Geoderma 290 (2017) 75–82

Contents lists available at ScienceDirect

Geoderma

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

Sampling scheme optimization to map soil depth to petrocalcic horizon at field scale

Marisa Beatriz Domenech [a,*], Mauricio Castro-Franco [b], José Luis Costa [c], Nilda Mabel Amiotti [d]

[a] Instituto Nacional de Tecnología Agropecuaria INTA, CEI Barrow. Ruta 3 Km 488, 7500 Tres Arroyos, Argentina
[b] Consejo Nacional de Investigaciones Científicas y Técnicas CONICET. Av. Rivadavia 1917, C1033AAJ Buenos Aires, Argentina
[c] Instituto Nacional de Tecnología Agropecuaria INTA, EEA Balcarce. Ruta 226. Km 73.5, Balcarce, Argentina
[d] Departamento de Agronomía, Universidad Nacional del Sur, Bahía Blanca. CERZOS-CONICET, Argentina

ARTICLE INFO

Article history:
Received 14 April 2016
Received in revised form 2 December 2016
Accepted 13 December 2016
Available online xxxx

Keywords:
Conditioned Latin hypercube
Digital soil mapping
Precision agriculture
Argentina
Ordinary cokriging

ABSTRACT

Soil depth has played a key role in the development of soil survey, implementation of soil-specific management and validation of hydrological models. Generally, soil depth at field scale is difficult to map due to complex interactions of factors of soil formation at field scale. As a result, the conventional sampling schemes to map soil depth are generally laborious, time consuming and expensive. In this study, we presented, tested and evaluated a method to optimize the sampling scheme to map soil depth to petrocalcic horizon at field scale. The method was tested with real data at four agricultural fields localized in the southeast Pampas plain of Argentina. The purpose of the method was to minimize the sample dataset size to map soil depth to petrocalcic horizon based on ordinary cokriging, five calibration sample sizes (returned by Conditioned Latin hypercube –cLHS-), and apparent electrical conductivity (ECa) or elevation as variables of auxiliary information.

The results suggest that (i) only 30% of samples collected on a 30-m grid are required to provide high prediction accuracy ($R^2 > 0.95$) to map soil depth to petrocalcic horizon; (ii) an independent validation dataset based on 50% of the samples on a 30-m grid is adequate to validate the most realistic accuracy estimate; and (iii) ECa and elevation, as variables of auxiliary information, are sufficient to map soil depth to petrocalcic horizon. The method proposed provides a significant improvement over conventional to map soil depth and allows reducing cost, time and field labour. Extrapolation of the results to other areas needs to be tested.

1. Introduction

The spatial distribution of soil depth affects the spatial dynamic of water storage capacity, runoff generation, subsurface flow, nutrients availability and crops yield (Stieglitz et al., 2003). For that reason, soil depth has played a key role in the development of soil survey, implementation of soil-specific management and validation of hydrological models (Tesfa et al., 2009). However, the spatial dynamic of soil depth at field scale is difficult to predict due to that the complex interactions of factors of soil formation at field scale such as topography, climate, parent material and land use (Jenny, 1941; Tesfa et al., 2010). As a result, the conventional sampling schemes to map soil depth are generally laborious, time consuming and expensive. Evidently, accurate and inexpensive sampling schemes are needed to map soil depth at field scale.

Terrain attributes obtained from digital elevation models and proximal soil sensors data are sources of inexpensive auxiliary information that have been used to map soil depth. For example, Tesfa et al. (2009) reported statistical models to map soil depth based upon the relationship between soil depth and terrain attributes. Ziadat (2010) reported that the modelling depth soil-landscape relationships using terrain attributes was a promising approach to map soil depth. On the other hand, Boettinger et al. (1997) and Bork et al. (1998) determined that electromagnetic induction data are potentially a powerful, inexpensive and quick tool to map soil depth. These examples suggest that terrain attributes and proximal soil sensor data are optimal sources of auxiliary information to map soil depth. However, there is little consensus on the optimal sampling scheme to map soil depth, especially where spatial soil depth pattern is highly variable.

The availability of auxiliary information is important to optimize sampling schemes (Hengl et al., 2004; Minasny and McBratney, 2006; Shaner et al., 2008) and to serve as ancillary variable in the local prediction of a soil property when using hybrid interpolation techniques such as cokriging (Vašát et al., 2010). This interpolation technique is used in cases where there are two or more spatially interdependent variables and incorporates those interdependent variables into spatial interpolation to obtain high prediction accuracy with limited sample data (Wang et al., 2013). Generally, cokriging needs two previous processes to improve prediction accuracy. The first process is a selection of the most important variables of auxiliary information characterized by high

* Corresponding author at: Ruta 3 Km 488 CC 50, 7500 Tres Arroyos, Argentina.
E-mail address: domenech.marisa@inta.gob.ar (M.B. Domenech).

http://dx.doi.org/10.1016/j.geoderma.2016.12.012
0016-7061/

interdependence with the variable to predict. At this respect, Behrens et al. (2010) proposed that using a variables selection technique based on Random Forest (RF), could help to reduce prediction model complexity while decreasing computation time and improving prediction accuracy. The second process is a selection of a model-based sampling scheme that allows quantifying of the spatial dependence and provide good area coverage for reliable prediction (Simbahan and Dobermann, 2006). According to that, several studies of digital soil mapping (DSM) have demonstrated that Conditioned Latin Hypercube sampling (cLHS) (Castro Franco et al., 2015; Minasny and McBratney, 2006; Mulder et al., 2013) could help to minimize the variance of the prediction error of geostatistical interpolation, with limited sample data. Although cokriging, RF and cLHS are being successfully applied as prediction models of several soil properties, their potential to optimize the sampling scheme to map soil-depth at field scale has been underexplored due to their novelty.

The southeast Pampas plain of Argentina, one of the most important cropping regions of the world, have about four million hectares that are underlain by a petrocalcic horizon which limits the soil depth (Pazos and Mestelan, 2002). Consequently, soil depth is the key factor that limits crop yield (Sadras and Calviño, 2001). At present, expensive and laborious sampling schemes are used to map soil depth at field scale. However, most of agricultural fields in the southeast Pampas plain of Argentina have wide availability of inexpensive auxiliary information because precision agriculture technologies have been rapidly adopted in the last decades (Swinton and Lowenberg-Deboer, 2001). In this context, the potential use of this auxiliary information to optimize the sampling schemes to map soil-depth at field scale requires to be evaluated and quantified.

