5.39	5.90	0.81	0.26	0.27	0.53	7.82	8.15	0.67
SBL**	**5.31**	**5.18**	**0.85**	**0.25**	**0.25**	**0.59**	**7.52**	**7.90**	**0.70**
Global soil spectral library (G–SSL)									
PLS*	12.44	12.95	0.71	1.35	1.08	0.50	10.30	9.79	0.51
SVM*	10.80	13.84	0.68	1.27	0.93	0.57	10.24	10.36	0.47
LWR**	10.72	12.81	0.73	0.82	1.02	0.56	8.97	10.94	0.49
LOCAL**	12.19	12.98	0.72	0.94	1.02	0.55	8.83	9.83	0.53
SBL**	**7.97**	**12.01**	**0.77**	**0.79**	**0.80**	**0.68**	**6.93**	**8.48**	**0.63**

*Global learning approaches; **Memory–based learning approaches; numbers in bold indicate the best results; [a]$RMSE_{tr}$ indicates the training RMSE; [b]$R^2_{pred.}$ and $RMSE_{pred.}$ indicate the R^2 and RMSE of the validation set respectively.

Overall, SBL outperformed the global calibration models (PLS and SVM) and the other memory–based learning approaches (LWR and LOCAL) in both spectral libraries (Table 3). In all cases the SBL produced the lowest training and prediction RMSEs ($RMSE_{tr}$ and $RMSE_{pred}$ in Table 3), as well as the highest prediction R^2 (R^2_{pred}) (Table 3). In terms of the $RMSE_{pred}$ the highest variability between results in the R–SSL were found for Ca^{++} predictions followed by CC predictions. The OC presented the lowest variability between the $RMSEs_{pred}$ produced by the regression algorithms. This is attributed to the fact that the OC presented a very low variation in the R–SSL. In the G–SSL the highest variability in terms of $RMSE_{pred}$ was observed for the OC predictions. In this case this can be related with the high OC variability observed in the G–SSL. For the Ca^{++}, the $RMSEs_{pred}$ producded by the different algorithms were also highly variable, while the variability of the $RMSEs_{pred}$ of the CC models was low. In summary, we observed a

trend in which the variability of the $RMSEs_{pred}$ produced by the algorithms increases with the variability of the soil attribute. Comparing the prediction errors of CC in both libraries we found that the differences between algorithms is much higher in the R–SSL than in the G–SSL. This can be probably due to the influence of OC in the spectral variability. We believe that the vis–NIR spectral variability of the R–SSL presents a low influence of the OC due to the fact that it presents a very low variation. This probably facilitates the modeling process of the CC since the "interference" of the OC is low. On the other hand, in the G–SSL the OC presents a large variability which influences the variability of the vis–NIR spectra and therefore can affect the modeling process of the CC.

For CC in both spectral libraries, the highest prediction errors were produced by the SVM models ($RMSE_{pred}$=6.40 % for the R–SSL and $RMSE_{pred}$=13.84% for the G–SSL), closely followed by the PLS models with $RMSEs_{pred}$ of 6.10 % and 12. 95% for the R–SSL and the G–SSL respectively. The error produced by the SBL model of OC in the R–SSL ($RMSE_{pred.}$=0.25%) was slightly lower than the error returned by the other algorithms. The R^2_{pred} of the SBL was higher than those returned by other algorithms. In all cases the R^2 of the soil OC predictions in the R–SSL were low (<0.60), which also is attributed to the small OC variation in that spectral library (this also explains why the low training and prediction errors of OC in this library). For the G–SSL, the $RMSEs_{pred}$ of the PLS ($RMSE_{pred}$=1.08), SVM ($RMSE_{pred}$=0.93%), LWR ($RMSE_{pred}$=1.02%) and LOCAL ($RMSE_{pred}$=1.02%) were similar and in all cases much higher than the $RMSE_{pred}$ obtained with the SBL model (0.80%) (Table 3).

Models of Ca^{++} calibrated with PLS, SVM and LWR produced similar prediction results with $RMSEs_{pred.}$ ranging from 9.11 to 9.84 $cmol_c\ kg^{-1}$ and 9.79 to 10.94 $cmol_c\ kg^{-1}$ for the R–SSL and the G–SSL respectively. For the R–SSL the error of LOCAL predictions of Ca^{++} ($RMSEs_{pred.}$= 8.15 $cmol_c\ kg^{-1}$) was slightly higher than the error returned by the SBL ($RMSEs_{pred.}$= 7.90 $cmol_c\ kg^{-1}$). For the G–SSL, we found much larger differences between LOCAL and SBL in terms of prediction error (LOCAL $RMSEs_{pred.}$= 9.83 $cmol_c\ kg^{-1}$; SBL $RMSEs_{pred.}$= 8.48 $cmol_c\ kg^{-1}$)

For the results obtained in the G–SLL we found that the SBL can produce competitive results in comparison with other approaches applied in global soil spectral libraries reported in the literature. For example, Brown *et al.*, (2006) calibrated models of OC using vis–NIR data in combination with sand content and pH as auxiliary predictors. They found an RMSE of 0.79% which is comparable to the one obtained by the SBL in the G–SSL, however in contrast to their approach, the SBL did not use other soil attributes as auxiliary predictors. Terhoeven-Urselmans *et al.* (2010) used a global soil mid infrared soil spectral library for modeling several soil attributes. In their validations they found RMSEs values of 12.6%, 0.90% and 10.2 $cmol_c\ kg^{-1}$ for CC, OC and Ca^{++} respectively. In case of CC the results are comparable to our SBL results, however for OC and Ca^{++} we obtained lower RMSEs in the G–SSL.

