Geoderma 226–227 (2014) 140–150

Contents lists available at ScienceDirect

Geoderma

journal homepage: www.elsevier.com/locate/geoderma

Sampling optimal calibration sets in soil infrared spectroscopy

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

[a] Swiss Federal Institute for Forest, Snow and Landscape Research WSL, Zürcherstrasse 111, 8903 Birmensdorf, Switzerland
[b] Institute of Terrestrial Ecosystems, Swiss Federal Institute of Technology (ETH) Zurich, Universitätstrasse 16, 8092 Zürich, Switzerland
[c] Institute of Geography, Physical Geography and Soil Science, University of Tübingen, Rümelinstraße 19-23, 72070 Tübingen, Germany
[d] 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
[e] Department of Soil Science, Luiz de Queiroz College of Agriculture, University of São Paulo, Av. Pádua Dias, 11 CP 9, Piracicaba, SP 13418-900, Brazil

ARTICLE INFO

Article history:
Received 14 July 2012
Received in revised form 25 January 2014
Accepted 2 February 2014
Available online 6 March 2014

Keywords:
Soil sensing
Sample size
Kennard–Stone sampling
Latin hypercube sampling
Fuzzy *c*-means

ABSTRACT

We investigated the effect of both the calibration set size (number of samples) and the calibration sampling strategy on the performance of vis–NIR models to predict 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), which are commonly used in spectroscopy and digital soil mapping. These algorithms were tested separately using a field-scale dataset and a regional scale dataset. For each dataset we randomly selected a validation subset and the remaining samples were used as candidates for calibration sampling. 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 38 different calibration set sizes varying from 10 to 380 samples. The vis–NIR models were calibrated by using the support vector regression machine (SVM) algorithm. The training root mean square error (RMSE), the normalized RMSE and the prediction RMSE were used to evaluate the sensitivity of the models to both the sampling algorithm and the calibration set size. In addition, we investigated the sample representativeness of each algorithm and we suggest a novel and simple methodology to identify an adequate calibration set size based only on the vis–NIR data (i.e. without prior knowledge of the response variables).

As expected, our results show that the error of the soil vis–NIR models depends on the calibration set size. When the number of calibration samples is relatively small the sampling algorithm may play an important role on the accuracy of the vis–NIR models. On the other hand, if the calibration set size is large enough, the sampling method is not a critical issue. Concerning the sample representativeness, we found for all the algorithms that the original distribution of the vis–NIR data can be better replicated by increasing the calibration set size. The results indicate that the calibration samples selected by the cLHS and by the FCMS algorithms better replicate the original vis–NIR distribution of all the samples, in comparison to those samples selected by the KSS algorithm.

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 infrared spectroscopy have a 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, for soil mapping purposes, 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 mapping accuracy without considerable increase in costs (Wetterlind et al., 2010). In this respect, for a 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 the spectra of a small but representative number of soil samples. Such models can be used to predict attributes efficiently over a large number of soil samples (collected within the same geographical domain) using only their 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 models with good generalization ability, especially when such models are calibrated from complex datasets such as soil spectral libraries (Ramirez-Lopez et al., 2013a).

Despite the well-known potential of vis–NIR spectroscopy to obtain soil information at high spatial resolution, research on both the sampling strategy and the adequate calibration set size has not received enough attention (Grinand et al., 2012; Kuang and Mouazen, 2012). In

* Corresponding author at: Institute of Terrestrial Ecosystems, Swiss Federal Institute of Technology (ETH) Zurich, Universitätstrasse 16, 8092 Zürich, Switzerland. Tel.: +41 44 63 27943.
E-mail address: leonardo.ramirez@usys.ethz.ch (L. Ramirez-Lopez).

http://dx.doi.org/10.1016/j.geoderma.2014.02.002
0016-7061/

principle, the optimal calibration set size could vary depending on the pedodiversity within the geographical domain under study. Hence, strategies for identifying the optimal calibration set size without an explicit prior knowledge of the response variable of interest are of great importance for the practical application of soil spectroscopy at the field scale.

In soil spectroscopy, a calibration set of n samples implies not only n number of spectra but n reference measurements of a given response variable (soil attribute). Despite that (in most of the cases) a large n may lead to reliable models, usually in practical applications n is rather small due to budget and/or time constraints related with the measurements of the response variable. Small calibration sets are usually prone to generate models with poor generalization ability. In this sense, Brown et al. (2005) indicated that the calibration sampling strategy is crucial when the number of samples that can be included in the calibration set is restricted. Furthermore, Minasny and McBratney (2010) stressed the importance to investigate the relation between the calibration sampling strategy and the generalization ability of the models.

Regarding the sampling strategies for obtaining a calibration set, the most common methods that have been 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 method that is 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. Nevertheless, several authors have shown that the strategies employed for covering the multivariate space can lead to different levels of prediction accuracy (e.g. Debaene et al., 2014; Fu et al., 2011; Rodionova and Pomerantsev, 2008; Siano and Goicoechea, 2007).

In this context the main objectives of this paper were: *i.* investigate 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^{++}); *ii.* analyze the sample representativeness on the basis of three different calibration sampling algorithms; *iii.* propose a method to identify an adequate calibration set size based only on the analysis of the vis–NIR data (i.e. without prior knowledge of the response variables).

2. Theory

2.1. Kennard–Stone sampling (KSS)

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

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

For distance computations in the KSS algorithm, the Euclidean distance is commonly used. However, due to the fact that the measurement of the similarity between samples in soil vis–NIR datasets is a very complex task (Ramirez-Lopez et al., 2013b), other strategies to measure the vis–NIR distance between samples can be adopted for the application of the KSS algorithm.