The objective of this study was to present, test and evaluate a method to optimize the sampling scheme to map soil depth to petrocalcic horizon at field scale, based on inexpensive auxiliary information, RF as algorithm of importance variables selection and cLHS as model-based sampling scheme. The integration of these algorithms offers a new approach to optimize the sampling scheme, to identify the most important variables of auxiliary information and to overcome the limitations of conventional methods. Also, the parameterization of cokriging, RF and cLHS is very simple and computationally slighter than other algorithms.

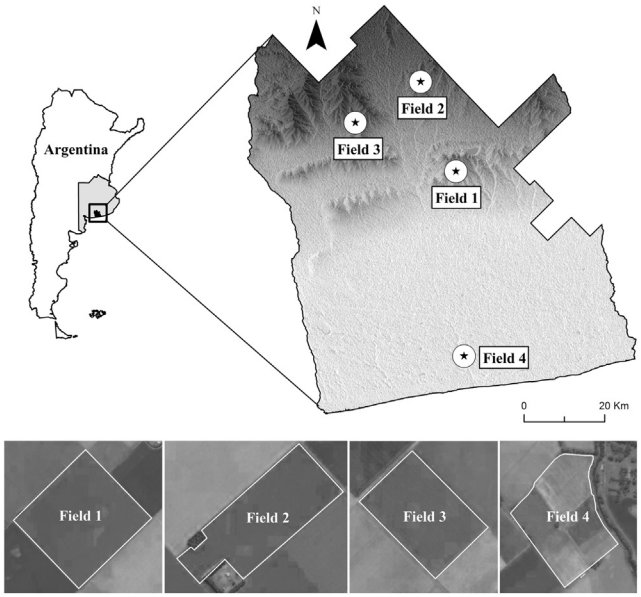

Fig. 1. Location of the study fields in the southeast Pampas of Argentina.

2. Materials and methods

2.1. Agricultural fields

The location of the fields used in this study is shown in Fig. 1. These fields were selected because they represent the variability of elevation, landscape position and spatial variability of soil depth usually found in the southeastern Pampas. The current crop rotations in all fields include corn, soybean or sunflower in summer and wheat or barley in winter (Costa et al., 2015). Specifically, the fields are located in the geological province locally termed "Sierras Septentrionales" in the southeast of Buenos Aires province of Argentina. In this zone, the loess deposits are from the Late Holocene and Pleistocene (Blanco and Stoops, 2007). The soils are classified as Subgroups Typic Argiudoll and Petrocalcic Argiudoll; Family fine, illitic, thermic (Soil Survey Staff, 2014). Table 1 shows the area and composition of soil mapping unit for each field.

The southeastern Pampas plain of Argentina has a frost-free period that extends from October to May. The mean annual temperature is 14.8 °C. It has a humid and subhumid hydric regime (Thornthwaite, 1948). The mean annual precipitation is about 756 mm. The rain regime is (i) rainy from October to March, (ii) moderately rainy in April, May and September, and (iii) scarcely rainy from June to August (Costa et al., 2015).

2.2. Auxiliary information measurement

ECa and elevation were used as auxiliary information to optimize sampling to map soil depth at field scale.

ECa measurements were collected at two different dates (July 18th, 2008 in Field 4 and June 23rd–30th, 2011 in Fields 1, 2 and 3) using a Veris® 3100 soil electrical conductivity sensor (Veris Technologies Inc., Salina, KS, USA). The accumulated rainfall reached 612 mm from January to July 2008, whereas only 211 mm were accumulated from December to June 2011. However, a rainfall of 8.5 mm occurred on June 22nd 2011. Precipitation data were provided by Agrometeorological Department of National Institute for Agricultural Technology of Argentina (INTA-CEI Barrow) from the nearest weather recording station for each field.

The coulter electrodes of the Veris® 3100 are configured as a Wenner array, an arrangement commonly used for geophysical resistivity surveys. In this sensor, the system records ECa in mS m^{-1} by electrical resistivity at a shallow depth (0–30 cm, ECa_30 cm) and a deep depth (0–90 cm, ECa_90 cm) (Moral et al., 2010). Veris® 3100 was pulled through the field by a pick-up truck. ECa measurements were made along parallel transects approximately 20 m apart on the surface of each agricultural field. An advance GPS Surveying instrument GPS Trimble® GeoXT™ handheld with submeter accuracy was used to georeferenced the ECa measurements. Latitude, longitude, ECa_30 cm and ECa_90 cm data were recorded in an ASCII text file and transferred to GIS software for further analysis. For more details of ECa measurements with Veris 3100® see Corwin and Lesch (2003), Corwin and Lesch (2005) and Allred et al. (2008).

Elevation was measured simultaneously with ECa, using an advance differential GPS Surveying instrument GPS Trimble®R3 (Trimble Navigation Limited, CA, USA), which is equipped with a GPS receiver, antenna and rugged handheld controller. Elevation data were post-processed with Trimble Business Center software V3.5 to produce a digital elevation model of spatial resolution of 10 m, in each field.

Experimental variograms were computed to describe the spatial variation of ECa and elevation following the procedure proposed by Diggle and Ribeiro (2007).

The adjusted experimental variogram was used to interpolate ECa and elevation by ordinary kriging in each field. The R package "geoR" was used to conduct the geostatistical interpolation (R Development Core Team, 2015). Finally, a 10 × 10 m grid square size was chosen for output maps.

Table 1
Agricultural fields and soil classification.

Field no	Area (ha)	Soil type			
		U.M.	Kind of U.M.	Soil map unit	Soil classification
Field 1	67.22	CM11	Complex	Claudio Molina (50%)	Typic Argiudoll
				El Gavilan (30%)	Petrocalcic Argiudoll
				Micaela Cascallares (10%)	Typic Argiudoll
		LPd13	Consociation	Laprida (100%)	Typic Argiudoll
Field 2	31.90	TA48	Association	Tres Arroyos (80%)	Petrocalcic Argiudoll
				Semillero Buck (20%)	Typic Natracuoll
Field 3	17.46	LPd11	Association	Laprida (50%)	Typic Argiudoll
				Tres Arroyos (50%)	Petrocalcic Argiudoll
Field 4	25.82	TA24	Association	Tres Arroyos (80%)	Petrocalcic Argiudoll
				Copetonas (20%)	Typic Argiudoll

2.3. Sampling scheme optimization

The sampling scheme optimization was developed in five steps:

2.3.1. Step 1: soil depth sampling scheme and calibration of sample size

Each field was divided using a regular square grid based on 30 × 30 m spacing, because this scale reflects the variability of soil depth associated with farm scale in the study area (Castro Franco et al., 2015). Three soil depth samples were collected at the nodes of a 30-m grid, by using a truck-mounted Giddings Soil Sampler (Model XHDGSRPST Giddings Machine Co., Fort Collins, CO, USA). The pitcher barrel sample was about 150 cm in length. Sample depths were measured and marked on the pitcher barrel to determine effective soil depth to petrocalcic horizon. The soil depth samples were separated into calibration and validation datasets.