5. Discussion

We found that SBL produces more accurate results than the rest of the algorithms tested. It results from the combination of two important characteristics of the SBL: *i.* a more appropriate neighbor selection is carried out by using the distance matrix computed with the oPC–M method, and *ii.* the inclusion in each local model of a $k \times k$ distance matrix as a source of additional predictor variables. Furthermore, the use of a GP algorithm with linear covariance function for local modeling seems to be a suitable regression approach.

One interesting characteristic which is proposed in SBL is the use of the distance matrix as a source of additional predictor variables. In other words, the SBL approach derives from the spectra additional predictive information that is not exploited in any of the other algorithms. For example, for the SBL model of CC in the R–SSL a number of 360 neighbors was used. This means that each local model of CC was calibrated with 360 new features or predictor variables in addition to the spectral features.

We assumed that the more similar two samples are in terms of their vis–NIR spectra, the more similar they could be in terms of soil compositional characteristics. This means that in a given set of samples, the variability of a soil attribute could be explained in part by the variability of the spectral similarity/dissimilarity scores with respect to a reference point (spectrum).

Each sample in the neighborhood is used as a reference point within the same neighborhood. The similarity/dissimilarity between the reference point and all the samples is estimated. Each new similarity/dissimilarity variable (or column of the distance matrix) represents new information about the position of the samples in the multivariate space. The exact position of the target sample within the neighborhood is known since the number of reference points is equal to the number of neighbor samples.

We think that the information about the position of the target sample in the neighborhood contains some information about the variability of the samples which cannot be easily captured by the regression algorithm when it is applied only on the vis–NIR variables. Zerzucha *et al*. (2012) showed that non-linear modeling problems can be resolved by simply applying partial least square regression on distance matrices. In general they concluded that the predictor variables based on distance measurements can increase the predictive power of models in complex datasets. In soil science, some works have shown that soil distances can be very useful for modeling soil variation. Kriging is a clear example of the application of distance matrices in the prediction of soil properties. In digital soil mapping, Minasny and McBratney (2007) suggested the use of soil taxonomic distances as source of predictor variables for soil classes. Furthermore, Carre and Jacobson *et al*. (2009) observed that pedological distances can be used as predictors for modeling soil available water capacity with high prediction performance. We think that the incorporation of soil distances in soil modeling is very promising approach which requires more research.

The neighbor selection for local models is an important step which can be viewed as a way to clean each local partition, i.e. noisy samples or samples which have low predictive information are ignored. In this respect the use of a reliable dis-

tance matrix can be critical for appropriate neighbor selections in local modeling (Gogé *et al.* 2012). Here we showed that in most of the cases the SBL used a low number of nearest neighbors (*k*) than LWR and LOCAL (Table 2) suggesting that the distance matrix computed in the SBL is more appropriate to find relevant neighbors for soil predictions.

In order to gain some insights about the importance of the oPC–M distance in the SBL approach we implemented the methodological framework carried out by the SBL with three different variations. In the first variation (Local GPL *a*) the oPC–M distance is used for neighbor selection and the local distance matrices are not used as source of additional predictors. The second variation (Local GPL *b*) follows the same methodological framework of the SBL approach but instead using the oPC–M distance it uses the standard principal component distance, this means that it uses the same distance matrix as in the LWR approach. The third variation (Local GPL *c*) also follows the methodological framework of the SBL but in this case the correlation distance is used, i.e. the same distance matrix used in LOCAL. Soil attribute predictions with these three Local GPL were carried out for both spectral libraries. The number of nearest neighbors was fixed to the same number of neighbors used for each soil attribute in the SBL method (see Table 2). In comparison to the prediction results found for the SBL (Table 3) we found that when the local oPC–M distance matrices are not used as source of additional predictors the accuracy is lower (Table 4). When the standard principal component distance or the correlation distance are used for both neighbor selection and source of additional predictors, the prediction performance is also lower than in the SBL approach (Table 4). For CC predictions, in the R–SSL the results returned by the three Local GPL were very similar while in the G–SSL the Local GPL *a* outperformed both, the Local GPL *b* and the Local GPL *c*. These results indicate that for local modeling, the oPC–M distance could perform better than the standard principal component distance and the correlation distance.