2.2. Conditioned Latin hypercube sampling (cLHS)

In soil spectroscopy, the cLHS (Minasny and McBratney, 2006) has been used for calibration sampling and uncertainty analysis (e.g. McBratney et al., 2006; Viscarra Rossel et al., 2008). The cLHS attempts to cover the multidimensional distribution corresponding to a set of predictor variables by using a stratified random sampling. 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 calibration subset $\mathrm{X_{tr}}$ 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, Eq. (1)) of the cLHS integrates two objective functions O_1 and O_2, so that

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

In this respect O_1 is given by Eq. (2):

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

where m is the number of variables, $\eta(q_j^i \le x_{trj} < q_j^{i+1})$ is the number of samples in $\mathrm{X_{tr}}$ 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 matrices for X and for $\mathrm{X_{tr}}$ respectively. The O_2 is calculated 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 $\mathrm{X_{tr}}$ that returns an O as close as possible to zero where the cumulative distribution of $\mathrm{X_{tr}}$ 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 sampling (FCMS)

Cluster-based sampling is a popular method to select representative calibration sets in vis–NIR spectroscopy as well as in soil science (de Gruijter et al., 2010; Naes, 1987). As its name implies, FCMS works on the basis of the fuzzy c-means clustering algorithm (Bezdek, 1981; Dunn, 1973) and a nearest neighbor search. The algorithm creates sample partitions (clusters) of a given dataset. For each sample in the dataset, its probability to belong to each cluster is computed. The highest probability found determines the (main) cluster membership of each sample. It is expected that samples belonging to the same cluster will share similar characteristics while the dissimilarity between samples in different clusters will be maximized. In this process, cluster centroids are calculated. The optimal fuzzy c-partitions and centroids are found by using the following objective function (Eq. (4)):

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

where $\mathrm{V} = \left\{\mathrm{v}_j\right\}_{j=1}^{c}$, is a matrix of prototypes of cluster centroids, $d(x_i,v_j)^2$ 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):

$$u_{ij} = \frac{d\left(x_i, v_j\right)^{-2/(m-1)}}{\sum_{q=1}^{c} d\left(x_i, v_q\right)^{-2/(m-1)}}. \quad (5)$$

In practice, different results of the final cluster center may be obtained due to different initializations of the cluster centers which may lead to different local minima of the objective function. Concerning the metrics used to compute the similarity between samples, the Euclidean and the Mahalanobis distances are the most commonly used ones, however other metrics can also be used.

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 $m = 2$ (Sun et al., 2012). When fuzzy c-means clustering is used for calibration sampling purposes it is necessary to carry out an additional step i.e. a nearest neighbor search in order 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

For this study, we used a field scale dataset and a regional scale dataset of legacy soil samples, which included their reflectance spectra. The datasets used here were not specifically collected for this study. These samples were collected and analyzed by other people for agronomic purposes.

3.1. Description of the field scale dataset

3.1.1. Field sampling

The field scale dataset covers an area of 5 km^2 and it is located in São Paulo State (Brazil, 22°24′30″S and 48°29′58″W) at altitudes ranging from 500 to 709 m. This area belongs to a single field which 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 campaign was carried out in 1999, and it was based on a dense regular grid of 100 × 100 m where soil samples were collected at 0–0.20 m (459 samples) and 0.80–1.00 m (452 samples). These are arbitrary depth intervals commonly used in this region of Brazil by surveyors and farmers in intensive soil sampling campaigns for soil fertility assessment in sugarcane farms.

3.1.2. Soil vis–NIR spectral scanning and pre-processing

Soil samples were oven-dried for 24 h at 45 °C, and sieved (2 mm mesh) prior to spectral scanning. To obtain the reflectance vis–NIR spectra of the samples, they were scanned using an Infrared Intelligent Spectroradiometer (Geophysical and Environmental Research Corporation, Buffalo, Ney York), which was the sensor available at the time (1999). The final spectrum of each sample was 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 in the range from 1004 to 2500 nm. The spectra of each sample comprised 830 spectral variables. Since the region between 400 and 450 nm presented unusual high reflectance values, we decided to exclude the variables within this spectral range from data. Therefore the final spectra of the field scale dataset comprised 773 variables.

We tested several spectroscopic transformation and pre-processing methods in order to enhance the spectra. Since such methods did not improve the spectral models the raw absorbance spectra obtained from the sensor system were used.

3.2. Description of the regional scale dataset

3.2.1. Area and samples

The samples of this dataset are spread over an area of approximately 464 km^2 which is located in the central–eastern portion of the state of São Paulo (Brazil, 22°51′51″S and 47°36′08″W). The field scale dataset (Section 3.1) falls within the area of the regional dataset.

In this dataset, soil samples correspond to 318 soil profiles collected in different soil surveys over the past 10 years in agricultural fields where sugarcane is the principal crop. The depth intervals at which these profiles were sampled are: 0–0.2 m (318 samples), 0.4–0.6 m (317 samples), and 0.8–1.0 m (291 samples).

In terms of parent material the area is dominated by sandstone, siltstone, and shale 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). Additional details on the terrain variability of this area are given in Behrens et al. (2014).

3.2.2. Soil vis–NIR spectral scanning and pre-processing

The samples were air-dried and sieved (2 mm). The spectral scanning was carried out in September 2008. The vis–NIR (400–2500 nm) reflectance spectra of the samples was measured 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 variables.

For this dataset we also tested several spectroscopic transformation and pre-processing methods in order to enhance the spectra. As in the field-scale dataset, such methods did not improve the spectral models, therefore the raw absorbance spectra were used in this study.

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., 1987) was used to measure the clay content.

3.4. Calibration: sampling, set size and SVM modeling

All the statistical analyses were carried out in R 2.15.3 (R Core Team, 2013).

In order to avoid multi-colinearity and high dimensionality related problems inherent to the vis–NIR spectra, all the sampling procedures were carried out on a principal component (PC) space of the vis–NIR spectra. The number of PCs retained in the analysis was based on the 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.

KSS, FCMS and cLHS were used for selecting calibration samples. For selecting the FCMS sets, a stand-alone function of this algorithm was implemented in R 2.15.3. The R package 'prospectr' (Stevens and Ramirez-Lopez, 2013) was used to select the KSS samples and 'clhs' (Roudier, 2012) to select the cLHS samples. Both packages include functions dedicated to calibration sampling and offer other options for selecting representative sets.

In the case of FCMS we used a fuzzy exponent of 2. For KSS and FCMS the Euclidean distance metric 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 selected a set of samples corresponding to 138 profiles (275 samples, 30% of the total of samples) from the field dataset and a set of samples corresponding to 83 profiles (249 samples, 27% of the total of samples) from the regional dataset. The remaining samples were used as candidate samples for the calibration of the vis–NIR models of soil attributes. To simplify the terminology, we will call this set of samples the "training population". For selecting these validation sets, we sampled by profiles instead of individual samples in order to avoid pseudo-replication of samples (Terhoeven-Urselmans et al., 2010). Fig. 1 shows the spatial distribution of the candidate profiles for calibration sampling as well as the validation profiles for both datasets. For each calibration sampling approach (KSS, cLHS and FCMS) we selected from the training population different calibration sets with sizes varying from 10 to 380 samples in steps of 10 samples.