In order to separate the calibration dataset, the cLHS algorithm was run using the R package "*clhs*" with ECa_30 cm and ECa_90 cm and elevation as auxiliary information (R Development Core Team, 2015; Roudier et al., 2012). cLHS is a stratified random procedure that picks samples based on the distribution of auxiliary information (Roudier et al., 2012). In cLHS, a Latin Hypercube is constructed by random sampling from the accumulative distribution of variables of auxiliary information, using a simulated annealing optimization approach, which focuses on preserving the correlation between variables of auxiliary information in the selected sampling set (Brungard and Boettinger, 2010; Minasny and McBratney, 2006). Recently, several studies have determined that cLHS is able to capture significant variation in soil properties by using limited samples (Castro Franco et al., 2015; Rad et al., 2014).

For calibration dataset, we tested five sample sizes (calibration sample sizes) equivalent to 10, 20, 30 40 and 50% of the total soil depth samples which were taken at the nodes of a 30-m grid, whereas as validation dataset, we tested a sample size equivalent to 50% at each field. The samples of the calibration were independent to the samples of the validation. Table 2 and Fig. 2 show the number of samples equivalent to each calibration and validation datasets, and their spatial distributions within each field.

Table 2
Number of samples for different percentages of soil depth samples on a 30-m grid for each field.

Field no	Number of samples on a 30-m grid					
	10%	20%	30%	40%	50%	100%
Field 1	74	147	221	295	368	736
Field 2	34	68	102	136	170	340
Field 3	20	40	61	80	101	202
Field 4	26	52	76	104	126	252

2.3.2. Step 2: evaluation of sampling representativeness

The performance of the calibration datasets obtained by cLHS, was evaluated from an analysis of representativeness (Ramirez-Lopez et al., 2014).

For this analysis, were compared the sample mean ($\bar{x}$) and the sample variance (s^2) of the variables of auxiliary information with the original mean (μ) and the original variance (σ^2) for each calibration sample size (i.e. 10, 20, 30, 40 and 50% of total soil depth samples). The absolute difference between means ($|\bar{x} - \mu|$) and the absolute difference between variances ($|s^2 - \sigma^2|$) were computed as:

$$|\bar{x} - \mu| = |\bar{x}_j - \mu|$$

$$|s^2 - \sigma^2| = |s_j^2 - \sigma^2|$$

where $\bar{x}_j$ and s_j^2 are the sample mean and sample variance of the *j*th variable of elevation, ECa_30 cm or ECa_90 cm, respectively, and μ_i and σ_i^2 are the original mean and original variance of elevation, ECa_30 cm or ECa_90 cm, respectively.

2.3.3. Step 3: importance of auxiliary information

The R package "*randomForest*" was used to establish the importance of elevation, ECa_30 cm ECa_90 cm to predict soil depth in each calibration sample size (Liaw and Wiener, 2002; R Development Core Team, 2015).

The *Random Forest* classifier (RF) uses numerous decision trees, n_{trees}, (e.g., CART or C4.5 algorithm). Each tree is constructed using bootstrap sampling, which is approximately 2/3 of the available data. The remaining 1/3 of available data are referred to as out-of-bag (OOB) and the proportion of misclassification of these samples (OOB_{error}) can be used as a measure of generalization errors (Breiman, 2001). At each binary split, the variable of auxiliary information that produces the best split is chosen from a random subset of the entire variable set. The number of predictors in each random subset is called *mtry*. The optimal n_{trees} and *mtry* must be identified by the user. For more details of the RF algorithm, see Breiman (2001) and Genuer et al. (2010).

Several studies of DSM have demonstrated that RF has a great ability to provide variable importance measures from auxiliary information (Castro Franco et al., 2015; Grimm et al., 2008; Rad et al., 2014). These measures are usually used for variable selection meant to differentiate relevant from non-relevant variables (Genuer et al., 2010; Grömping, 2009).

The RF algorithm was run 20 times with ECa_30 cm, ECa_90 cm and elevation as predictors, and soil depth as the target variable. The average of predictor importance was estimated across these 20 runs. Finally, the most important predictor was selected for each field.

2.3.4. Step 4: soil depth mapping

OCK was used to generate a map of soil depth to petrocalcic horizon for each field. OCK is a geostatistical interpolation method that allows

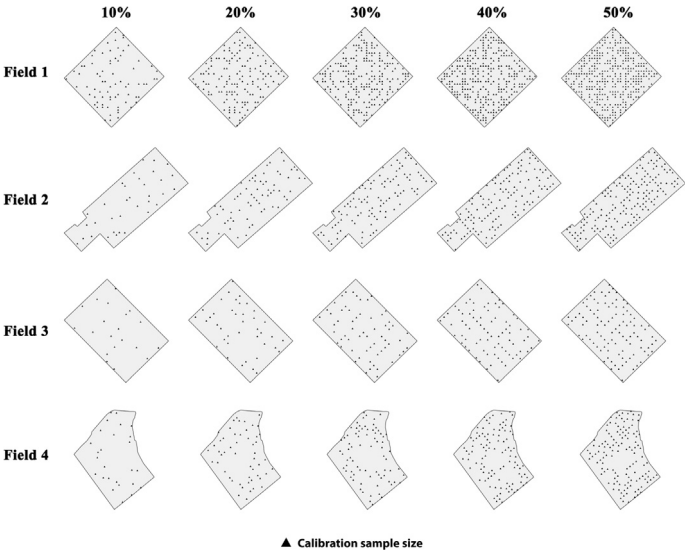

Fig. 2. Spatial distribution of five different percentages of soil depth samples on a 30-m grid for each field.

incorporating variables of auxiliary information (Wang et al., 2013). OCK is most commonly applied when a variable of auxiliary information has been densely sampled and it is inexpensive, fast and easy to measure. Normally, OCK determines the coregionalization between two or more variables. When this coregionalization exists, then it is feasible to use the variables of auxiliary information to improve the predictions of the target variable through OCK (Pang et al., 2009).

In this study, OCK was carried out to predict the soil depth value at unsampled locations from variables of auxiliary information for each calibration sample size. The OCK prediction of soil depth was computed as:

$$\overline{Z}_u\,(X_0) = \sum_{i=1}^{N} \lambda^u_{u,i} Z_u(X_i) + \sum_{i=1}^{P} \lambda^u_{v,i} Z_v(X_i),$$

where $\overline{Z}_u\,(X_0)$ is the estimated value of soil depth at location X_0; the values of $\lambda^u_{u,i}$ and $\lambda^u_{v,i}$ are OCK weights; $Z_u(X_i)$ and $Z_v(X_i)$ are the target variable and variable of auxiliary information, respectively; and N is the number of measured values of $Z_u(X_i)$ and P is the number of measured values of $Z_v(X_i)$ used in estimation at location X_0, respectively.

