Geomorphology 220 (2014) 58–67

Contents lists available at ScienceDirect

Geomorphology

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

Geostatistical approach for identifying scale-specific correlations between soil thickness and topographic attributes

Hocine Bourennane [a,*], Sébastien Salvador-Blanes [b], Alain Couturier [a], Caroline Chartin [b], Catherine Pasquier [a], Florent Hinschberger [b], Jean-Jacques Macaire [b], Joël Daroussin [a]

[a] INRA, UR0272 Unité de Science du Sol, 2163 Avenue de la Pomme de Pin CS 40001 Ardon, F-45075 Orléans Cedex 2, France
[b] Université François-Rabelais de Tours, EA 6293 GéHCO, Faculté des Sciences et Techniques, Parc de Grandmont, F-37200 Tours, France

ARTICLE INFO

Article history:
Received 17 June 2013
Received in revised form 15 May 2014
Accepted 28 May 2014
Available online 5 June 2014

Keywords:
Factorial kriging analysis
Soil thickness
Topographic attributes
Partial least square regression
Extrapolation
LiDAR

ABSTRACT

This paper investigates how the spatial correlations between topographic attributes and a soil thickness can be improved by focusing on the relationships between them at specific spatial scales. In addition, this paper examines the effects of the topographic attribute data sources that are used as explanatory variables for modeling the response variable, and considers the possibility of model extrapolation for mapping beyond the area where the model was established. Here, factorial kriging analysis (FKA) and partial least square regression (PLSR) analysis are used to separate nuggets and small- and large-scale structures in data including four topographic attributes and soil thickness (*ST*). These analyses were conducted at different scales to analyze the relationships between *ST* and the selected topographic attributes in the southwest region of the Parisian Basin. The structural correlation coefficients from the FKA show strong correlations between the variables. These correlations, which change as a function of spatial scale, are not revealed by the linear correlation coefficients. The Eigen vectors from the principal component analysis that was performed on the small-scale and large-scale structures of the linear co-regionalization model are used to obtain *ST* and the topographic attributes at both spatial scales over the study area. The *ST* models are built as a function of topographic attributes using PLSR. Results have shown that the models built using variables that were assessed at a specific scale are better at predicting the target variable than models that were built using raw data. Regarding the models that were built using raw data, the structural correlations that occur at different spatial scales are merged together and the variance–covariance matrix of the nugget that represents data noise is not filtered out. Measures of model performance that are based on a validation data set have shown that the model based on small-scale structure (Model-S) is better for predicting soil thickness than the model based on large-scale structure (Model-L). The effects of topographic attribute data sources as explanatory variables for modeling *ST* are less significant than the effects of the two models for mapping. Moreover, extrapolation of the model-S beyond the area where it was generated is appropriate. The decomposition process is associated with a modeling approach, such as the PLSR, which accounts for the collinearity between predictor variables and leads to an efficient prediction model. These results are important for modeling soil properties based on topographic attributes and for spatially generalizing models that have been established over small to large areas. Thus, in the presence of nested variogram models, the correlations between variables of interest and auxiliary information should be improved by filtering out some of the spatial structures by factorial kriging. The information filtered is associated with an appropriate approach for modeling when collinearity occurs between the predictor variables and provides a suitable model for predicting and spatially generalizing a locally established model.

1. Introduction

Topographic attributes like slope gradient are commonly used by soil surveyors to delineate soil types in the landscape and to model the spatial distribution of various soil properties (e.g., Moore et al., 1993; Bell et al., 1994; Knotters et al., 1995; Bourennane et al., 1996; King et al., 1999; McBratney et al., 2003; Lagacherie et al., 2007; Debella-Gilo and Etzelmüller, 2009; Kim and Zheng, 2011). However, good correlations that are qualitatively described in the field between certain soil properties and topographic attributes may not be reproduced by quantitative data because the two types of information can be measured at supports with various sizes (e.g., Stevenson et al., 2010; Kim and Zheng, 2011). Therefore, topographic attributes should be computed at an appropriate scale to represent a particular process that occurs in soils. If these scale effects are not considered, the computed attributes

* Corresponding author. Tel.: +33 2 38 41 48 28; fax: +33 2 38 41 78 69.
E-mail address: Hocine.Bourennane@orleans.inra.fr (H. Bourennane).

http://dx.doi.org/10.1016/j.geomorph.2014.05.026
0169-555X/

may be meaningless and the processes of interest may be masked. Therefore, it is important to determine the appropriate scale for analyzing relationships between soil variables and topographic attributes.

Wide use of digital elevation models (DEMs) for modeling environmental processes has resulted in several research papers regarding the following topics: (1) digital elevation data sources, (2) DEM accuracy, (3) algorithms for deriving topographic attributes, and (4) obtaining optimal DEM resolution (e.g., Desmet and Govers, 1996; Wilson et al., 2000; Thompson et al., 2001; Claessens et al., 2005; Erskine et al., 2006; Smith et al., 2006; Wechsler, 2007; Li et al., 2011; Shi et al., 2012). In contrast, this paper focuses on removing noise from DEMs to extract relevant information and determine an appropriate scale for relating topographic attributes to soil properties.