Table 4. Prediction results using different local Gaussian process regression strategies. The number of neighbors used for modeling each soil attribute in each spectral library was the same as in the SBL approach (see Table 2).

Algorithm	CC		OC		Ca^{++}	
	$RMSE_{pred.}$	$R^2_{pred.}$	$RMSE_{pred.}$	$R^2_{pred.}$	$RMSE_{pred.}$	$R^2_{pred.}$
Regional soil spectral library (R–SSL)						
Local GPL *a**	5.65	0.82	**0.25**	**0.58**	**8.11**	**0.69**
Local GPL *b***	**5.60**	0.82	0.27	0.54	8.94	0.61
Local GPL *c****	5.66	0.82	0.26	0.56	8.22	0.66
Global soil spectral library (G–SSL)						
Local GPL *a**	**12.36**	**0.75**	**0.89**	**0.61**	**8.80**	**0.61**
Local GPL *b***	12.87	0.72	0.97	0.59	9.22	0.57
Local GPL *c****	13.02	0.72	1.00	0.54	9.01	0.59

Numbers in bold indicate the best results;
*Local GPL *a*: local Gaussian process regressions using the oPC–M for neighbor selection and excluding the local distance matrices as additional predictors.
**Local GPL *b*: local Gaussian process regressions using the standard principal component distance for neighbor selection and local distance matrices as additional predictors.
***Local GPL *c*: local Gaussian process regressions using the correlation distance for neighbor selection and local distance matrices as additional predictors.

In terms of computational cost, in LWR and LOCAL the number of PLS factors in each local model needs to be optimized and this represents one important drawback. The SBL is more efficient than LWR and LOCAL regarding the optimization of *k*. This is due to the fact that the SBL (unlike LWR and LOCAL) does not require any internal optimization in each local model and for cross validation only the most similar samples (nearest neighbor) of the prediction set found in the training set are used. Despite SBL uses much more predictor variables than LWR and LOCAL, we found that the computational cost of the SBL is lower. For example for optimizing *k* for models of both models (using the grid of 30 to 400 *k* by steps of 10), SBL was 2.21 and 2.24 times faster than LWR and LOCAL respectively.

In general, despite LWR, LOCAL and SBL belong to the memory–based learning methods, the LWR and LOCAL did not showed clear evidence to perform better than the global models. This is mainly attributed to the distance matrices used which does not represent correctly the similarity between samples and therefore can fail in the neighbor selection. However we showed through the SBL results

that when appropriate similarity measurements are performed, memory–based learning is a reasonable approach for modeling soil attributes in complex vis–NIR datasets. Memory based learners such our SBL are very flexible approaches since they offer the possibility to combine other approaches. For example in each partition a local outlier analysis can be carried out in order to obtain a better identification of those samples. Another possibility is to calculate the prediction intervals for each partition, and then identifying the samples outside those ranges.

Soil spectroscopy research should be focused on bridging the gap between modeling algorithms and theories of the interactions between soil components and electromagnetic radiation. With the development of the SBL algorithm, we attempt to stimulate the use of memory–based learning which represents a straightforward strategy for integrating theory and algorithms.

It is worth to mention that in extrapolation cases (i.e. when samples to be predicted are far away from the spectral library), the SBL (as any other regression algorithm) would be prone to fail in producing reliable model soil predictions. In this sense, Shepherd and Walsh (2002) propose a very simple (non-modeling) solution to this problem: take the samples which are far away from the spectral library and perform soil routine analyses of the target soil attributes, then include those samples in the spectral library (which means library improvement).

6. Conclusions

Two important goals in this research were to introduce a new high performance memory–based learning approach and also to point out the importance of studying local modeling approaches for modeling soil attributes using visible and near infrared spectral libraries.

The main conclusions derived from this research are: *i.* our SBL is a new and reliable approach which returned the best prediction results (lowest RMSEs and highest R^2s) for both spectral libraries, in comparison to the global calibration models (PLS and SVM) and the other memory–based learning approaches (LWR and LOCAL); *ii.* the oPC–M distance matrices (used in the SBL approach) proba-

bly represent better the compositional similarity between samples than the conventional principal component distance matrices (used in the LWR approach). This indicates that probably the SBL performs a better neighbor selection than LWR; *ii.* the use of local oPC–M distance matrices as source of additional predictive variables do not degrade the prediction performance, instead it can result in an increment of it. *iii.* in terms of computational time for optimizing the number of nearest neighbors, the SBL is more efficient than both LWR and LOCAL; and *iv.* both LWR and LOCAL did not show clear evidence to perform better than the global calibration algorithms (PLS and SVM).

Acknowledgments

We wish to thank Henrique Bellinaso and Suzana Romeiro who spent an enormous amount of time working on the implementation the soil vis–NIR library of State of São Paulo. We also thank ICRAF and ISRIC for making the global soil vis–NIR library available; it is a very nice present to the soil science community.