2.3.5. Step 5: accuracy of soil depth maps

The accuracy of each soil depth map was assessed by comparing between predicted soil depth values at interpolation points and measured soil depth values of the validation datasets. The normalized root mean square error (nRMSE) and the adjusted coefficient of determination (Adjusted R^2) between the predicted soil depth values and measured soil depth values of the validation datasets were used to verify global prediction accuracy. nRMSE was estimated by

$$nRMSE = \frac{RMSE}{y_{max} - y_{min}}$$

where y_{max} and y_{min} are the maximum and the minimum values of the measured soil depth in each calibration sample size. RMSE was estimated by

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

where y_i is the measured soil depth value of the ith sample, $\overline{y}_i$ is its corresponding predicted soil depth value, and n is the number of samples.

Finally, a linear plateau regression between R^2, nRMSE and each calibration sample size was carried out to determine the optimal percentage of samples to map soil depth to petrocalcic horizon at field scale.

3. Results and discussion

3.1. ECa and elevation values

Table 3 shows the summary statistics of ECa and elevation values for each field.

As we expected, ECa values were different among fields. In Field 1, mean values of ECa_30 cm were lower, whereas in Field 4 were higher. In Field 3, mean values of ECa_90 cm were the highest. These results could be due to differences in soil water content at each ECa measurement date; and also to differences in the vertical spatial variability of soil among fields. At this respect, numerous works have reported that ECa values are governed by the status of the soil water content (Corwin and Lesch, 2005; McCutcheon et al., 2006). On the other hand, some works have reported that the vertical spatial variability of soil depth to cemented or argillic horizons could be the primary cause of ECa variation (Boettinger et al., 1997; Myers et al., 2010). It is important to note that minor differences between ECa_30 cm and ECa_90 cm in Field 1 and 3 could indicate a more homogeneous soil profile, in

Table 3
Statistical summary of ECa_30 cm, ECa_90 cm and elevation for each field.

Farm field	Auxiliary information	Units	Mean	S.D.	Min*	Max†
Field 1	Soil depth	m	0.95	0.16	0.64	1.44
	ECa_30 cm	mS m^{-1}	19.95	0.95	15.67	23.66
	ECa_90 cm	mS m^{-1}	20.49	1.94	13.77	26.35
	Elevation	m	133.09	4.68	125.39	141.44
Field 2	Soil depth	m	0.65	0.18	0.23	1.20
	ECa_30 cm	mS m^{-1}	26.94	5.72	1.74	41.21
	ECa_90 cm	mS m^{-1}	24.49	5.05	3.05	34.59
	Elevation	m	169.76	3.44	164.34	176.53
Field 3	Soil depth	m	0.64	0.19	0.41	1.21
	ECa_30 cm	mS m^{-1}	27.55	5.87	17.02	39.28
	ECa_90 cm	mS m^{-1}	28.53	2.63	20.67	33.82
	Elevation	m	156.40	0.58	155.30	157.95
Field 4	Soil depth	m	0.72	0.17	0.43	1.10
	ECa_30 cm	mS m^{-1}	30.26	6.10	15.31	45.91
	ECa_90 cm	mS m^{-1}	25.90	4.27	15.15	35.65
	Elevation	m	35.94	1.59	32.60	37.92

* Minimum value.
† Maximum value.

comparison with Field 2 and 4 (Bork et al., 1998). Minimum values of both ECa_30 cm and ECa_90 cm in the Field 2 could be associated with a very shallow soil depth to petrocalcic horizon.

Table 4 shows a comparison of variogram model parameters of elevation, ECa_30 cm and ECa_90 cm for all fields.

Clearly, Matérn variogram model provided the best fit for ECa and elevation. Several studies have determined that the Matérn model is flexible and can be used to describe many isotropic soil spatial processes because it may adequately describe an experimental variogram over small lags (Minasny and McBratney, 2005). Similar parameters variogram values of ECa and elevation were found by Carroll and Oliver (2005) and Kumhálová et al. (2011), respectively. It is interesting to note that the exponential model fitted to ECa_90 cm values in Field 2, might indicate a complex spatial pattern of soil depth.

Large nugget value of ECa_30 cm in Field 3 may be due to changes of soil depth over short distances (Pazos and Mestelan, 2002) (Table 4). The range of spatial dependence of ECa_90 cm in Field 2 was considerably larger not only for changes in soil depth over long distances but also for a very shallow soil depth to petrocalcic horizon. Low sill values in both ECa_30 cm and ECa_90 cm in Field 1 reflected less variance of soil depth to petrocalcic horizon, confirming the results shown in Table 3. The range of spatial dependence of elevation was considerably larger in Field 1 and 2 due to changes in elevation over long distances. On the other hand, low sill and range values in Fields 3 and 4 indicate small variance at short distance. These changes in variograms parameters of elevation suggest that it would not have a stable predictive performance to map soil depth to petrocalcic horizon.

Surprisingly, the spatial relationships between ECa and elevation were inconsistent (Fig. 3). Fields 1 and 4 showed a trend to higher values of ECa in low elevation zones, whereas Fields 2 and 3 showed high values of ECa either low or upper elevation zones. These inconsistences could be attributed to differences in the soil water content at ECa measurement date; and also to complex spatial interactions between topography and effective soil depth, mainly in Fields 2 and 3.

3.2. Calibration sample sizes effect

Fig. 4a shows the plot of the absolute difference between the sample variance (S^2) and the original variance (σ^2) and the calibration sample sizes. In terms of the ECa_30 cm and as we expected, it is clear that the smallest percentage (i.e. 10%) of calibration sample size yields the largest $S^2-\sigma^2$ for all fields (e.g. Field 1–7). Conversely, when the largest percentage (i.e. 50%) of calibration sample size yields the smallest $S^2-\sigma^2$ as exemplified in Fields 1 and 2. In practical terms however, it is clear that when an intermediate percentage (i.e. 30%) of calibration sample sizes is considered, yields $S^2-\sigma^2$ values (~3.25) equivalent to the largest percentage of calibration sample size.

Table 4
Model parameters for fitted variograms of ECa_30 cm, ECa_90 cm and elevation for all fields.