For each calibration set selected by each sampling method, the support vector regression machine (SVM) algorithm (Drucker et al., 1996) was used for calibrating models of clay content and Ca^{++}. The reason for using SVM instead of the classical partial least square (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 the need to compute it explicitly) with linear or nearly linear structure. We used the linear function kernel (LFK) or dot product in order to keep the models as simple as possible. The LFK 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 accuracy of the SVM models. The optimal C parameter was chosen as the one that minimized the training root mean square error (RMSE; Eq. (6)):

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

where y_i is the observed value of the ith sample, $\hat{y}_i$ is its correspondent predicted value, and n is the number of samples. The normalized RMSE (nRMSE) was also calculated for the calibrations (Eq. (7)).

$$\mathrm{nRMSE} = \frac{\mathrm{RMSE}}{y_{max.} - y_{min.}} \tag{7}$$

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 cLHS is a stochastic method, in order to obtain better estimates of its performance, we repeated the calibration set sampling procedure 10 times for each calibration set size with its corresponding SVM calibration. The averages of the 10 repetitions are the final RMSEs and nRMSEs reported here. The same procedure was carried out for the FCMS, because the random initialization of the clustering algorithm may produce different sampling results (as explained in Section 2.3).

3.5. Sampling representativeness

In order to further evaluate the performance of the sampling algorithms we carried out an analysis of the representativeness of the calibration sets in the PC space. As explained in Section 2.2, the cLHS algorithm ensures that the selected sample sets properly represent the original probability distributions of the explanatory variables observed in the training population. 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 ($\bar{x}$) 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 of the training population. Note that σ^2 is equivalent to 1 and μ to 0 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 ($|\bar{x} - \mu|$) were computed (Eqs. (8) and (9)):

$$\left|s^2 - \sigma^2\right| = \left|s^2 - 1\right| = \frac{1}{k}\sum_{j=1}^{k}\left|s_j^2 - 1\right| \tag{8}$$

$$|\bar{x} - \mu| = |\bar{x} - 0| = \frac{1}{k}\sum_{j=1}^{k}\left|\bar{x}_j - 0\right| \tag{9}$$

where s_j^2 and $\bar{x}_j$ are the sample variance and the sample mean of the jth PC variable (i.e. x_j), μ_j is the original mean of the jth PC and k is the total number of PCs retained.

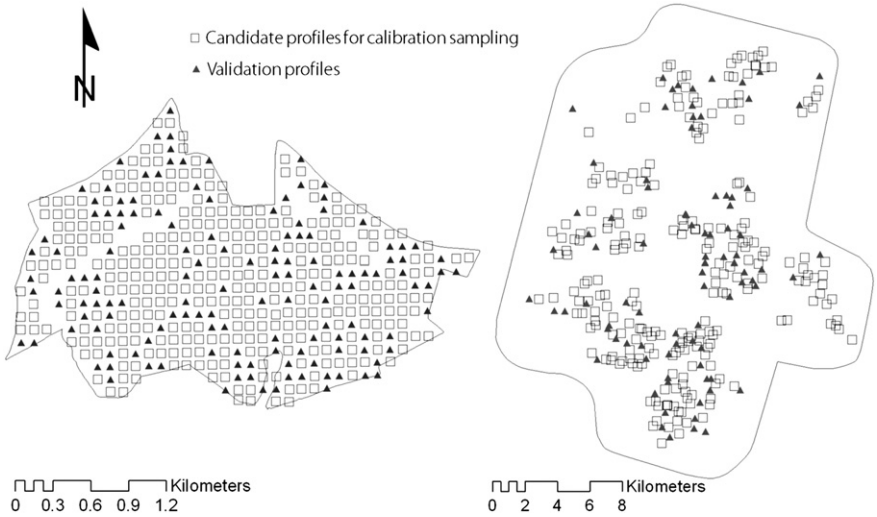

Fig. 1. Spatial distribution of both the candidate profiles for calibration sampling (training population) and the validation profiles in the field scale dataset (left) and regional scale dataset (right).

We also evaluated the mean squared Euclidean distance (msd) as a measure of dissimilarity between the estimates of the probability density functions of the training population and the estimates of the probability density function of a given calibration sample set in the PC space (Eqs. (10) and (11)):

$$\mathrm{msd} = \frac{1}{k}\sum_{j=1}^{k} d^2\left(P_s\left(x_j \in \mathrm{cs}\right), P_p\left(x_j\right)\right) \quad (10)$$

where

$$d^2(P_s(x_j \in \mathrm{cs}), P_p(x_j)) = \int_a^b (P_p(x_j) - P_s(x_j \in \mathrm{cs}))^2 dx_j. \quad (11)$$

$P_s(x_j \in \mathrm{cs})$ is the estimated probability density function of the jth PC of a given set of samples cs (i.e. a calibration set), $P_p(x_j)$ is the probability density function of the jth PC for the whole training population, a and b represent the range of the jth PC, d^2 represents the squared Euclidean distance between the two single distributions being compared, k is the total number of PCs retained and msd is the mean squared distance between probability density functions. To estimate $d^2(P_s(x_j \in \mathrm{cs}), P_p(x_j))$, the density values are estimated for some fixed points between a and b. Brungard and Boettinger (2010) showed that the analysis of the probability distributions of the predictor variables in calibration sets can be useful for selecting adequate calibration set sizes in digital soil mapping.

Overall, the evaluation of the representativeness in the predictor or explanatory variable space at different sample set sizes through the Eqs. (9)–(11), constitute a novel methodology which can be useful for identifying an adequate calibration sample set size (as well as an adequate sampling algorithm) without prior knowledge of the response variables.

4. Results and discussion

4.1. Soil attributes and vis–NIR characteristics

Soil attributes in both datasets show a large variation (Fig. 2). In terms of clay content and Ca^{++}, the large variation in the field scale dataset is mainly due to the high topographic variability. For instance, the range in elevation of the area corresponding to the field scale dataset (500 to 709 m.a.s.l) is almost equivalent to the range in elevation of the regional scale area (489 to 709 m.a.s.l). In the case of the regional scale dataset, the influence of soil forming factors on the soil attributes should be dominated by differences in both parent material and topography.