References

Aizerman, M., Braverman, E., Rozonoer, L. 1964. Theoretical foundations of the potential function method in pattern recognition learning. Automation and Remote Control 25, 821–837.

Akbari, M., van Overloop, P.J., Afshar, A. 2011. Clustered K Nearest Neighbor Algorithm for Daily Inflow Forecasting. Water Resources Management 25, 1341-1357.

An, A. Classification Methods. In J. Wang (Ed.), Encyclopedia of Data Warehousing and Mining, Idea Group Inc., 2005. 144-149.

Bellinaso, H., Demattê, J.A.M., Romeiro, S.A. 2010. Soil spectral library and its use in soil classification. Revista Brasileira de Ciencia Do Solo 34, 861–870.

Ben–Dor, E., J.R. Irons, and G.F. Epema. 1999. Soil reflectance. p. 111–188. In N. Rencz (ed.) Remote sensing for the earth sciences: Manual of remote sensing. Vol. 3. John Wiley & Sons, New York.

Ben-Dor, E., Taylor, R.G., Hill, J., Demattê, J.A.M., Whiting, M.L., Chabrillat, S., Sommer, S. 2008. Imaging spectrometry for soil applications. Advances in Agronomy 97, 321–392.

Bezdek, J.C. 1981. Pattern Recognition with Fuzzy Objective Function Algorithms, Plenum, NY.

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

Camargo, A.O., Moniz, A.C., Jorge, J.A., Valadares, J.M. 2009. Métodos de análise química, mineralógica e física de solos do IAC. Campinas, Instituto Agronômico de Campinas, 77p. (Boletim Técnico, 106).

Carré, F., Jacobson, M. 1999. Numerical classification of soil profile data using distance metrics. Geoderma 148, p. 336–345.

Chang, C.W., Laird, D.A., Mausbach, M.J. & Hurburgh, C.R. 2001. Near-infrared reflectance spectroscopy-principal components regression analyses of soil properties. Soil Science Society of America Journal, 65, 480-490.

Chen, T., Morris, J., Martin, E. 2007. Gaussian process regression for multivariate spectroscopic calibration, Chemometrics and Intelligent Laboratory Systems 87, 59–71.

Christy, C.D., Dyer, S.A. 2006. Estimation of soil properties using a combination of spectral and scalar sensor data. Conference Record - IEEE Instrumentation and Measurement Technology Conference , art. no. 1700262, pp. 729-734.

Davies, A.M.C., Fearn, T. 2006. Quantitative analysis via near infrared databases: Comparison analysis using restructured near infrared and constituent data-deux (CARNAC-D). Journal of Near Infrared Spectroscopy 14, 403–411.

De Maesschalck, R., Jouan-Rimbaud, D., Massart, D.L. 2000. The Mahalanobis distance. Chemometrics and Intelligent Laboratory Systems 50, 1-18.

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

Demattê, J.A.M., Garcia, G.J. 1999. Alteration of soil properties through a weathering sequence as evaluated by spectral reflectance. Soil Science Society of America Journal 63, 327-342.

Drucker, H., C. J. C. Burges, L. Kaufman, A. Smola, and V. Vapnik. 1996. Support vector regression machines. In: Advances in Neural Information Processing Systems (NIPS), Vol. 10. pp. 155–161.

Fearn, T., Davies, A.M.C. 2003. Locally-biased regression. Journal of Near Infrared Spectroscopy 11, 467–478.

Fernández Pierna, J.A., Dardenne, P. 2008. Soil parameter quantification by NIRS as a Chemometric challenge at 'Chimiométrie 2006'. Chemometrics and Intelligent Laboratory Systems 91, 94–98.

Genot, V., Colinet, G., Bock, L., Vanvyve, D., Reusen, Y., Dardenne, P. 2011. Near infrared reflectance spectroscopy for estimating soil characteristics valuable in the diagnosis of soil fertility. Journal of Near Infrared Spectroscopy 19, 117–138.

Gneiting, T., Sasvári, Z., Schlather, M. 2001. Analogies and correspondences between variograms and covariance functions. Advances in Applied Probability 33, 617-630.

Gogé, F., Joffre, R., Jolivet, C., Ross, I., Ranjard, L. 2012. Optimization criteria in sample selection step of local regression for quantitative analysis of large soil NIRS database. Chemometrics and Intelligent Laboratory Systems 110, 168–176.

Guerrero, C., Zornoza, R., Gómez, I., Mataix–Beneyto, J. 2010. Spiking of NIR regional models using samples from target sites: Effect of model size on prediction accuracy. Geoderma 158, 66–77.

Heanes, D.L. 1984. Determination of total organic-C in soils by an improved chromic acid digestion and spectrophotometric procedure. Communications in Soil Science and Plant Analysis 15, 1191–1213.

Igne, B., Reeves III, J.B., McCarty, G., Hively, W.D., Lund, E., Hurburgh, C.R. 2010. Evaluation of spectral pretreatments, partial least squares, least squares support vector machines and locally weighted regression for quantitative spectroscopic analysis of soils. Journal of Near Infrared Spectroscopy 18, 167–176.