Field no	Predictor	Model	Co†	Co + C‡	a (m)§
Field 1	ECa_30 cm	Matern	0.00	0.72	119.54
	ECa_90 cm	Matern	0.00	2.83	194.27
	Elevation	Matern	0.00	207.15	1388.80
Field 2	ECa_30 cm	Matern	0.00	18.84	126.72
	ECa_90 cm	Exponential	0.02	23.54	443.69
	Elevation	Matern	0.00	6.97	6565.00
Field 3	ECa_30 cm	Matern	0.91	59.96	203.49
	ECa_90 cm	Matern	0.28	4.83	56.35
	Elevation	Matern	0.00	0.66	237.83
Field 4	ECa_30 cm	Matern	0.00	36.67	100.50
	ECa_90 cm	Matern	0.00	16.72	103.73
	Elevation	Matern	0.00	5.68	311.72

† Nugget variance.
‡ Sill.
§ Range.

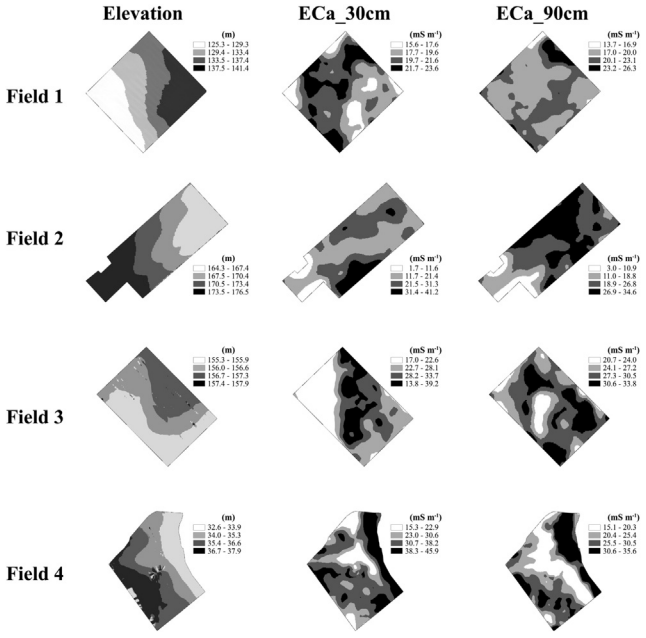

Fig. 3. Maps of elevation, ECa_30 cm and ECa_90 cm for each field.

Fig. 4b shows the plot of the absolute difference between the sample mean ($\bar{x}$) and the original mean (μ) and the calibration sample sizes. In terms of the ECa_30 cm, the smallest percentage (i.e. 10%) yields the largest $\overline{X}-\mu$ for all fields (e.g. Field 1 ~0.25). Conversely, when the largest percentage (i.e. 50%) of calibration sample size yields the smallest $\overline{X}-\mu$, as is shown in Field 1 (~0). As with the $S^2-\sigma^2$, it is clear that when an intermediate percentage (i.e. 30%) of calibration sample size is considered, yields $\overline{X}-\mu$ values (~0.25) equivalent to the largest percentage of calibration sample size.

The results for ECa_30 cm are often replicated and are equivalent for ECa_90 cm and elevation. In this study, note that the highest $S^2-\sigma^2$ and $\overline{X}-\mu$ values for ECa data in Field 4 were most likely a function of soil water content at measurement date. Clearly, the transient nature of soil water complicates the characterization of ECa variability by altering its response to a given soil depth during ECa measurement. The results with respect to the elevation were large in Field 1 because of its higher elevation variability than Fields 2, 3 and 4. At this respect, we observed that in Field 1, cLHS tends to select a wider range of elevation values. Therefore, the highest $S^2-\sigma^2$ and $\overline{X}-\mu$ values for elevation data may be interpreted by cLHS algorithm as highly dissimilar samples with respect to the population measured.

In general, these results suggest that (i) cLHS is an adequate algorithm to replicate the original distribution of both ECa and elevation at field scale and (ii) the density distribution of both ECa and elevation was adequately replicated only when the calibration sample size was larger than 30% on a 30-m grid for all fields. Similar results have been reported by Ramirez-Lopez et al. (2014), who found that the error of models to predict soil properties at field scale using vis-NIR depends on the calibration sample size, and that when it is small, the sampling algorithm may play an important role in the accuracy of the models. Based on our results, we consider that the simultaneous use of the cLHS algorithm and the calibration of an adequate percentage of samples is a reasonable strategy to identify an optimal sample size to predict soil depth to petrocalcic horizon at farm-field scale. This strategy guarantees a good coverage of the soil depth to petrocalcic horizon and a good replication of the variables of auxiliary information in any condition.

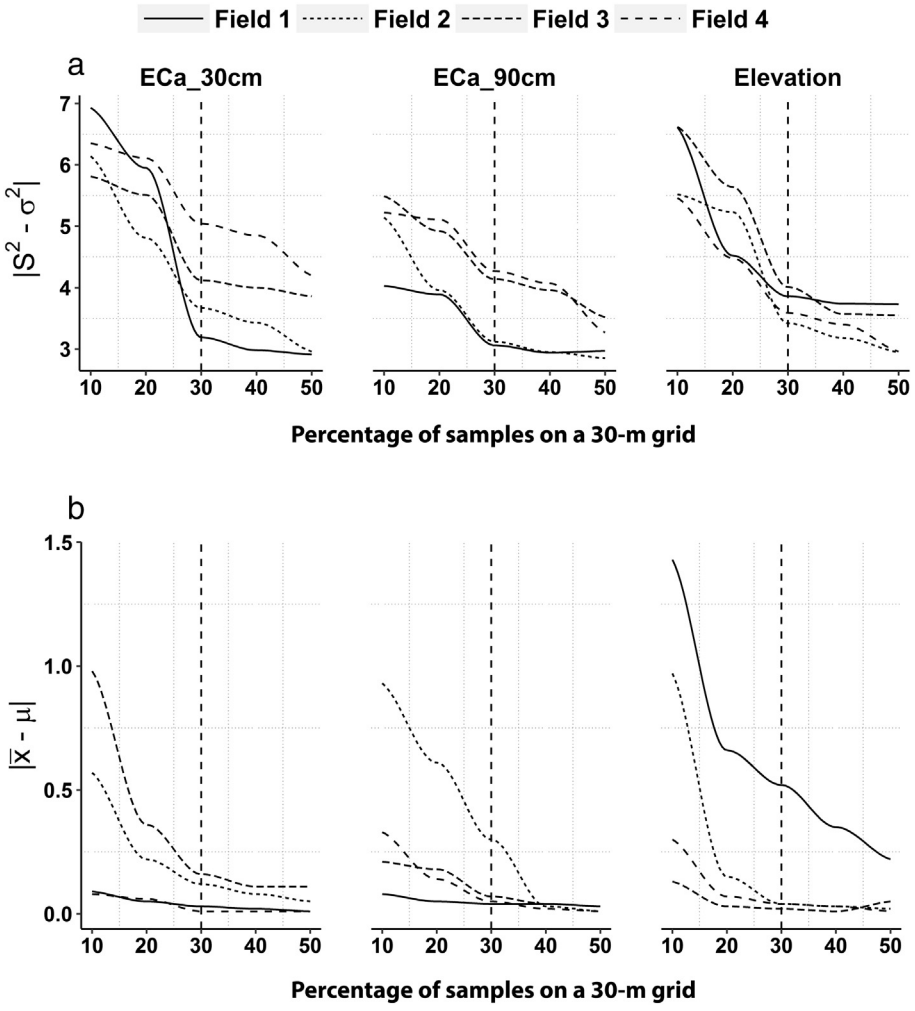

Fig. 4. Calibration sample sizes effect on the absolute difference between the sample variance (S^2) and the original variance (σ^2), the absolute difference between the sample mean ($\bar{x}$) and the original mean (μ), and between the probability density functions of the calibration and the probability density functions of the validation.