The vis–NIR reflectance spectra of the field dataset (Fig. 3a) showed well defined absorption features near to 1455 and 1915 nm. These are features relating to structural OH in clays as well as hygroscopic OH (Ben-Dor et al., 2008; Demattê and Garcia, 1999; Petit et al., 1999). 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 1415 nm, 2207 nm and 2160 nm which are related to the kaolinite content (Demattê et al., 2004; Petit et al., 1999; 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

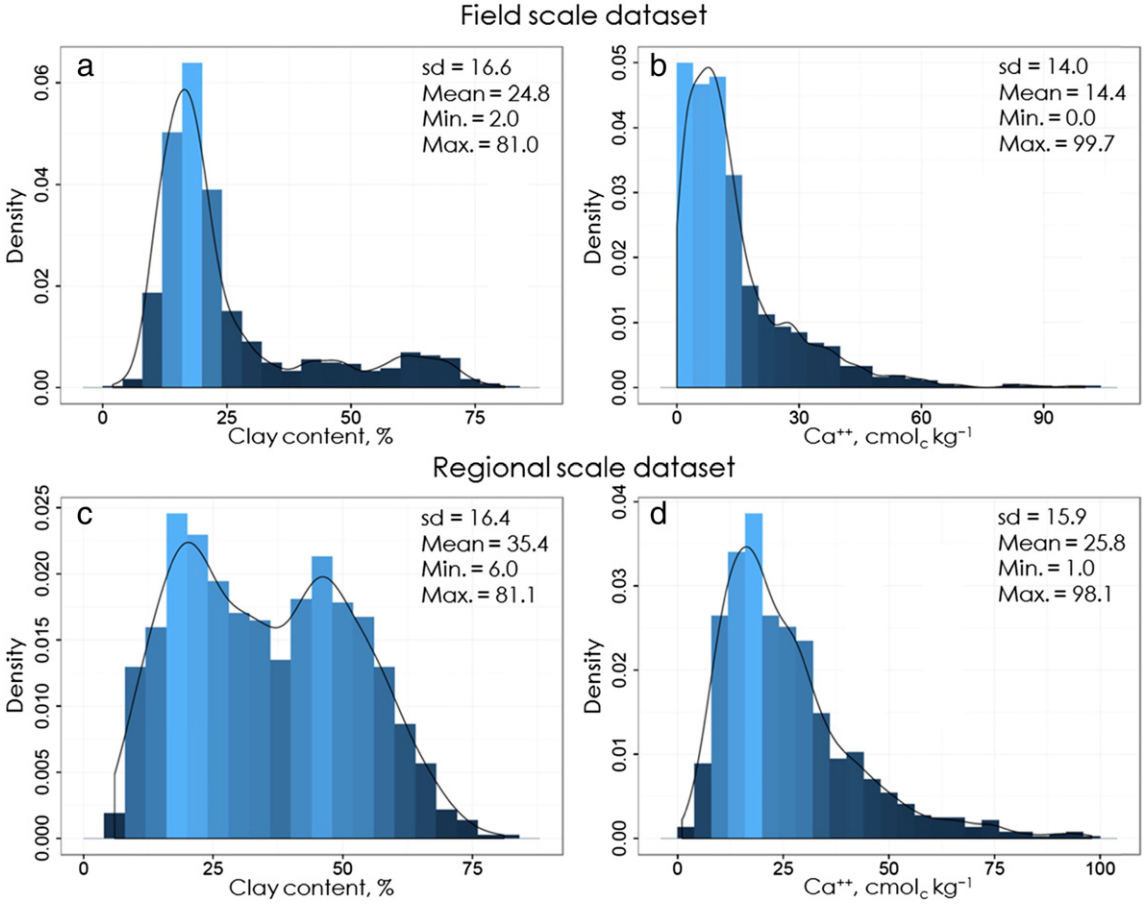

Fig. 2. Density plots of the soil attributes in both datasets.

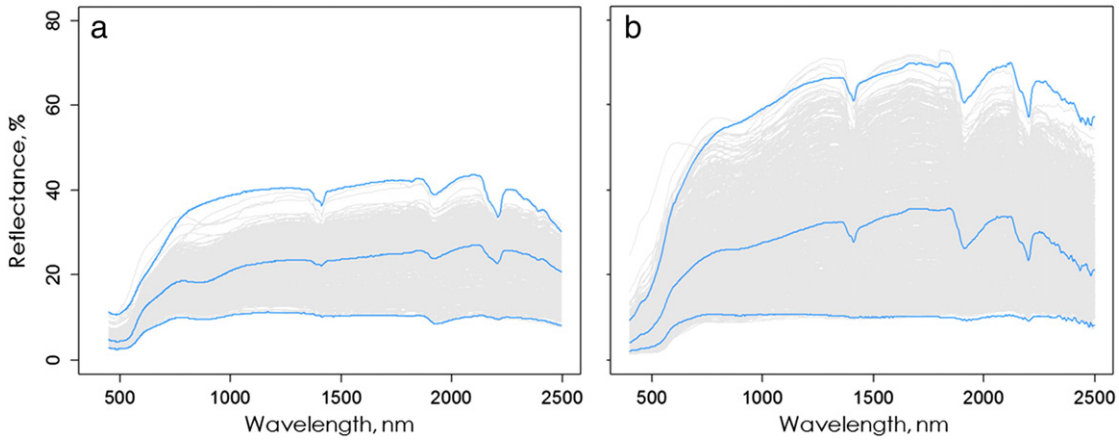

Fig. 3. Reflectance spectra of the field scale dataset (a) and regional scale dataset (b). The highlighted spectra correspond to the samples with the lowest mean reflectance, the highest mean reflectance and the closest samples to the mean of the mean reflectance values.

fact that soils with high clay content present high energy absorption, while soils with high sand content present higher albedo due to higher amount of quartz (White et al., 1997). Concerning the inverse correlation observed for Ca^{++}, 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 the soil samples in the field scale showed similar shape and localization of the absorption features but with considerable variability in their absorption intensities. In this respect, we presume that there should be a specific group of clay minerals that dominate the soils of the area, and that the main spectral differences between samples are associated to the variability in the proportions in which these minerals are mixed.

The vis–NIR spectra of the regional dataset (Fig. 3b) presented larger variation in comparison to the field dataset (Fig. 3a). In the regional dataset we also observed that most of the samples showed an absorption feature in the 1415 nm and 2207 nm related to the presence of kaolinite (Petit et al., 1999). Furthermore, most of the samples presented typical characteristics of energy absorption at 1455 nm and 1915 nm assigned to hygroscopic water. Other features related 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 PCs were retained for the field dataset and 8 for the regional dataset. In both cases, they accounted for 99.9% of the total vis–NIR variation.

4.2. Sampling algorithms and the effect of the calibration set size

As expected, the results showed that the accuracy of the spectral models can be maximized (up to certain point) by increasing the number of samples used to calibrate the models. This has been already reported by several authors (e.g. Brown et al., 2005; Debaene et al., 2014; Grinand et al., 2012; Kuang and Mouazen, 2012; Shepherd and Walsh, 2002). 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 (Fig. 4). 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.