The main objective of this paper is to investigate how the spatial relationships between a soil variable and topographic attributes can be improved by obtaining stronger correlations between them at specific spatial scales. In addition, the following aspects are examined: (i) the effects of the data source used to derive the topographic attributes as explanatory variables for modeling the response variable and (ii) the extrapolation of a soil thickness variation model beyond the area where it was established. Factorial kriging analysis (FKA) was used to separate the data at different scales to analyze the relationships between soil thickness (*ST*) and the topographic attributes using partial least square regression (PLSR).

The main benefits of FKA relative to traditional spatial smoothers (e.g., median filtering) and classical multivariate data analysis (e.g., principal component analysis: PCA) are that FKA filters noise from the data and can be used to decompose structured data components into several spatial components (i.e., local versus regional variability) based on semivariogram models. For example, important features like the range of spatial correlation are not accounted for by classical PCA. In addition, filtering data noise is an important issue. Thus, it would be useful to decompose the structured variability according to the corresponding spatial scale rather than by using a filtering algorithm to eliminate hotspots. Filtration and decomposition could both be reached using FKA. Furthermore, an approach that accounts for collinearity between the predictor variables (PLSR) is considered here for modeling the target variable.

Several authors (e.g., Goovaerts and Webster, 1994; Bourennane et al., 2003, 2012; Castrignano et al., 2012; Muñoz and Kravchenko, 2012) have addressed the scale issue previously using FKA. These authors have shown that weak correlations for a given spatial scale can mask actual correlations between variables. Thus, this paper investigates modeling at specific scales and uses developed approaches to extrapolate the model beyond its established area. Furthermore, this study attempts to confirm previous results regarding the usefulness of decomposing the structured variability and filtering noise to extract relevant features from data and strengthen the correlations between variables.

2. Study area and data

This study was conducted over a 17 ha (Fig. 1) southeast-facing hillslope near the village of Seuilly (in the southwestern region of the Parisian Basin, 47°08.31′N, 0°10.97′E). This site is part of a network of French sites where soil erosion has been studied and the soils have been sampled frequently. In addition, this area represents a calcareous landscape of Western Europe. The elevation of the study area varies from 37 to 80 m, and the slope length is 750 m.

Accurate coordinates and elevations were obtained at 1550 points by post-treatment (differential correction using GPS Pathfinder ® Office) of data that were recorded by DGPS (Trimble ® ProXRS), which was used as a base station and as a mobile recorder. Next, the DEM was estimated on a grid at a resolution of 2 m. Details regarding the methods and parameters that were used to generate the DEM are presented by Chartin et al. (2011). Finally, four topographic attributes, slope gradient (S), curvature (C), plan curvature (C_l) and profile curvature (C_r), were derived from the DEM by using the algorithms (Zeverbergen and Thorne, 1987; Moore et al., 1991) implemented in ArcGIS 9.3.1.

Soil thickness was measured with a manual auger according to two sampling schemes. The first sampling scheme measured soil thickness over well-described lynchets and undulations (Bolline, 1971; Macaire et al., 2002; Salvador-Blanes et al., 2006; Chartin et al., 2011) across the study area. These features were easily identified in the field. In contrast, the second sampling scheme measure soil thickness across the study area by randomly considering the soil samples from each 25 × 25 m square within the grid. Both sampling designs included 734 samples. Twenty percent of the observations (148 samples) were randomly selected and used as the validation set. The remaining 80% of the dataset (586 samples) was used as the prediction set of *ST*. Ordinary kriging was used throughout the study area (17 ha) with a 2 m resolution grid that was consistent with the DEM and its derived attributes. For *ST* kriging, an experimental variogram was computed and fit by a model using the weighted least squares method. Cross validation was used on the original data to validate the variogram model. Every known point was estimated by using the surround points, but not the point itself. After performing these calculations, the mean error should