IUSS Working Group WRB 2006. World Reference Base for Soil Resources 2006. 2nd edition. Rome FAO, World Soil Resources Reports No. 103.

Janik, L.J., Skjemstad, J.O., Shepherd, K.D., Spouncer, L.R., 2007. The prediction of soil carbon fractions using mid-infrared-partial least-square analysis. Australian journal of soil research 45, 73–81.

Kang, P., Cho, S. 2008. Locally linear reconstruction for instance-based learning. Pattern Recognition 41, 3507–3518.

Lopez de Mantaras, R., McSherry, D., Bridge, D., Leake, D., Smyth, B., Craw, S., Faltings, B., Maher, M.L., Cox, M., Forbus, K., Keane, M., Aamodt, A., and Watson, I. 2006. Retrieval, Reuse, Revision and Retention in Case-Based Reasoning. The Knowledge Engineering Review 20, 215–240.

Madejova, J., 2003. FTIR techniques in clay mineral studies. Vibrational Spectroscopy 31, 1–10.

McBratney, A. B., Hart, G. A., McGarry, D. 1991. The use of region partitioning to improve the representation of geostatistically mapped soil attributes. Journal of Soil Science 42, 513–532.

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

Meyer, C.R., Flanagan, D.C. 1992. Application of case-based reasoning concepts to the WEPP soil erosion model. AI Applications in Natural Resource Management 6 (3) , pp. 63-71

Minasny, B., McBratney, A.B. 2007. Incorporating taxonomic distance into spatial prediction and digital mapping of soil classes. Geoderma 142, 285–293.

Mitchell, TM. Machine Learning. McGraw-Hill, New York, 1997.

Naes, T., Isaksson, T., Kowalski, B. 1990. Locally weighted regression and scatter correction for nearinfrared reflectance data. Analytical Chemistry 62, 664–673

Naes, T.; Isaksson, T.; Fearn, T.; Davies T. 2002. A User-friendly guide to multivariate calibration and classification. NIR Publications, Chichester, UK. 420 p.

Oliveira, J.B.; Camargo, M.N.; Rossi, M.; Calderano Filho, B. 1999. Mapa pedológico do Estado de São Paulo. Campinas: Instituto Agronômico, escala 1:500.000.

Ostfeld, A., Salomons, S. 2005. A hybrid genetic - Instance based learning algorithm for CE-QUAL-W2 calibration. Journal of Hydrology 310, 122-142.

Qi, F., Zhu, A.-X., Harrower, M., Burt, J.E. 2006. Fuzzy soil mapping based on prototype category theory. Geoderma 136, 774-787.

R Development Core Team. 2011. R: A language and environment for statistical computing. Vienna: R Foundation for Statistical Computing.

Ramirez-Lopez, L., Behrens, T., Schmidt, K., Viscarra Rossel, R., Demattê, J.A.M., Scholten, T. Distance and similarity-search metrics for use with soil vis–NIR spectra, Geoderma (2012), http://dx.doi.org/10.1016/j.geoderma.2012.08.035

Rasmussen, C.E., Williams, C.K. Gaussian Processes for Machine Learning. Massachusetts Institute of Technology: MIT-Press, 2006.

Savvides,A., Corstanje, R., Baxter, S.J., Rawlins, B.G. Lark, R.M. 2010. The relationship between diffuse spectral reflectance of the soil and its cation exchange capacity is scale-dependent. Geoderma 154, 353–358.

Schmidt, K., Behrens, T., Friedrich, K., Scholten, T. 2010. A method to generate soilscapes from soil maps. Journal of Plant Nutrition and Soil Science 173, 163–172.

Schölkopf, B., Smola. A. 2002. Learning with Kernels. MIT Press, Cambridge, MA.

See, L. 2008. Data fusion methods for integrating data-driven hydrological models. Studies in Computational Intelligence 79, 1-18.

Shenk, J.S., Westerhaus, M.O., Berzaghi, P. 1997. Investigation of a LOCAL calibration procedure for near infrared instruments. Journal of Near Infrared Spectroscopy 5, 223–232.

Shi, X., Zhu, A-X., Burt, J.E., Qi, F., Simonson, D. 2004. A Case–based Reasoning Approach to Fuzzy Soil Mapping. Soil Science Society of America Journal 68, 885–894.

Shi, X., Long, R., Dekett, R., Philippe, J. 2009. Integrating different types of knowledge for digital soil mapping. Soil Science Society of America Journal 73, 1682-1692

Shepherd, K.D., Palm, C.A., Gachengo, C.N., Vanlauwe, B., 2003. Rapid characterization of organic resource quality for soil and livestock management in tropical agroecosystems using near-infrared spectroscopy. Agronomy Journal 95, 1314-1322.

Shepherd, K,D., Walsh, M.G. 2002. Development of Reflectance Spectral Libraries for Characterization of Soil Properties. Soil Science Society of America Journal 66, 988-998.