The highest training RMSEs for the field dataset were returned by the KSS algorithm and the differences between KSS and both cLHS and FCMS were markedly high (Fig. 4a,e). In the case of clay content, for calibration set sizes ≤ 90 samples the cLHS returned better results than the FCMS (Fig. 4a). 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 for 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 compared to the cLHS.

The highest training RMSEs in the regional dataset were also produced by the samples selected by the KSS algorithm (Fig. 4h,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 larger 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 (Fig. 4b,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 nRMSE of 0.73 and the FCMS a value of 0.78. For calibration set sizes > 100 samples, the nRMSE of the models of Ca^{++} was 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 (Fig. 4i). Nevertheless, for calibration set sizes ≤ 20 the KSS samples produced much higher nRMSEs in comparison to the cLHS and the FCMS samples. Moreover, in the case of the models of Ca^{++}, the three sampling algorithms produced comparable results in terms of nRMSE (Fig. 4l).

The reason why the differences between the sampling algorithms were considerably wider for the $RMSEs_{tr}$ than for the nRMSEs is due to the range of the soil attribute values of the samples selected by the algorithms. For example, the ranges of the clay content values selected, cLHS and FCMS were 6–81%, selected with the KSS, 7–75% with the cLHS and 7–74% with the FCMS (calibration set size = 150 samples). Similarly for the 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

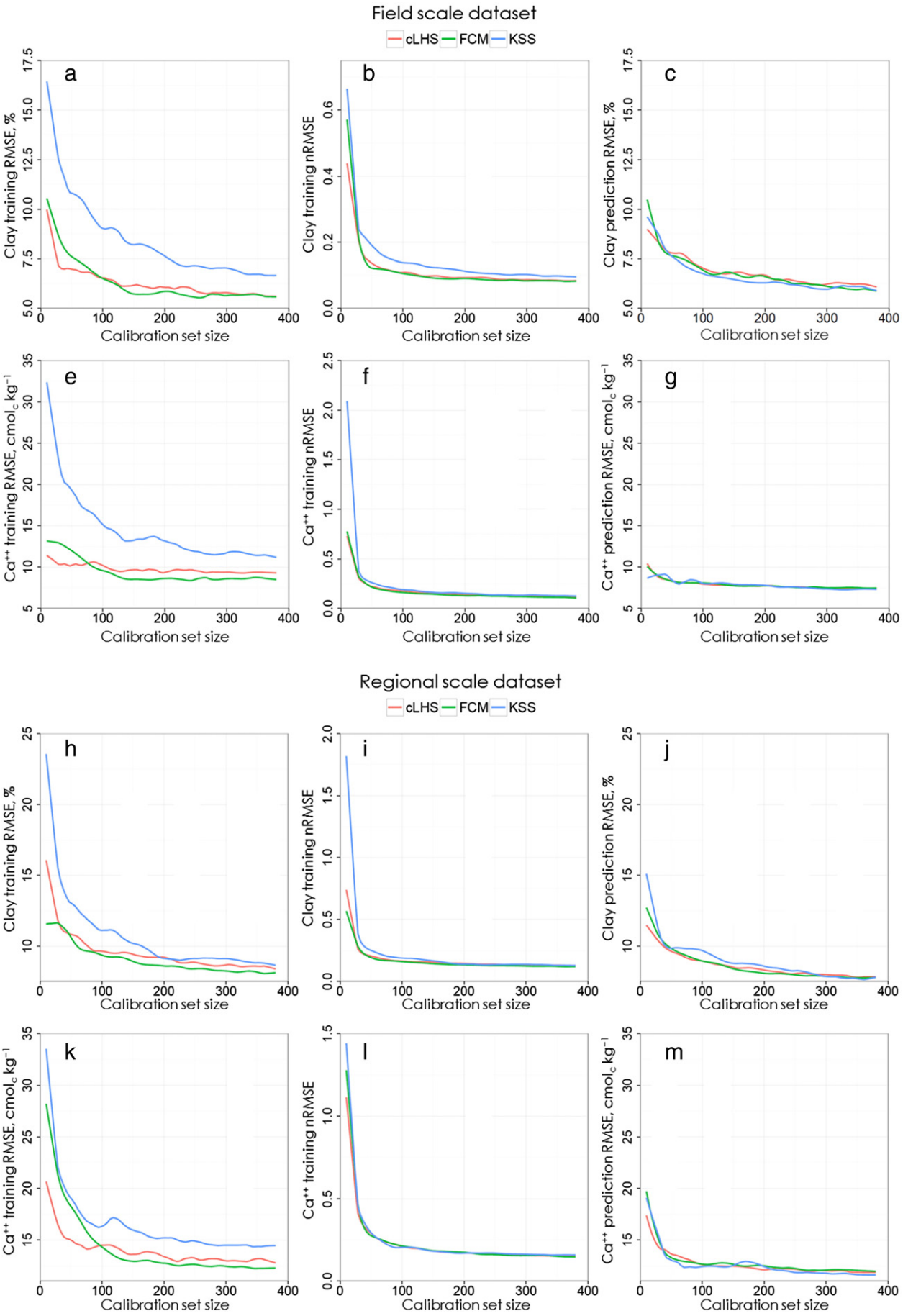

Fig. 4. Training RMSE, nRMSE and prediction RMSE of clay content and exchangeable Ca^{++} against calibration set size in both datasets.

calibration in some cases, especially when the relationship between the predictors and the soil attribute is known, which is not often the case (Minasny and McBratney, 2010). Moreover, in the case of soil vis–NIR spectra collected in the field where spectral measurement errors (due to uncontrolled conditions) are common, the use of the KSS is not recommended. After all, the spectra with large error measurements may be interpreted by the KSS algorithm as highly dissimilar samples with respect to the population measured and therefore some of them could be selected.

Based on the prediction RMSEs, our results are not conclusive. 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. Furthermore, by comparing the estimated uncertainties (training RMSEs) and the generalization errors (prediction RMSEs) (Fig. 4), we also found that the KSS selects samples that lead to overestimated values of the uncertainty.

Despite the similar trends in both datasets, the errors corresponding to regional scale dataset are larger to the ones observed in the field scale dataset. This is probably due to a larger complexity in the regional scale dataset in comparison to the complexity in the field scale dataset. For example the spectra of the regional scale dataset (Fig. 3b) suggest that the mineralogical variability of the samples of this dataset is larger than the mineralogical variability corresponding to the samples in the field scale dataset (Fig. 3a).