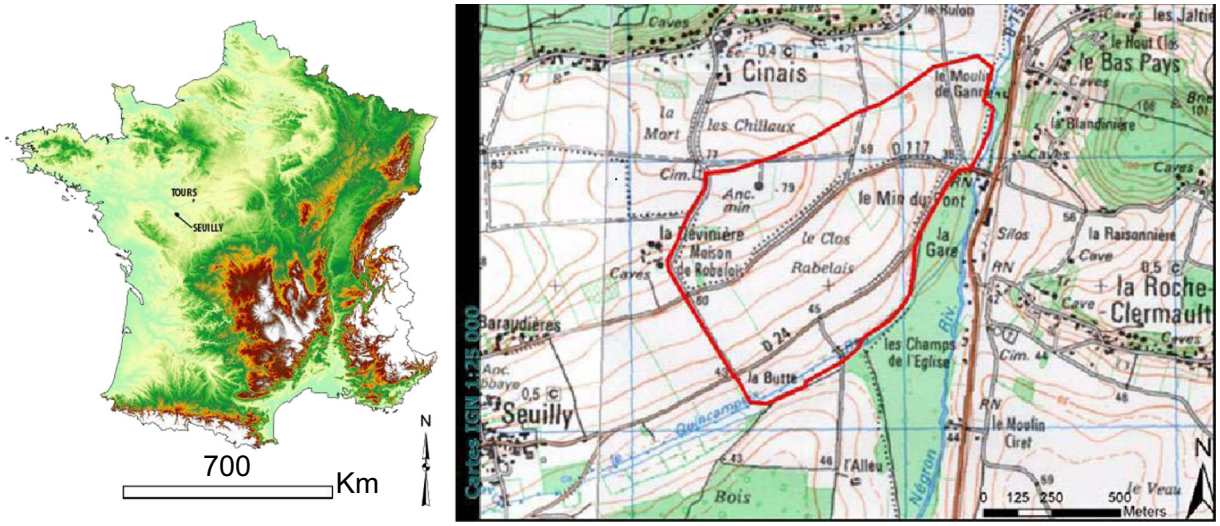

Fig. 1. The location of the study area within France (source of the maps: SRTM 90 m Digital Elevation Data and I.G.N. in 1988 and 2005).

be near zero and the ratio of the mean squared error to the kriging variance should be near unity.

3. Methods

3.1. Factorial kriging analysis (FKA)

A geostatistical method such as FKA can be used to decompose raw variables into several components according to different spatial scales. Consequently, modeling could be performed for the values of the predictor variables that are strongly correlated with the response variable. In addition, other methods could be considered, such as the explicit use of residuals in the statistical model. However, comparing the approaches that address these correlated predictor variables is beyond the scope of this study.

FKA isolates and displays variation sources that act at different spatial scales with different correlation structures. The theory that underlies FKA has been described in several publications (e.g., Goovaerts, 1997; Wackernagel, 1998). Below, we summarize the major steps of this geostatistical technique.

Methods based on second-order distribution moments (such as the FKA) are sensitive to skewed data (e.g., Goovaerts, 1997; Chilès and Delfiner, 1999). Therefore, in FKA, raw to Gaussian transformations were conducted before data analysis. Finally, back-transformation (Gaussian to raw) to the original unit was performed to validate and present the results.

FKA begins by analyzing the co-regionalization of a set of variables to define a linear co-regionalization model (LMC). The $p\,(p+1)/2$ experimental direct and cross variograms of the p variables require prior modeling that uses the linear combination of the same set of variograms that are standardized to a unit sill $g^u(\boldsymbol{h})$. The $\boldsymbol{h}$ parameter represents the vector (lag) that separates any pair of measurements made at locations u_α and $u_{\alpha+\boldsymbol{h}}$, and $g^u(\boldsymbol{h})$ represents the different variogram functions considered in the LMC. Thus, for any couple of variables i and j, the variogram γ_{ij} takes the following form:

$$\gamma_{ij}(\boldsymbol{h}) = \sum_{u=1}^{N_s} b_{ij}^u g^u(\boldsymbol{h}), \tag{1}$$

where b_{ij}^u are the coefficients that must be determined by the data. Using matrix notation, the LMC can be rewritten as:

$$\Gamma(\boldsymbol{h}) = \sum_{u=1}^{Ns} \boldsymbol{B}^u g^u(\boldsymbol{h}), \tag{2}$$

where $\Gamma(\boldsymbol{h})$ is a $p \times p$ symmetric matrix whose diagonal and off-diagonal elements are the direct and cross-variogram values, respectively, for a given lag $\boldsymbol{h}$. $\boldsymbol{B}^u$ is the $p \times p$ symmetric matrix of the coefficients b_{ij}^u, and is referred to as the co-regionalization matrix.

To ensure that the variances of the finite linear combinations of the random functions are positive, the iterative procedure developed by Goulard (1989) was used. Thus, the best LMC, regarding the weighted residual sum of squares (Goulard and Voltz, 1992), was chosen by comparing the goodness of fit for several combinations of $g^u(\boldsymbol{h})$ functions with different ranges.

Each variogram function in $g^u(\boldsymbol{h})$ indicated the spatial variance for the individual variable measurements and the spatial covariance between the measurements for a pair of variables over a given range. This range defines a spatial scale. Thus, for a given variogram function of LMC, we can estimate the relationship between a pair of variables at a particular spatial scale given by the range.

The second step consists of an analysis of the structural correlation coefficients, which is conducted using each co-regionalization matrix $\boldsymbol{B}^u$. The values of each $\boldsymbol{B}^u$ describe the relationships between the chosen variables at the particular spatial scale that is defined by the basic variogram function of $g^u(\boldsymbol{h})$. However, the structural correlation coefficient r_{ij}^u, defined as follows:

$$r_{ij}^u = \frac{b_{ij}^u}{\sqrt{b_{ii}^u b_{jj}^u}}, \tag{3}$$

is more revealing as it is a unit-free measure of correlation between any two variables at different spatial scales and is defined when modeling the co-regionalization. However, the value of r_{ij}^u depends entirely on modeling the co-regionalization between each pair of variables.