Solomatine, D.P., Maskey, M., Shrestha, D.L. 2008. Instance-based learning compared to other data-driven methods in hydrological forecasting. Hydrological Processes 22, 275–287.

Stenberg, B., Viscarra Rossel, R.A., Mouazen, A.M., Wetterlind, J. Visible and Near Infrared Spectroscopy in Soil Science. In: Donald L. Sparks, Ed. Advances in Agronomy, Vol. 107, Burlington: Academic Press, 2010, pp. 163–215.

Stevens, A., Udelhoven, T., Denis, A., Tychon, B., Lioy, R.,Hoffmann, L., van Wesemael, B. 2010. Measuring soil organic carbon in croplands at regional scale using airborne imaging spectroscopy. Geoderma 158, 32–45.

Stoner, E.R., Baumgardner, M.F. 1981. Characteristic variations in reflectance of surface soils. Soil Science Society of America Journal 45, 1161–1165.

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, 1792–1799.

Van Reeuwijk L.P. (ed.), 2002. Procedures for Soil Analysis. 6th Edition. International Soil Reference and Information Centre, Wageningen, and Food and Agricultural Organization of the United Nations, Rome.

Viscarra Rossel, R., and Behrens, T. 2010. Using data mining to model and interpret soil diffuse reflectance spectra. Geoderma 158, 46–54.

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

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

Wetterlind, J., Stenberg, B. 2010. Near–infrared spectroscopy for within–field soil characterization: Small local calibrations compared with national libraries spiked with local samples. European Journal of Soil Science 61, 823–843.

Williams, C. K. I. and Rasmussen, C. E. Gaussian processes for regression. In: D. S. Touretzky, M. C. Mozer, and M. E. Hasselmo, editors, Advances in Neural Information Processing Systems 8, pages 514-520. The MIT Press, Cambridge, MA, 1996.

Wold, S., Martens, H., Wold, H., 1983. The multivariate calibration method in chemistry solved by the PLS method. In: Ruhe, A., Kagstrom, B. (Eds.), Proc. Conf. Matrix Pencils, Lecture Notes in Mathematics. Springer-Verlag, Heidelberg, pp. 286–293.

World Agroforestry Centre (ICRAF) and ISRIC – World Soil Information. 2010. ICRAF–ISRIC Soil vis–NIR spectral Library. Nairobi, Kenya: World Agroforestry Centre (ICRAF).

Zerzucha, P., Daszykowski, M., Walczak, B. 2012. Dissimilarity partial least squares applied to non-linear modeling problems, Chemometrics and Intelligent Laboratory Systems 110, 156–162.

Zhu, A-X., Liu, J. Soil property mapping over large areas using sparse ad-hoc samples. American Geophysical Union, Fall Meeting 2011.

Acknowledgments

Being in Germany and Belgium during these three years of doctoral research was a great experience from which I learned that uncertainty, complexity and difficulties brings along challenges and opportunities. I believe this work would not have been possible without the outstanding company and support of many people (my *k*-nearest neighbors) and as well as without the numerous challenges I faced during this time.

Firstly, I would like to express my sincere gratitude to Prof. Dr. Thomas Scholten for receiving me in his research group, for the constructive and critical discussions and for all the guidance and support during my doctoral work.

I am deeply grateful to Dr. Thorsten Behrens. All his invaluable ideas, support, guidance and challenging questions during all these years contributed substantially to the development of this thesis. I also would like to give a very special thank you to Dr. Karsten Schmidt who always was willing to help me and to discuss with me on thousands of questions. I admire the innovative work of these two pedometricians who showed me their "voodoo magic" and taught me a lot on how to dig for knowledge.

The first two years of my doctoral work were founded by the iSOIL project (Interactions between soil related sciences – Linking geophysics, soil science and digital soil mapping, Grant Agreement No. 211386) which was co-funded by the Research DG of the European Commission within the RTD activities of the FP7 Thematic Priority Environment. Once again I express my gratitude to Dr. Thorsten Behrens and Prof. Dr. Thomas Scholten for offering me the possibility to work within this project.

My sincere appreciation and gratitude goes to Prof. Dr. Bas van Wesemael who received me in his research team at the Université Catholique de Louvain (Belgium) during the last year of my doctoral work. His support and patience were essential for finishing this thesis. I also enjoyed discussing with him and learning from him during the endless field campaigns we carried out in Luxembourg. During this last year in Belgium I was also lucky to work with Dr. Antoine Stevens.

Our endless and very interesting discussions on soil spectroscopy and programming influenced large part of this thesis.

I am grateful to Prof. Dr. Jose Dematte for all his academic support and for introducing me to the field of soil sensing when I started my master studies. I also thank my colleague Jonas Daummann who helped me a lot when I arrived in Germany. I owe many thanks to Ann-Kathrin Schatz, Dr. Karsten Schmidt, Dr. Thosten Behrens and Felix Stumpf who helped me with one of the hardest parts of this thesis (the "Zusammenfassung"). I am grateful to: Amparo, Jesus, Leonardo, Mauricio, Nelson, Monica, Anibal, Carolina, Joel, Richi, Geraldine, Marcelo, Luisa, Felipe, Diana and Angela for their sincere friendship.