3.3. Importance of the variables of auxiliary information

Fig. 5 shows the plot of importance measures as determined by RF against the percentage of data available and with respect to elevation, ECa_30 cm and ECa_90 cm. In Field 1, the most important variable of auxiliary information to map soil depth was elevation and as shown by the fact that the importance measure was a maximum for all percentages. For example, the importance measure for 50% of the elevation data was a maximum (~60). This was closely followed by ECa_30 cm data. From here, the importance measure systematically decreased for both sources of auxiliary information.

Fig. 5 shows an equivalent systematic reduction in the importance measure with percentages decrease in auxiliary data available except herein the best auxiliary information is the ECa_30 cm (i.e. Fields 2 and 3). In terms of the ECa_30 cm data, in Field 4 showing results equivalent to Field 1. The reason behind this lesser importance can be attributed to that smaller calibration sample sizes returned by cLHS, not always included the number of samples required to predict soil depth.

3.4. Maps of soil depth to petrocalcic horizon

Soil depth maps obtained from OCK for each field had similar spatial patterns with all calibration sample sizes (Fig. 6). However, the visual analysis shows that the soil depth maps produced with 10 and 20% of samples were slightly less efficient that the other sample sizes, especially in Field 4.

In general, these results suggest that (i) sampling design optimization using the cLHS algorithm and variables of auxiliary information such as elevation and ECa, leads to allocate of sampling points that can be considered as the optimum to interpolate with OCK; and (ii) precisely adjusting the percentage of samples on a 30-m grid and determining the most important variable of auxiliary information, ensure that we can obtain accurate soil depth maps at field scale (Castro Franco et al., 2015; Ramirez-Lopez et al., 2014; Schmidt et al., 2014).

3.5. Soil depth prediction accuracy

Fig. 7 shows the comparison of prediction accuracies (R^2 and *nRMSE*) between predicted soil depth values at OCK interpolation points and measured soil depth values of the validation datasets, for all fields.

OCK interpolations using variables of auxiliary information selected by the importance measure of the RF algorithm and a calibration sample size larger than 30%, were consistently the most accurate ($R^2 > 0.95$) to map soil depth to petrocalcic horizon for all fields. In addition, the median and standard error of R^2 and nRMSE for 30% of samples indicated a more stable prediction. These results also suggest that an independent validation dataset based on 50% of samples, which were not used in the OCK interpolation, is the most realistic accuracy estimate.

The linear plateau regression showed the optimal calibration sample size to map soil depth to petrocalcic horizon, where the intersections of the plateau from the regression and the slope were regarded as equivalent to the suitable percentage of samples on a 30-m grid. Based on

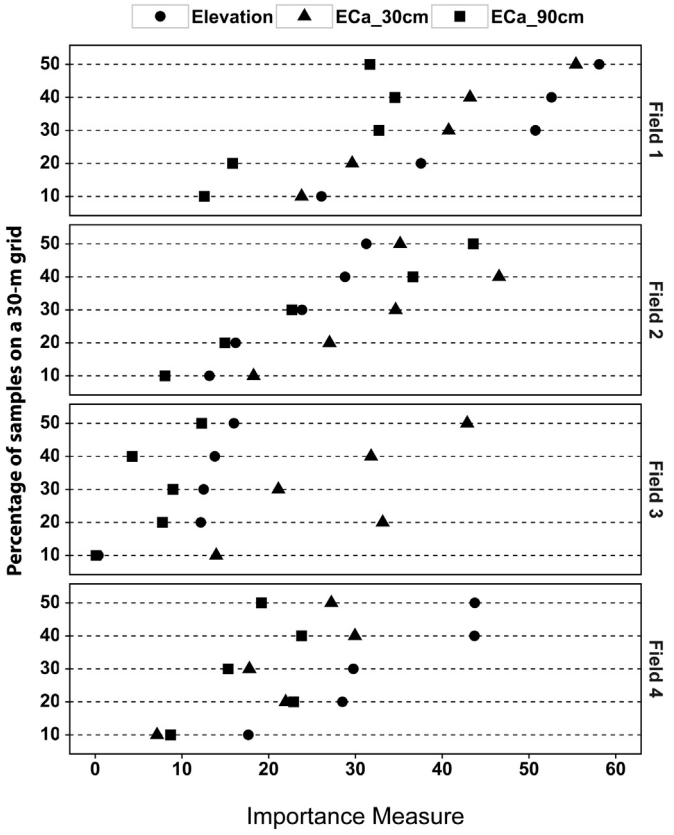

Fig. 5. Importance measure plots for ECa_30 cm, ECa_920 cm and elevation for five different percentages of soil depth samples for each field.

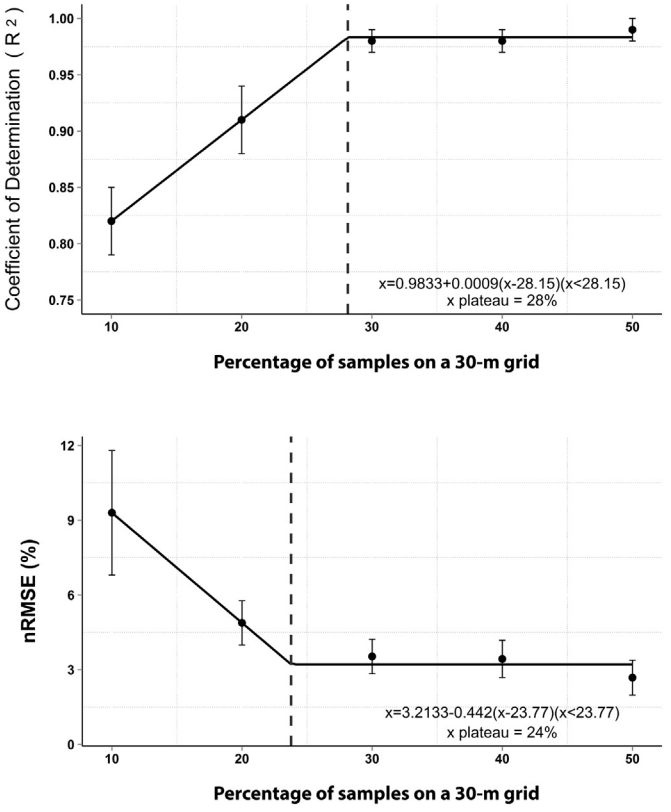

Fig. 7. Comparison of prediction accuracies (R^2 and *nRMSE*) between the predicted soil depth values at OCK interpolation points with the measured soil depth values of the validation datasets, and a linear plateau model that shows the optimal calibration sample size to map soil depth to petrocalcic horizon.