The final step in FKA consists of a principal component analysis (PCA), which is applied to the co-regionalization matrices. These co-regionalization matrices are the variance–covariance matrices that describe the correlation structure of a set of variables at different spatial scales. Unlike the PCA that is performed for a classical variance–covariance matrix, the PCA performed for the co-regionalization matrices yields sets of spatial components (regionalized factors) for each spatial scale u. Accordingly, the Eigen vectors from the PCA of small-scale and large-scale structures can be used to infer input data for modeling at different spatial scales over a study area.

3.2. Partial least square regression (PLSR)

PLSR was used to model the response variable as a function of topographic attributes at different spatial scales. The underlying theory of PLSR has been described in several statistical textbooks and papers (e.g., Wold et al., 1984; Höskuldsson, 1988; Tenenhaus, 1998). Here, we describe the major steps for the PLSR algorithm.

PLSR can be described as a generalized multiple linear regression (Gerlach et al., 1979). However, in contrast to multiple linear regression, PLSR can analyze data that are collinear, noisy, and have numerous ***X***-variables (***X***: a set of predictor variables). Moreover, PLSR can simultaneously model several response (***Y***) variables (***Y***: a set of response variables).

The goal of PLSR is to predict values of ***Y*** from values of ***X*** and to describe their common structure. This goal could be achieved using ordinary multiple regression. However, when the number of predictors is large, ***X*** is likely singular and the regression approach is no longer feasible (i.e., due to multicollinearity). Several approaches have been developed to cope with this problem. One approach is to eliminate some predictors, e.g., using stepwise methods. Another approach is called principal component regression, which is used to perform a PCA of the ***X*** matrix before using the principal components of ***X*** as regressors for ***Y***. The orthogonality of the principal components eliminates the multicollinearity problem. However, choosing an optimum subset of predictors remains problematic. One possible strategy is to keep only a few of the first components. However, the latter components are chosen to explain ***X*** rather than ***Y***. Thus, nothing guarantees that the principal components (which explain ***X***) are relevant for ***Y***.

In contrast, PLSR finds components from ***X*** that are also relevant for ***Y***. Specifically, PLSR searches for a set of components (called latent vectors) that simultaneously decompose ***X*** and ***Y*** with the constraint that these components explain as much of the covariance between ***X*** and ***Y*** as possible. This step corresponds to generalizing the PCA. In addition, this step is followed by a regression step where the decomposition of ***X*** is used to predict ***Y***.

Cross validation is used to determine the number of significant PLSR components (e.g., Tenenhaus, 1998). With cross validation, several observations are placed aside during model development. The response variable for these unused observations is predicted by the model and compared with the actual values. This procedure is repeated several times until every observation has been placed aside once. The prediction error sum of squares (*Press*) is the squared differences between the observed and predicted values when the observations are placed aside. Based on the *Press*, Q^2 (the fraction of the total variation of the

dependent variables that can be predicted by a component) and Q^2_{cum} (cumulative Q^2) can be calculated as follows:

$$Q^2 = 1.0 - Press/SS \quad (4)$$

$$Q^2_{cum} = 1.0 - \prod \left(\frac{Press}{SS}\right)_a, \quad (5)$$

where $a = 1, 2, \ldots k$ and SS are the residual sum of squares for the previous dimension, $\prod \left(\frac{Press}{SS}\right)_a$, which is the product of $Press/SS$ for each individual component a.

The tested PLSR component is significant when $Press/SS \leq 0.95^2$ or $Q^2 \geq (1 - 0.95^2) = 0.0975$. The model is considered to have a good predictive ability when Q^2_{cum} is greater than 0.5 (Tenenhaus, 1998).

The variable importance in the projection (VIP) is a parameter that shows the importance of a variable (a predictor variable) in the PLSR model. The methods for calculating the VIP are presented by Tenenhaus (1998). The first PLSR analysis with all predictor variables was performed using these methods. Next, the variable with the lowest VIP value was eliminated and the PLSR analysis was performed again. This procedure was repeated until only two variables remained in the PLSR model. Finally, the obtained PLSR models that had the highest Q^2_{cum} values and the fewest predictor variables were selected as the optimal models.

3.3. Model performance

The performance of the models that were obtained through PLSR and the variables that were inferred at different spatial scales were examined regarding their prediction abilities. The *ST* was estimated at each point of the validation set (148 sites of the validation set) by the previously mentioned models. A scatter plot of the measured versus predicted *ST* values at each validation site was created and the mean error (*ME*) and root mean square error (*RMSE*) were calculated.

3.4. Effects of the data sources of the predictor variables when modeling the response variable: the developed approach

The validity of the established models regarding the source of the DEM that was used to derive the predictor variables was examined by considering a DEM established from airborne Light Detection and Ranging (LiDAR). The LiDAR was conducted over an area of 3 km^2 that included the 17 ha described in Section 2 and was used to establish the *ST* spatial variation models.

Two scenarios were tested to examine the validity of the models that were developed by the approaches described in Subsections 3.1 and 3.2. These scenarios used predictor variables that were derived from LiDAR rather than DGPS.