Regarding the set size, the differences between the errors produced by samples selected by the different algorithms are larger for small calibration set sizes. Apparently large calibration set sizes ensure a good coverage of the PC space and for this reason the differences between sampling strategy in terms of the modeling error are lower.

4.3. Sample representativeness in the PC space of the vis–NIR data

For the three algorithms the absolute difference between the sample variance and the training population variance ($|s^2 - \sigma^2|$) as well as the difference between the sample mean and the training population mean ($|\bar{x} - \mu|$) 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 $\bar{x}$ of the calibration sets selected by the cLHS showed the highest similarity to their training population equivalents (σ^2 and μ) (Fig. 5). The density distribution of the PCs in the training population was better replicated by the samples selected with the cLHS algorithm which showed the lowest msd values (Fig. 5). In contrast, the density distributions of the PCs of the samples selected with the KSS algorithm were considerably dissimilar to the density distributions of the PCs of the training population, especially in the case of the regional scale dataset. Fig. 5 shows that in the case of the FCMS algorithm, the density distributions of the PCs of the training population were well replicated only when the number of samples was larger than 90.

To further illustrate the performance of the algorithms Fig. 6 shows an example of the distribution of the first PC in the training population and its correspondent distribution in a small calibration set ($n = 30$) as well in a large calibration set ($n = 380$) selected with each sampling algorithm. The probability distribution of the first PC in the training population is poorly reproduced by the samples selected with the KSS and the FCMS algorithms when the number of samples is too small (Fig. 6a,c). In contrast when the number of samples is large enough all the algorithms replicate reasonably well the original probability distribution (Fig. 6b,c).

In general, these results demonstrate the superiority of the cLHS in terms of the replication of the original density distribution of the PC variables in the training population. These results were expected since the cLHS is a stratified sampling algorithm based on the cumulative probability distribution of the variables, while the KSS and the FCMS are distance-based algorithms.

We consider that a sampling strategy that is based only on the maximization of the spectral dissimilarity between calibration samples does not necessarily guarantee a model with good prediction performance. A good sampling strategy must ensure both a good coverage of the predictor space and a good replication of the original distribution of the predictor variables. When the number of samples is large enough, these requirements are fulfilled by the FCMS algorithm.

In general, for calibration set sizes < 130 samples the cLHS outperformed the FCMS in terms of the replication of the distribution of the training population (i.e. mean, variance and probability distribution; Fig. 5). However, at calibration set sizes ≥ 130 samples the cLHS and FCMS produced analogous results. Comparing the KSS to both the cLHS and the FCMS in terms of the differences in the probability distributions, the KSS was largely outperformed by the other algorithms.

4.4. Sample representativeness and predictive performance of the vis–NIR models

Regardless of the calibration sampling algorithm, the calibration set size has a similar effect on both the accuracy of the vis–NIR models (Fig. 5) and the representativeness of the calibration sets (Fig. 6). The larger the calibration set, the higher are the representativeness and the accuracy up to certain limits. One could assume a similar sample set size limit at which both the representativeness of a given calibration dataset and the accuracy become stable. For example, in the regional scale dataset the msd values returned by both the cLHS and FCMS algorithms and the errors of the models calibrated for Ca^{++} become relatively stable above 90 samples (Fig. 4f,k,l,m). A similar tendency can be observed for the training RMSE and nRMSE of the vis–NIR models of clay content (Fig. 4h,i). However, for the prediction RMSE this was not clear (Fig. 4j). Although the error is largely reduced with 90 samples, a slight reduction tendency continues up to 200 samples.

Concerning the sampling algorithms, the tendencies observed for the training RMSE and nRMSE are similar to the tendencies observed in the comparisons of the distributions (Figs. 5 and 6). The highest RMSEs and nRMSEs returned by the KSS are consistent with the poorest representative results returned by this algorithm. However, this is not the case for the prediction RMSE where our results were not conclusive. Overall, we consider that the comparison between the sample distribution and the original distribution of a given training population at different calibration set sizes can be very useful for identifying an adequate calibration set size when there is no prior knowledge about the response variable.