I will always be in debt to all my beloved family: Elvia, Celiano, Diana, Martha Lucia, Mafe, Lucia, Patricia, Eunice and Rebeca. I thank them for being there whenever I needed it. I am also very grateful to Eva, Uli, Helmut, Elke and Sebastian for opening your homes and making me feel part of the family.

Last but not least, I would like to express my deepest and sincere gratitude to Katrin for all her love, constant support, patience and understanding during the development of this thesis.

Curriculum vitae

Leonardo Ramirez-Lopez, M.Agr.
Nationality: Colombia
Date and place of birth: 24.11.1980. Fusagasugá (Colombia),

Education

2009 –	Doctoral student University of Tübingen. Faculty of Mathematics and Natural Sciences. Soil Science and Geomorphology.
2007- 2009	Master in agronomy (Soil Science) University of São Paulo. Escola Superior de Agricultura "Luiz de Queiroz" Brazil Title of M.Sc. Dissertation: Quantittive pedology: VIS–NIR -SWIR spectrometry and digital soil mapping (in Portuguese)
2001 – 2006	Agronomic engineering (5 years degree) University of Cundinamarca, Colombia Undergraduate project: Study of spatial variability of some soil physical attributes in Colombian east floodplains.

Professional experience

1. Swiss Federal Institute for Forest, Snow and Landscape Research WSL, ETH Domain (Switzerland)

2013 –	Researcher Forest soils and biogeochemistry unit. Project: Digital hydropedological mapping in Swiss forests.

2. Université Catholique de Louvain (Belgium)

2011 – 2012	Researcher The Georges Lemaître Centre for Earth and Climate Research Project: SOC 3D: Three dimensional soil organic carbon (SOC) monitoring using VNIR reflectance spectroscopic techniques

3. University of Tübingen – iSoil Project (Germany)

2009 – Doctoral student
Chair of Soil Science and Geomorphology.

Project:
iSOIL project: Interactions between soil related sciences – Linking geophysics, soil science and digital soil mapping.

4. University of São Paulo (USP) - The State of São Paulo Research Foundation (FAPESP) (Brazil)

2007 – 2009 M.Sc. Student/Researcher
Soil Science Department.
Laboratory of remote sensing and geoprocessing applied to soils and land use planning.

Projects:
Digital soil assessment for improving soil management in sugarcane plantations.
Global soil spectral library (Project from the soil spectroscopy group, South America).

5. El Palmar del Llano S.A. (Colombia) (oil palm plantation)

2007 Assistant soil surveyor
Agronomy department

6. Colombian Oil Palm Research Center – cenipalma (Colombia)

2006 Undergraduate research intern
Soil and water management division

Main research areas of interest

1 Pedometrics, (spatial) soil data mining and machine learning
2 Soil sensing
3 Digital soil mapping
4 Quantitative analysis of soil-landscape formation
5 Geomorphometry
6 Chemometrics

Awards

2008 The State of São Paulo Research Foundation Scholarship for Master Studies

2003 Best undergraduate academic performance award during the first semester of 2003. Agronomy program, University of Cundinamarca

Invited talks and courses

Training course on reflectance spectroscopy and multivariate calibration in R. Carried out at: Natural Resources Management Cluster, Centre for Development and Environment, University of Bern. Switzerland. (Dates in which the course will be carried out: 12- 14th November, 2012). Lecturers: Leonardo Ramirez-Lopez and Antoine Stevens.

Proximal soil sensing and digital soil mapping. At: V international lectures on engineering. National University of Colombia. Colombia, 2011. Lecturer: Leonardo Ramirez-Lopez

Pedometrics and proximal soil infrared spectroscopy. At: Scientific seminars. University of Cundinaraca. Colombia, 2011. Speaker: Leonardo Ramirez-Lopez

Peer reviews

Geoderma 2011-2012 (2 papers)
International Journal of Remote Sensing 2011-2012 (2 papers)

Articles in scientific journals

Ramirez-Lopez, L.; Behrens, T.; Schmidt, K.; Stevens, A.; Demattê, J.A.M.; Scholten. 2013. The spectrum-based learner: a new local approach for modeling soil vis–NIR spectra of complex datasets. Geoderma 195-196, 268-279.

Ramirez-Lopez, L.; Behrens, T.; Schmidt, K.; Viscarra Rossel, R.; Demattê, J.A.M., Scholten, T. Distance and similarity-search metrics for use with soil vis–NIR spectra. Geoderma (Special issue on proximal soil sensing). doi: 10.1016/j.geoderma.2012.08.035. Accepted on August 2012.