The first scenario entailed selecting topographic attributes within the 17 ha area based on the LiDAR when these values are in the same range as the topographic attributes that were derived from the DEM and established by DGPS. After selection, the models were applied to determine the soil thickness across the 17 ha area. The soil thickness maps that were obtained from the first scenario were called Model-S-LiDAR-Strict and Model-L-LiDAR-Strict. In the first scenario, *ST* was not inferred for many pixels because the selection of topographic attributes was conditioned by the range of topographic attributes that were derived from the DEM and established by DGPS.

In the second scenario, the models were applied over the same area (17 ha), but without constrains regarding the topographic attributes that were derived from the LiDAR. In this case, the inferred soil thickness maps were called Model-S-LiDAR-Large and Model-L-LiDAR-Large. In addition, *ST* was estimated for all pixels in the 17 ha area. The four maps that resulted from the two scenarios were evaluated using a validation set and the criteria mentioned in Subsection 3.3.

3.5. Extrapolation approach and data

The extrapolation approach consists of the following steps: (i) selecting a model among the models that were developed by using the predictor variables derived from the DEM and established by the DGPS measurements on the 17 ha area, and (ii) applying the selected model to an area of 3 km^2 where the predictor variables are derived from a LiDAR survey. Data from 231 individuals were used to measure the performances of the model for extrapolation beyond the area where it was generated. This data set consisted of 148 individuals from the validation set across the 17 ha area and an additional 83 locations where *ST* was measured specifically for validation. Here, the criteria mentioned in Subsection 3.3 were used to measure model performance in extrapolation.

4. Results and discussion

4.1. Spatial estimates and modeling of ST

The measured *ST* values varied from 0.22 to 1.85 m (Table 1: first line) with a mean of 0.60 m and a standard deviation (S.D.) of 0.31 m. *ST* was mapped (Fig. 2c) by ordinary kriging on a regular grid at a resolution of 2 m across the entire study area. The spatial autocorrelation of *ST* (Fig. 2b), which was quantified through the semivariogram, indicated that the experimental variogram was not flat and presented a sill for a distance of at least 400 m. Overall, 586 data points (Fig. 2a) were used to estimate the experimental variogram with a lag size of 22 m. The number of pairs used to compute the average semivariance per lag varied from 715 pairs to 13,013 pairs for the first and sixth lags, respectively. The average semivariance for the last lag was computed from 4339 pairs. The tolerance for lag distance was 50% of the lag size. The nested model (the nugget plus a Gaussian and spherical model) that was fit to the experimental variogram is relevant. The cross validation results indicated that the mean error was nearly zero (−0.002) and the ratio of the mean squared error to the kriging variance was nearly 1 (1.011). Various models were tested to fit the experimental model. Here, we note the model that provided the most accurate prediction criteria. For example, the model composed of a nugget and a spherical model provided a mean error of −0.004 with a mean square error to kriging variance error ratio of 0.937.

Table 1 summarizes the statistics obtained from the raw variables used to model *ST* as a function of four topographic attributes (*S*, *C*, C_l, and C_r). PLSR was used to model *ST* throughout the study area by using these attributes as predictor variables. This analysis was conducted using 44,052 observations and resulted in a model that only explained 17% ($R^2 = 0.17$) of the total *ST* variance (Table 2). In addition, the standardized coefficient values of the PLSR model obtained on raw data (Table 2) indicated that C_r and *C* were not significant terms. The standard errors of the C_r and *C* coefficients were similar to the coefficients themselves. From a statistical standpoint, this result was expected because the Pearson's correlation coefficients (Table 3) indicated weak correlations between *ST* and the topographic attributes (predictor variables) and strong correlations between some topographic attributes. These

Table 1
Summary statistics of soil thickness and several topographic attributes that were derived from a DEM. g_1: skewness.

Variable	Unit	Count	Mean	Std	Min	Max	g_1
Soil thickness: prediction set	m	586	0.60	0.31	0.22	1.85	1.48
Soil thickness: validation set 1	m	148	0.70	0.39	0.25	2.23	1.38
Soil thickness: validation set 2	m	83	1.03	0.58	0.21	2.30	0.74
Soil thickness: spatial estimate	m	44,052	0.61	0.26	0.11	1.72	1.27
Curvature from DEM (C)	m^{-1}	44,052	0.00	0.33	−3.35	4.38	3.49
Plan-curvature from DEM (C_l)	m^{-1}	44,052	0.00	0.12	−1.83	1.31	−1.22
Profile-curvature from DEM (C_r)	m^{-1}	44,052	0.00	0.27	−3.82	3.35	−4.08
Slope gradient from DEM (S)	%	44,052	2.85	1.46	0.03	6.93	0.51