5. Conclusions

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

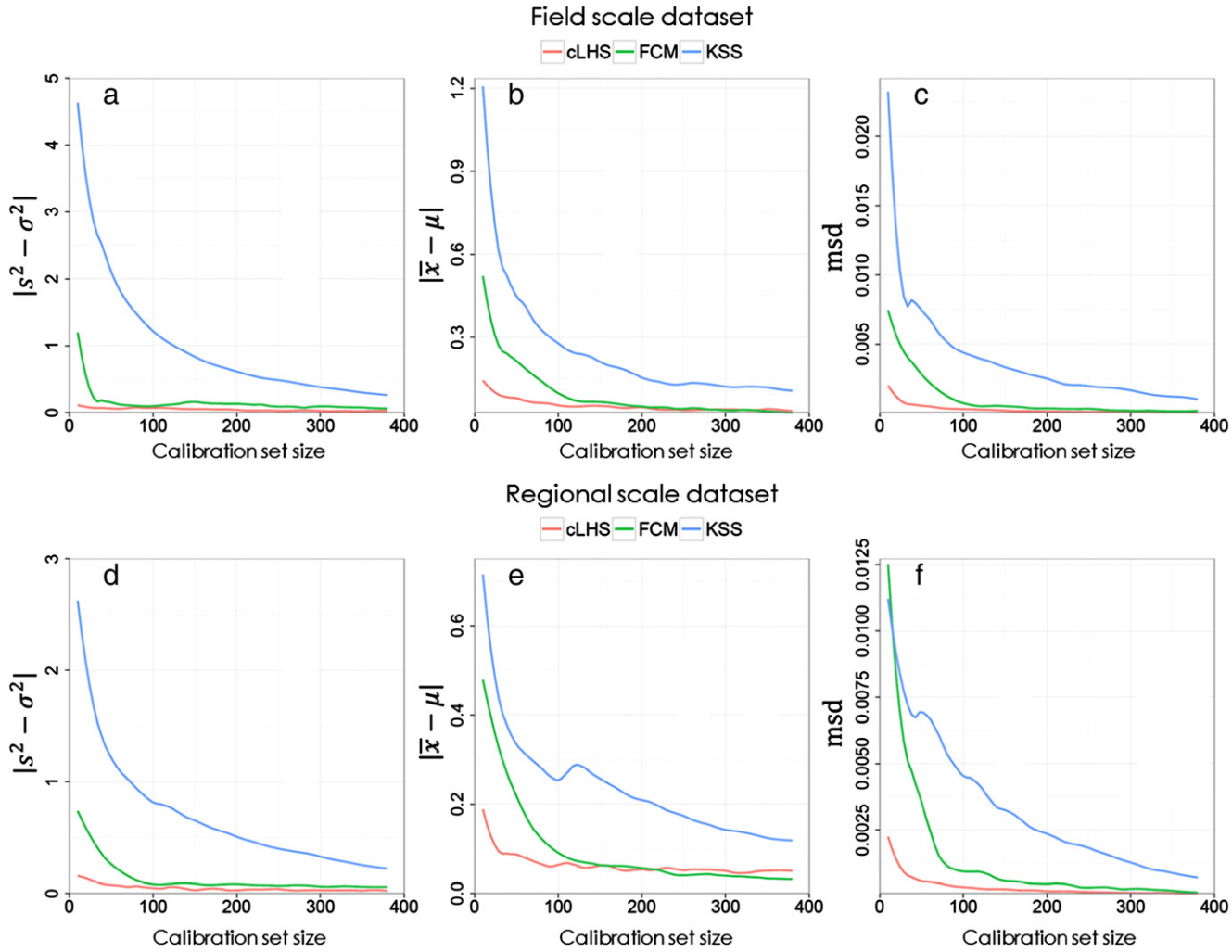

Fig. 5. Calibration set size against the absolute difference between the sample variance (s^2) and the training population variance (σ^2); absolute difference between the sample mean ($\bar{x}$) and the training population mean (μ); and mean squared Euclidean distance (msd) between the probability density functions of the calibration sets and the probability density functions of the training population.

we proposed a method for identifying the optimal calibration set size based on the analysis of the vis–NIR data (i.e. without prior knowledge of the soil attributes to be predicted).

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 spectral space) in the calibration set may be beneficial when the dataset does not contain outlier samples due to spectral measurement errors. However, for datasets containing outliers, the KSS should be avoided or at least a careful outlier removal procedure should be performed prior selecting the samples with the KSS algorithm.

In terms of the prediction errors, the three sampling algorithms returned comparable results. Further research is necessary in this area before a firm conclusion can be reached in this respect.

As expected, we found that the error of the soil vis–NIR models depends on the calibration set size. Particularly for small calibration set sizes, the errors are higher probably due to insufficient coverage of the predictor space and/or a poor representation of the population in which the models are supposed to be applied. Although our results are not entirely conclusive in this respect, we consider that when the number of calibration samples is relatively small the sampling algorithm may play an important role on the accuracy of the vis–NIR models. On the other hand, if the calibration set size is 'large enough', the sampling method is not a critical issue. In this respect, it is particularly relevant to define how large is large enough in terms of the number of samples required to calibrate reliable soil vis–NIR models.

Concerning the sample representativeness in the principal component (PC) space, for all the algorithms we found (as expected) that the original distribution of the vis–NIR data in the 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 small calibration set sizes the cLHS better replicated the original distribution of the PCs in comparison to the FCMS. However, at calibration set sizes $\geq$ 130 the cLHS and the FCMS produced comparable results.

In our study, when the number of samples was large enough, the FCMS algorithm guaranteed both a good coverage of the PC space and a good replication of the original distribution of the predictor variables.

We consider that the comparison between the distribution of the calibration set and the original distribution of the training population offers a solution for solving the question of how large is large enough. This method provides a reasonable strategy for identifying an optimal calibration set size without any explicit knowledge of the soil attributes to be predicted. Furthermore, at least for the estimation of the uncertainty of the vis–NIR models, it can be beneficial to select a calibration

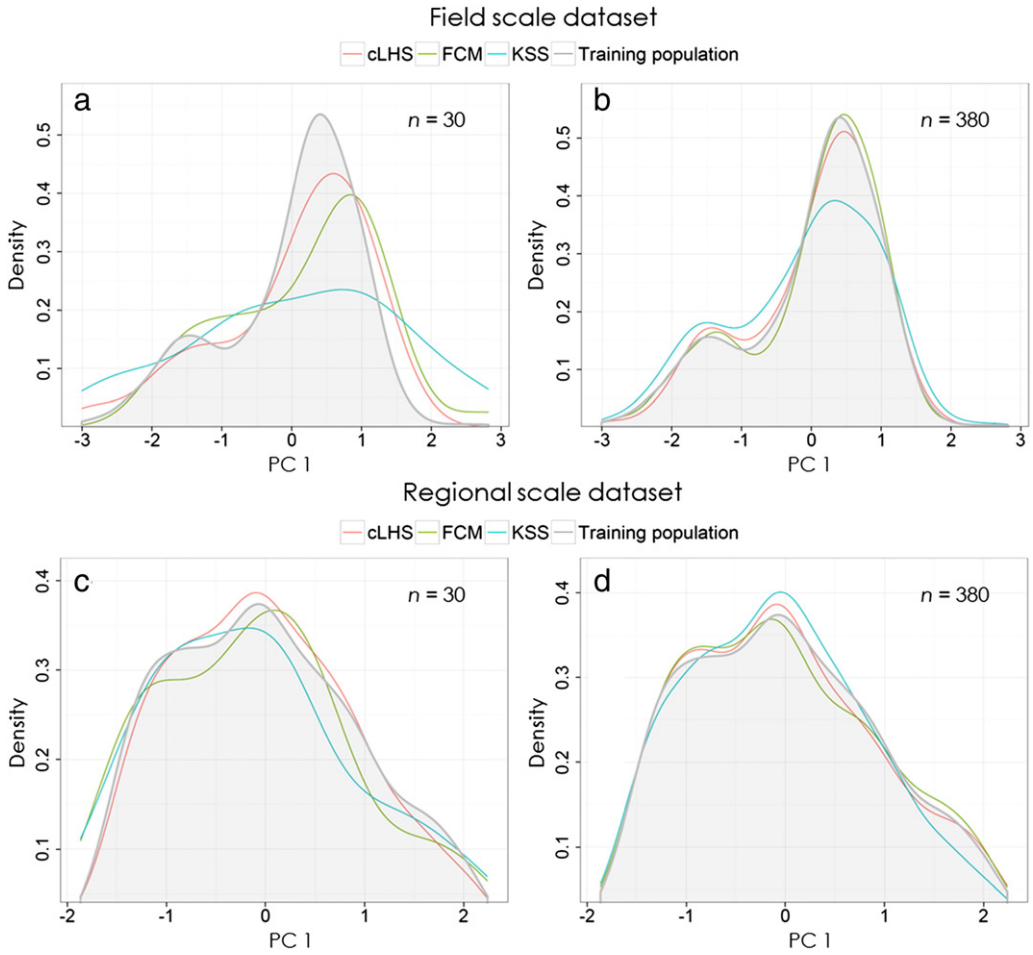

Fig. 6. Distribution of the first PC of the vis–NIR data in the training population and its correspondent distribution in a small calibration set ($n = 30$) as well in a large calibration set ($n = 380$) selected with each sampling algorithm.

sample set whose distribution is close or equal to the distribution of a given population.