Articles under review

Ramirez-Lopez, L.; Demattê, J.A.M.; Schmidt, K. Behrens, T.; van Wesemael, B.; Scholten, T. Calibration sampling and calibration set size for soil vis–NIR modeling and mapping. (Submitted to Geoderma in July 2012).

Behrens, T.; Schmidt, K.; Ramirez-Lopez, L.; Gallant, J.; A-Xing Zhu; Scholten, T. Hyper-scale digital soil mapping and soil formation analysis. (Submitted to Geoderma July 2012, special issue on pedometrics conference 2011).

Schmidt, K.; Behrens, T.; Daumman, J.; Ramirez-Lopez, L.; Scholten, T. A comparison of calibration sampling schemes at the field scale for proximal gamma ray spectroscopy (Submitted to Geoderma July 2012, special issue on pedometrics conference 2011).

Conference presentations

Ramirez-Lopez, L., Knauer, K. SOC3D: Three dimensional soil organic carbon monitoring using VNIR reflectance spectrometry. In: BruHyp Airborne Imaging Spectroscopy Workshop 2012. Brugge (Belgium).

Ramirez-Lopez, L., Behrens, T., Schmidt, K., Demattê, J.A.M., ,Scholten, T. Optimal calibration set size and sampling strategies for modeling vis–NIR spectra at the field sacle. In: EUROSOIL 2012. Bari (Italy)

Ramirez-Lopez, L., van Wesemael, B., Stevens, A., Doetterl, S., Van Oost, K., Behrens, T., Schmidt, K. Integrating depth functions and hyper-scale terrain analysis for 3D soil organic carbon modeling in agricultural fields at regional scale. In: European Geosciences Union General Assembly 2012. Vienna (Austria).

Behrens, T., Schmidt, K., Ramirez-Lopez, L., Scholten. Contextual mapping approaches for terrain based digital soil mapping. In: Pedometrics conference. 2011. Trest (Czech Republic)

Ramirez-Lopez, L., Behrens, T., Schmidt, K., Viscarra Rossel, R, Scholten, T. Learning a new soil vis–NIR distance metric by using a manifold based approach. In: Pedometrics conference. 2011. Trest (Czech Republic).

Vasques, G.M., Dematte, J.A.M., Viscarra Rossel, R.A., Ramirez-Lopez, L., Terra, F.S., Rizzo, F. Enhancing digital soil mapping in southeaster Brazil: incorporating stream density and soil reflectance from multiple depths. In: Pedometrics conference. 2011. Trest (Czech Republic).

Vasques, G., Demattê, J.A.M, Ramirez-Lopez, L., Terra, F. Soil classification from visible/near-infrared diffuse reflectance spectra at multiple depths. XXXIII Brazilian soil science congress. 2011. Uberlandia (Brazil).

Ramirez-Lopez, L., Behrens, T., Schmidt, K., Viscarra Rossel, R, Scholten, T. New approaches of soil similarity analysis and manifold learning of proximal vis–NIR sensing data. In: Global Workshop in Proximal Soil Sensing. 2011. Montreal (Canada).

Behrens, T., Schmidt, K., Ramirez-Lopez, L., Werban, U., Scholten, T . Digital Soil Sensing and Mapping - Lessons from the iSOIL Project. In: European Geosciences Union General Assembly 2011. Vienna (Austria).

Schmidt, K., Behrens, T., Ramirez-Lopez, L., Werban, U., Scholten, T. Digital Soil Mapping und Geophysik – Erfahrungen aus dem iSoil-Projekt. In: German Digital Soil Mapping Workshop. 2011. Muenchen (Germany)

Ramírez-López, L., Demattê, J.A.M. Terra, F. Bortoletto, M.A. Proximal soil sensing on digital soil fertility mapping for precision agriculture. In: Brazilian Remote sensing symposium. 2009. Natal (Brazil)

Ramírez-López, L.; Demattê, J.A.M. Pedometric methodologies for digital prediction of detailed soil maps. In: Brazilian Remote sensing symposium. 2009. Natal

(Brazil)

Demattê, J.A.M, Fiorio, P.R, Ramirez-Lopez, L. Different strategies on soil mapping by using laboratory and orbital spectral information In: European Geosciences Union General Assembly 2008. Vienna (Austria).

Araújo, S.R, Ramirez-Lopez, L, Demattê, J.A.M, Bellinaso, H Use of proximal soil sensing techniques for the identification of chemical alterations induced by plants and lime applications. In: Brazilian precision agriculture congress– ConBAP, 2008, Piracicaba (Brazil).

Ramirez-Lopez, L., A. Cristancho R., Tovar, J.P, Navia, E., Gutierrez, D. Spatial relationship between oil palm plants affected by "mortal disease" and some soil factors. In: Colombian soil science congress, 2008. Bogota (Colombia)

Ramirez-Lopez, L., Reina, G. A., Camacho Tamayo, J.H. Spatial variation of soil impedance as a factor of several physical attributes. In: V International congress of agricultural engineering, Chile. 2006.