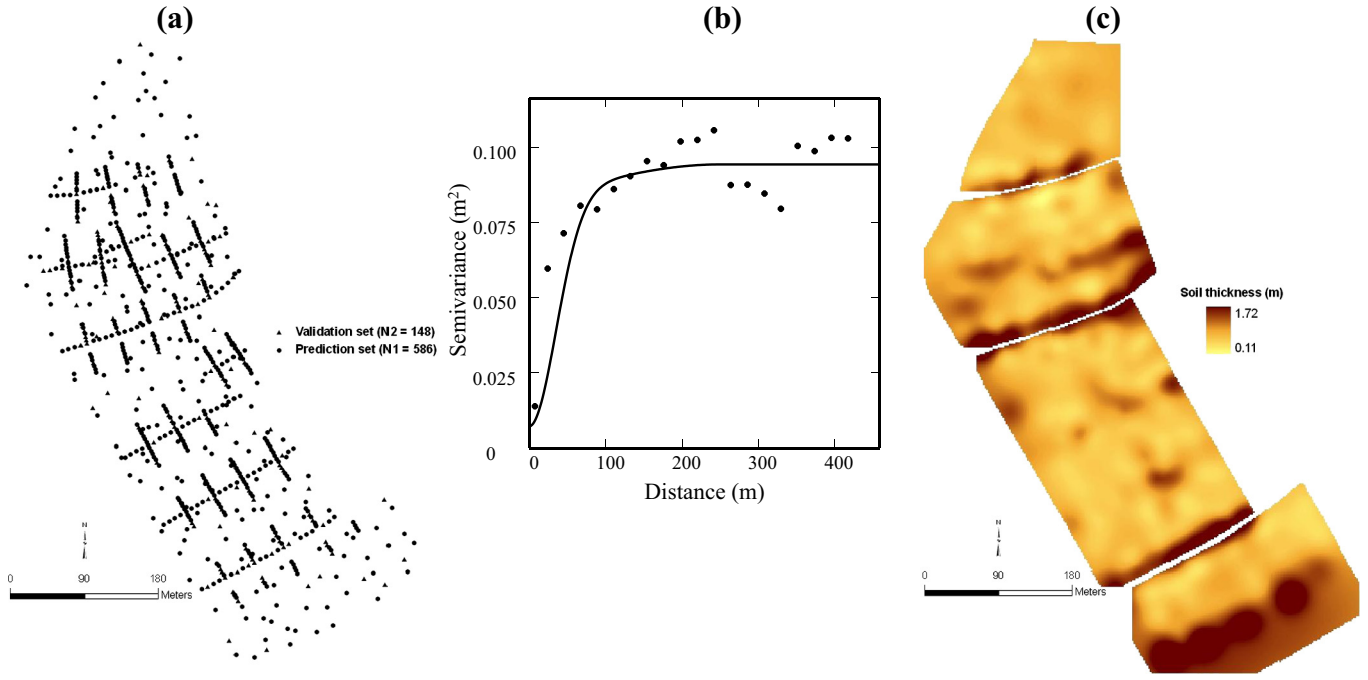

Fig. 2. Mapping soil thickness (*ST*) over the entire study area. (a) The global sampling pattern of *ST*. (b) The experimental variogram of *ST* (dots) with the theoretical model fits (solid line). (c) *ST* estimated by ordinary kriging using the prediction set.

statistical results contradict results from previous studies that were conducted in this area (e.g., Bellemlih, 1999; Chartin et al., 2011). In addition, these results contradict the survey that was conducted within this study. However, both studies mentioned that the soil distribution and thickness are mainly elucidated by topographic attributes.

4.2. Co-regionalization analysis of ST and topographic attributes

The analysis of the variograms and cross-variograms (not shown) suggested the presence of three basic structures at different spatial scales for the five transformed variables (*ST*, *S*, *C*, C_r and C_l). In addition to the nugget term, the two elementary variogram functions that were used for the co-regionalization model were spherical models with ranges of 90 and 150 m. These models were called small-scale and large-scale structures, respectively.

The linear correlation coefficient did not reveal actual relationships among the variables (Table 3: Pearson correlation coefficient) because it averaged the distinct changes in the correlation structures that occurred at different spatial scales and included the measurement errors that were inherent in the nugget effect. Thus, filtering the different components resulted in strong correlations between the variables (Table 3: small-scale and large-scale structures) that changed as a function of the spatial scale. For example, Table 3 shows a weak linear correlation between C_r and *S* ($R = 0.01$). Nevertheless the structural correlation coefficients (Table 3: small-scale and large-scale structures) were much greater between these two variables once the nugget effect was filtered out. Likewise, the large correlation between *ST* and *S* in the small-scale structure (−0.90) was hidden by the lack of correlation in the large-scale structure (Table 3: large-scale structure).

The coefficients of the small-scale and large-scale structures were used to perform two distinct principal component analyses (PCA). The two first components (Fig. 3) accounted for more than 86% of the total variance in the matrix. Fig. 3a shows that the first component for the small-scale structure was positively correlated with C_r and *ST* and negatively correlated with *S*, *C*, and C_l.

Similarly, the first component for the large-scale structure was strongly and positively correlated with *S* and C_l (Fig. 3b). In contrast, *ST* and *C* contributed more to the second component. However, C_r was positively related to both components. Comments regarding the nugget scale are omitted here due to the measurement errors that were inherent at the nugget spatial scale. In geostatistics, the nugget structure

Table 2
Standardized coefficient values of the PLSR model when using raw data. *C*: curvature; C_l: plan curvature; C_r: profile curvature; and *S*: slope gradient.