linear plateau regressions, the results suggest that the optimal calibration sample size was between 23% and 28% for all fields. These results indicate that the methodology of sampling scheme optimization proposed could reduce up to 72% the number of samples necessary to map soil depth at field scale in the southeast Pampas. Clearly, the method proposed provides a significant improvement to map soil depth over conventional methodologies and seems to be a good approach to potentially reduce cost, time and field labour.

4. Conclusions

This study presented, tested and evaluated a method to optimized sampling schemes, which used OCK interpolation with ECa or elevation selected by the importance measure of the RF algorithm, as covariable, five calibration sample sizes (based on the cLHS algorithm) and soil depth to petrocalcic horizon as target variable.

The results suggest that (i) only 30% of samples on a 30-m grid based on an appropriate soil-sampling scheme model are required to provide high prediction accuracies ($R^2 > 0.95$) for soil depth to petrocalcic horizon at field scale, in the conditions of the southeast Pampas of Argentina; (ii) an independent validation dataset based on 50% of the samples using a regular square gird, is adequate to validate the most realistic accuracy estimate; and (iii) ECa and elevation, as variables of auxiliary information, are sufficient to map soil-depth to petrocalcic horizon.

This study demonstrates that the sampling scheme optimization proposed could be successfully applied in situations where (i) ECa and elevation data are available, (ii) the area is not regularly shaped, (iii) the area has different topographic characteristics and (iv), the spatial

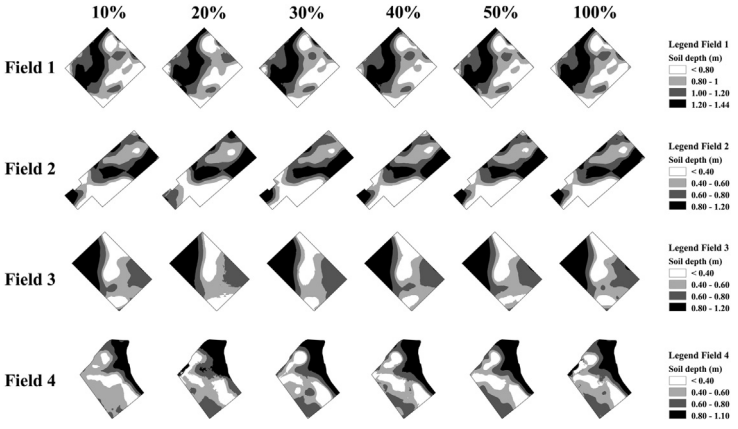

Fig. 6. Maps of soil depth to petrocalcic horizon predicted with ordinary cokriging (OCK) for each percentage of soil depth samples on a 30-m grid.

variability of the soil properties is determined for the petrocalcic horizon. This sampling optimization technique may be applied in the implementation of site-specific management and hydrological models.

The method proposed provides an improvement for soil depth mapping over conventional methods and could be a good approach to reduce cost, time and field labour. However, the extrapolation of the results within the southeast Pampas or to other areas should be tested. Also, additional studies are needed to test the performance of new variables of auxiliary information.

Conflict of interest

The authors confirm and sign that there is no conflict of interests with networks, organizations and data centers referred in the paper

Acknowledgements

This study was supported by INTA (PNSUELO-1134023), Argentina. We would like to thank José Massigoge for collecting field data, Virginia Aparicio for providing Veris® 3100 equipment and Luis Alonso for his assistance with field measurements.