Acknowledgments

The present research was partly funded by the State of São Paulo Research Foundation (FAPESP, process 07/58656-8). We thank the anonymous reviewers their helpful comments. We would also like to thank Siul Ruiz for proofreading the present paper.

References

Aizerman, M., Braverman, E., Rozonoer, L., 1964. Theoretical foundations of the potential function method in pattern recognition learning. Autom. Remote. Control. 25, 821–837.
Behrens, T., Schmidt, K., Ramirez-Lopez, L., Gallant, J., Zhu, A., Scholten, T., 2014. Hyper-scale digital soil mapping and soil formation analysis. Geoderma 213, 578–588.
Ben-Dor, E., Taylor, G.R., Hill, J., Demattê, J.A.M., Whiting, M.L., Chabrillat, S., Sommer, S., 2008. Imaging spectroscopy for soil applications. Adv. Agron. 97, 321–392.
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. Commun. Soil Sci. Plant Anal. 36, 9–22.
Brungard, C.W., Boettinger, J.L., 2010. Conditioned Latin Hypercube Sampling: Optimal Sample Size for Digital Soil Mapping of Arid Rangelands in Utah, USA. In Digital Soil Mapping (pp. 67-75). Springer, Netherlands.
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., Kauffman, J.H., 1987. Classificação de solos utilizada em levantamentos pedológicos no Brasil. 12. Boletim Informativo da Sociedade Brasileira de Ciência do Solo, Viçosa, Campinas 11–13 (n.1).
Daszykowski, M., Walczak, B., Massart, D.L., 2002. Representative subset selection. Anal. Chim. Acta. 468, 91–103.
de Gruijter, J.J., McBratney, A., 2010. Sampling for high-resolution soil mapping. In: Viscarra Rossel, R.A., McBratney, A.B., Minasny, B. (Eds.), Proximal Soil Sensing, Progress in Soil Science. Springer Netherlands, Netherlands, pp. 3–14.
De Maesschalck, R., Jouan-Rimbaud, D., Massart, D.L., 2000. The Mahalanobis distance. Chemom. Intell. Lab. Syst. 50, 1–18.
Debaene, G., Niedźwiecki, J., Pecio, A., Żurek, A., 2014. Effect of the number of calibration samples on the prediction of several soil properties at the farm-scale. Geoderma 214, 114–125.
Demattê, J.A.M., Garcia, G.J., 1999. Alteration of soil properties through a weathering sequence as evaluated by spectral reflectance. Soil Sci. Soc. Am. J. 63, 327–342.
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 21, 95–112.
Drucker, H., Burges, C.J.C., Kaufman, L., Smola, A., Vapnik, V., 1996. Support vector regression machines. 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. J. Cybern. 3, 32–57.
Fernandes, R.B.A., Barrón, V., Torrent, J., Fontes, M.P.F., 2004. Quantificação de óxidos de ferro de Latossolos brasileiros por espectroscopia de reflectância difusa. Rev. Bras. Cienc. Solo. 28, pp. 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. 4, pp. 3411–3421.
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). Eur. J. Soil Sci. 63, 141–151.
IUSS Working Group WRB, 2006. World Reference Base for Soil Resources 2006, 2nd edition. FAO, Rome 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., Hummel, J.W., 2009. Soil macronutrient sensing for precision agriculture. J. Environ. Monit. 11, 1810–1824.
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. Eur. J. Soil Sci. 63, 421–429.
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.
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 21, 239–245.
Minasny, B., Mcbratney, A.B., 2006. A conditioned Latin hypercube method for sampling in the presence of ancillary information. Comput. Geosci. 32, 1378–1388.
Minasny, B., McBratney, A., 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, Progress in Soil Science. Springer Netherlands, Netherlands, pp. 15–28.
Næs, T., 1987. The design of calibration in near infra-red reflectance analysis by clustering. J. Chemom. 1 (2), 121–134.
Petit, S., Madejova, J., Decarreau, A., Martin, F., 1999. Characterization of octahedral substitutions in kaolinites using near infrared spectroscopy. Clay Clay Miner. 47, 103–108.
R Core Team, 2013. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria3-900051-07-0 (URL) http://www.R-project.org/.
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. Fundação Cargill, Campinas 170.
Ramirez-Lopez, L., Behrens, T., Schmidt, K., Stevens, A., Demattê, J.A.M., Scholten, T., 2013a. The spectrum-based learner: a new local approach for modeling soil vis–NIR spectra of complex datasets. Geoderma 195, 268–279.
Ramirez-Lopez, L., Behrens, T., Schmidt, K., Viscarra Rossel, R.A., Demattê, J.A.M., Scholten, T., 2013b. Distance and similarity-search metrics for use with soil vis–NIR spectra. Geoderma 199, 43–53.
Rodionova, O.Y., Pomerantsev, A.L., 2008. Subset selection strategy. J. Chemom. 22, 674–685.
Roudier, P., 2012. Introduction to conditioned Latin hypercube sampling with the clhs package. (URL) http://cran.r-project.org/web/packages/clhs/.
Shepherd, K.D., Walsh, M.G., 2002. Development of reflectance spectral libraries for characterization of soil properties. Soil Sci. Soc. Am. J. 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. Chemom. Intell. Lab. Syst. 88, 204–212.
Stevens, A., Ramirez-Lopez, L., 2013. An introduction to the prospectr package. (URL) http://cran.r-project.org/web/packages/prospectr/.
Sun, X.L., Zhao, Y.G., Wang, H.L., Yang, L., Qin, C.Z., Zhu, A., Gan-Lin, Z., Tao, P., Li, B.L., 2012. Sensitivity of digital soil maps based on FCM to the fuzzy exponent and the number of clusters. Geoderma 171, 24–34.
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 Sci. Soc. Am. J. 74, 1792–1799.
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., Jeon, Y.S., Odeh, I.O.A., McBratney, A.B., 2008. Using a legacy soil sample to develop a mid-IR spectral library. Aust. J. Soil Res. 46, 1–16.
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.
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 Sens. Environ. 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. Chemom. Intell. Lab. Syst. 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. Spectrochim. Acta A Mol. Biomol. Spectrosc. 74, 344–348.