Variable	Coefficient	Standard error	The lower limit of the 95% confidence interval	The upper limit of the 95% confidence interval
C	−0.038	0.021	−0.079	0.003
C_l	−0.052	0.008	−0.067	−0.037
C_r	0.022	0.024	−0.025	0.070
S	−0.401	0.102	−0.600	−0.201

$R^2_{ajusted} = 0.17$.

Table 3
Linear correlation coefficients and structural correlation coefficients. *C*: curvature; C_l: plan curvature; C_r: profile curvature; *S*: slope gradient; and *ST*: soil thickness.

	C	C_l	C_r	*S*
(a) Pearson correlation coefficient				
C_l	0.60			
C_r	−0.93	−0.27		
S	0.01	0.05	0.01	
ST	−0.04	−0.05	0.02	−0.42
(b) Small-scale structure (Spherical: 90 m)				
C_l	0.65			
C_r	−0.95	−0.37		
S	0.50	0.13	−0.57	
ST	−0.21	−0.20	0.21	−0.90
(c) Large-scale structure (Spherical: 150 m)				
C_l	0.29			
C_r	−0.56	0.62		
S	−0.38	0.44	0.61	
ST	−0.72	−0.11	0.57	−0.15

Number of observations: 44052.

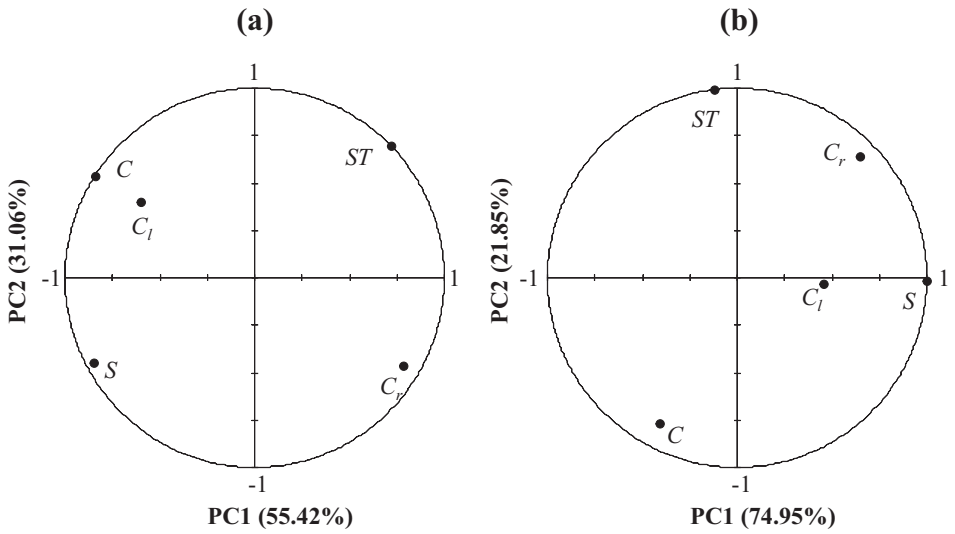

Fig. 3. Correlation circles corresponding to (a) small-scale structure and (b) large-scale structure.

represents the measurement errors and the structures that were not captured during sampling. Thus, in FKA, the variance–covariance matrix that corresponds to the nugget structure is generally omitted during analysis. The model input data were inferred for different spatial scales by using the Eigen vectors of the PCA for the small-scale structure and the PCA for the large-scale structure. Next, PLSR was used to quantify how well the topographic attributes that were decomposed into the small-scale and large-scale structures could reflect the variations in *ST* that were also decomposed into the two structures. According to the cross validation results, two PLSR components were appropriate when modeling the relationships between *ST* for the small-scale structure and the topographic attributes that were inferred from the small-scale and large-scale structures. In addition, the VIP parameter indicated that the *S* values for the small-scale structure contributed the most to the *ST* variation in the small-scale structure model, followed by C_r in the large-scale structure model. Finally, the PLSR model explained 76% of the *ST* variations for the small-scale structure and was expressed according to the unscaled regression coefficients of the predictor variables and a constant that was transformed from the PLSR results (Table 4, first line: termed Model-S in the sequel of the manuscript).

When modeling the relationships between *ST* for the large-scale structure and with topographic attributes that inferred at the small-scale and large-scale structures, the cross validation results indicated that only one PLSR component was appropriate when modeling. In addition, the VIP parameter indicated that the *S* values at the small-scale structure and the *C* values at the large-scale structure mainly contributed to model the *ST* variation for the large-scale structure. The PLSR model, which is expressed based on the predictor variables and a constant transformed from PLSR results (Table 4, last line: termed

Table 4
Standardized and non-standardized coefficient values of the PLSR models according to the spatial scale. *ST*: soil thickness; *S-short*: slope gradient at small-scale structure; C_r-*long*: profile curvature at large-scale structure; and *C-long*: curvature at large-scale structure.