References

Allred, B., Daniels, J.J., Ehsani, M.R., 2008. Handbook of Agricultural Geophysics. CRC press, Florida, USA.
Behrens, T., Zhu, A., Schmidt, K., Scholten, T., 2010. Multi-scale digital terrain analysis and feature selection for digital soil mapping. Geoderma 155 (3), 175–185.
Blanco, M.D.C., Stoops, G., 2007. Genesis of pedons with discontinuous argillic horizons in the Holocene loess mantle of the southern Pampean landscape, Argentina. J. S. Am. Earth Sci. 23 (1), 30–45.
Boettinger, J.L., Doolittle, J.A., West, N.E., Bork, E.W., Schupp, E.W., 1997. Nondestructive assessment of rangeland soil depth to petrocalcic horizon using electromagnetic induction. Arid Soil Res. Rehabil. 11 (4), 375–390.
Bork, E.W., West, N.E., Doolittle, J.A., Boettinger, J.L., 1998. Soil depth assessment of sagebrush grazing treatments using electromagnetic induction. J. Range Manag. 469–474.
Breiman, L., 2001. Random forests. Mach. Learn. 45 (1), 5–32.
Brungard, C., Boettinger, J., 2010. Conditioned Latin Hypercube Sampling: Optimal Sample Size for Digital Soil Mapping of Arid Rangelands in Utah, USA, Digital Soil Mapping. Springer, pp. 67–75.
Carroll, Z.L., Oliver, M.A., 2005. Exploring the spatial relations between soil physical properties and apparent electrical conductivity. Geoderma 128 (3–4), 354–374.
Castro Franco, M., Costa, J.L., Peralta, N., Aparicio, V., 2015. Prediction of soil properties at farm scale using a model-based soil sampling scheme and random forest. Soil Sci. 180, 1–12.
Corwin, D.L., Lesch, S.M., 2003. Application of Soil Electrical Conductivity to Precision Agriculture. Agron. J. 95, 455–471.
Corwin, D.L., Lesch, S.M., 2005. Characterizing soil spatial variability with apparent soil electrical conductivity: I. Survey protocols. Comput. Electron. Agric. 46 (1–3), 103–133.
Costa, J., Aparicio, V., Cerdà, A., 2015. Soil physical quality changes under different management systems after 10 years in the Argentine humid pampa. Solid Earth 6 (1).
Diggle, P.J., Ribeiro, P.J., 2007. Model Based Geostatistics. Springer Series in Statistics, New York.
Genuer, R., Poggi, J.-M., Tuleau-Malot, C., 2010. Variable selection using random forests. Pattern Recogn. Lett. 31 (14), 2225–2236.
Grimm, R., Behrens, T., Märker, M., Elsenbeer, H., 2008. Soil organic carbon concentrations and stocks on Barro Colorado Island — digital soil mapping using random forests analysis. Geoderma 146 (1–2), 102–113.
Grömping, U., 2009. Variable importance assessment in regression: linear regression versus random forest. Am. Stat. 63 (4).
Hengl, T., Rossiter, D.G., Stein, A., 2004. Soil sampling strategies for spatial prediction by correlation with auxiliary maps. Soil Research 41 (8), 1403–1422.
Jenny, H., 1941. Factors of soil formation, A System of Quantitative Pedology. McGraw-Hill Book Company, New York, London.
Kumhálová, J., Kumhála, F., Kroulík, M., Matějková, Š., 2011. The impact of topography on soil properties and yield and the effects of weather conditions. Precis. Agric. 12 (6), 813–830.
Liaw, A., Wiener, M., 2002. Classification and Regression by Random Forest. R News. 2(3) pp. 18–22.
McCutcheon, M.C., Farahani, H.J., Stednick, J.D., Buchleiter, G.W., Green, T.R., 2006. Effect of soil water on apparent soil electrical conductivity and texture relationships in a dryland field. Biosyst. Eng. 94 (1), 19–32.
Minasny, B., McBratney, A.B., 2005. The Matérn function as a general model for soil variograms. Geoderma 128 (3–4), 192–207.
Minasny, B., McBratney, A.B., 2006. A conditioned Latin hypercube method for sampling in the presence of ancillary information. Comput. Geosci. 32 (9), 1378–1388.
Moral, F.J., Terrón, J.M., Silva, J.R.M.D., 2010. Delineation of management zones using mobile measurements of soil apparent electrical conductivity and multivariate geostatistical techniques. Soil Tillage Res. 106 (2), 335–343.
Mulder, V., de Bruin, S., Schaepman, M.E., 2013. Representing major soil variability at regional scale by constrained Latin hypercube sampling of remote sensing data. Int. J. Appl. Earth Obs. Geoinf. 21, 301–310.
Myers, D.B., Kitchen, N.R., Sudduth, K.A., Grunwald, S., Miles, R.J., Sadler, E.J., Udawatta, R.P., 2010. Combining proximal and penetrating soil electrical conductivity sensors for high-resolution digital soil mapping. In: Viscarra Rossel, R.A., McBratney, A.B., Minasny, B. (Eds.), Proximal Soil Sensing. Progress in Soil Science. Springer, Netherlands, pp. 233–243.
Pang, S., Li, T.-X., Wang, Y.-D., Yu, H.-Y., Li, X., 2009. Spatial interpolation and sample size optimization for soil copper (Cu) investigation in cropland soil at county scale using cokriging. Agric. Sci. China 8 (11), 1369–1377.
Pazos, M.S., Mestelan, S.A., 2002. Variability of depth to Tosca in Udolls and soil classification, Buenos Aires Province, Argentina. Soil Sci. Soc. Am. J. 66 (4), 1256–1264.
R Development Core Team, 2015. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Viena, Austria.
Rad, M.R.P., Toomanian, N., Khormali, F., Brungard, C.W., Komaki, C.B., Bogaert, P., 2014. Updating soil survey maps using random forest and conditioned Latin hypercube sampling in the loess derived soils of northern Iran. Geoderma 232, 97–106.
Ramirez-Lopez, L., Schmidt, K., Behrens, T., van Wesemael, B., Demattê, J.A., Scholten, T., 2014. Sampling optimal calibration sets in soil infrared spectroscopy. Geoderma 226, 140–150.
Roudier, P., Beaudette, D., Hewitt, A., 2012. A Conditioned Latin Hypercube Sampling Algorithm Incorporating Operational Constraints, Digital Soil Assessments and Beyond. pp. 227–231.
Sadras, V.O., Calviño, P.A., 2001. Quantification of grain yield response to soil depth in soybean, maize, sunflower, and wheat. Agron. J. 93 (3), 577–583.
Schmidt, K., Behrens, T., Daumann, J., Ramirez-Lopez, L., Werban, U., Dietrich, P., Scholten, T., 2014. A comparison of calibration sampling schemes at the field scale. Geoderma 232, 243–256.
Shaner, D.L., Khosla, R., Brodahl, M.K., Buchleiter, G.W., Farahani, H.J., 2008. How well does zone sampling based on soil electrical conductivity maps represent soil variability? Agron. J. 100 (5), 1472–1480.
Simbahan, G.C., Dobermann, A., 2006. Sampling optimization based on secondary information and its utilization in soil carbon mapping. Geoderma 133 (3), 345–362.
Soil Survey Staff, 2014. Keys to Soil Taxonomy. 12th ed. United States Department of Agriculture - Natural Resources Conservation Service, Washington, DC.
Stieglitz, M., Shaman, J., McNamara, J., Engel, V., Shanley, J., Kling, G.W., 2003. An approach to understanding hydrologic connectivity on the hillslope and the implications for nutrient transport. Glob. Biogeochem. Cycles 17 (4), 1105.
Swinton, S.M., Lowenberg-Deboer, J., 2001. Global adoption of precision agriculture technologies: who, when and why. Proceedings of the 3rd European Conference on Precision Agriculture. Citeseer, pp. 557–562.
Tesfa, T.K., Tarboton, D.G., Chandler, D.G., McNamara, J.P., 2009. Modeling soil depth from topographic and land cover attributes. Water Resour. Res. 45 (10).
Tesfa, T.K., Tarboton, D.G., Chandler, D.G., McNamara, J.P., 2010. A generalized additive soil depth model for a mountainous semi-arid watershed based upon topographic and land cover attributes. In: Boettinger, J., Howell, D., Moore, A., Hartemink, A., Kienast-Brown, S. (Eds.), Digital Soil Mapping. Progress in Soil Science. Springer, Netherlands, pp. 29–41.
Thornthwaite, C.W., 1948. An approach toward a rational classification of climate. Geogr. Rev. 55–94.
Vašát, R., Heuvelink, G.B.M., Borůvka, L., 2010. Sampling design optimization for multivariate soil mapping. Geoderma 155 (3–4), 147–153.
Wang, K., Zhang, C., Li, W., 2013. Predictive mapping of soil total nitrogen at a regional scale: a comparison between geographically weighted regression and cokriging. Appl. Geogr. 42, 73–85.
Ziadat, F.M., 2010. Prediction of soil depth from digital terrain data by integrating statistical and visual approaches. Pedosphere 20 (3), 361–367.