Variable	Standardized coefficient	Non-standardized coefficient	R^2 (%)
(a) ST at the small-scale structure versus the topographic attributes			
Constant		3.96	76
S-short	−0.99	1.38	
C_r-*long*	−0.27	−4.82	
(b) ST at the large-scale structure versus the topographic attributes			
Constant		2.58	94
S-short	−0.68	−0.97	
C-long	−0.49	−5.07	

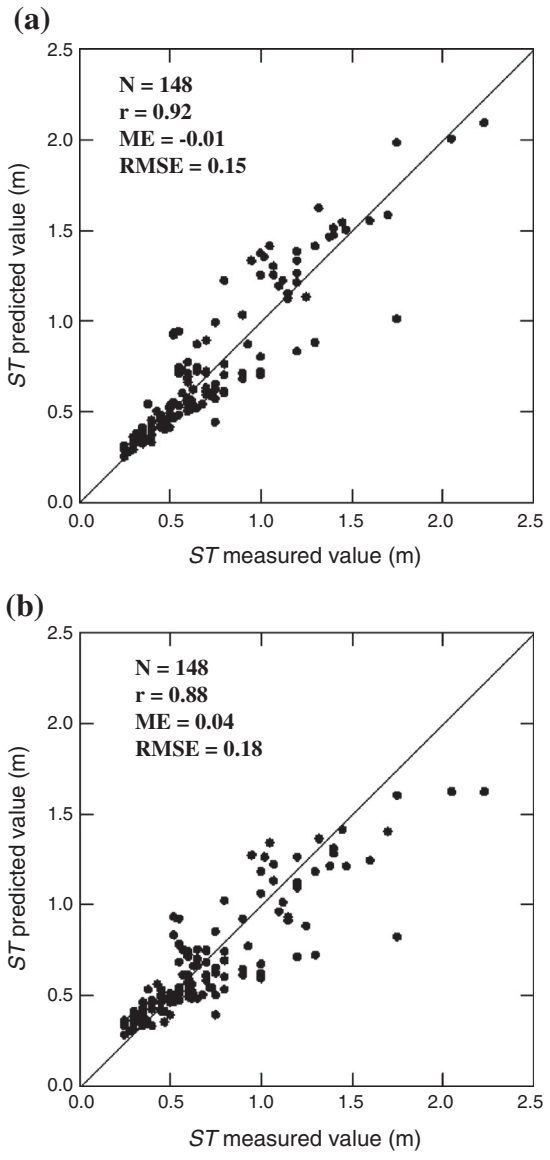

Fig. 4. The soil thickness (*ST*) values of the validation set versus the *ST* values that were inferred by (a) Model-S and (b) Model-L.

Table 5
Comparison of *ST* values inferred by the Model-S and Model-L with the *ST* values of the validation set. *ST*: soil thickness.

	N	*Min*	*Max*	*Mean*	*Std*	*ME*	*RMSE*	*R*
ST measured value (m)	148	0.25	2.23	0.70	0.39			
ST from Model-S (m)	148	0.25	2.09	0.71	0.40	−0.01	0.15	0.92
ST from Model-L (m)	148	0.28	1.62	0.66	0.32	0.04	0.18	0.88

Model-L in the sequel of the manuscript), explained 94% of the *ST* variations for the large-scale structure.

4.3. Evaluation of model performances based on the validation data set

Two sets of *ST* values were inferred from the models summarized in Table 4. These *ST* values were compared to the *ST* values in the validation set (148 points) that were collected across the 17 ha area. The differences between the models were not clear according to *ME*, *RMSE* and the correlation coefficient (*R*) (Fig. 4 and Table 5). Nevertheless, Table 5 indicates that the model based on small-scale structure analysis (Model-S) has reproduced better mean and extreme values. In addition, the variance observed in the measured values of the validation set was better reproduced in this model than in the model based on the large-scale structure analysis (Model-L). Thus, the Model-S is more appropriate for predicting soil thickness across the 17 ha study area. According to the statistics exhibited in Table 4, this result was not expected. Model-L explained more of the *ST* variability than Model-S. This finding likely resulted from the importance of the main predictor variable '*S*' in Model-S relative to Model-L (Table 4). Consequently, Model-L was potentially more sensitive to variations in the second predictor variable, *C*, relative to Model-S. In addition, Model-S should be less sensitive to the variations of the second predictor variable, C_r, due to the weight of the first predictor variable, *S*.

4.4. Validity of the models with respect to the DEM source and extrapolation

The two scenarios addressed for testing the validity of the models regarding their ability to map *ST* using explanatory variables derived from a LiDAR survey have resulted in the maps presented in Fig. 5. The validation results (Fig. 6) revealed that applying Model-S to the topographic attributes derived from LiDAR according to the first scenario (Model-S-LiDAR-Strict: Fig. 6a) resulted in better predictions than the other three situations (Fig. 6b–d). All statistical indicators listed in Fig. 6 are for Model-S-LiDAR-Strict. In addition, the analysis of variance (ANOVA) that was conducted on the residuals of each case (Table 6) showed significantly influence of the model used on the residuals because the *F*-ratio was significant. Thus, we concluded that the models significantly differed regarding their effects on residuals. However, we

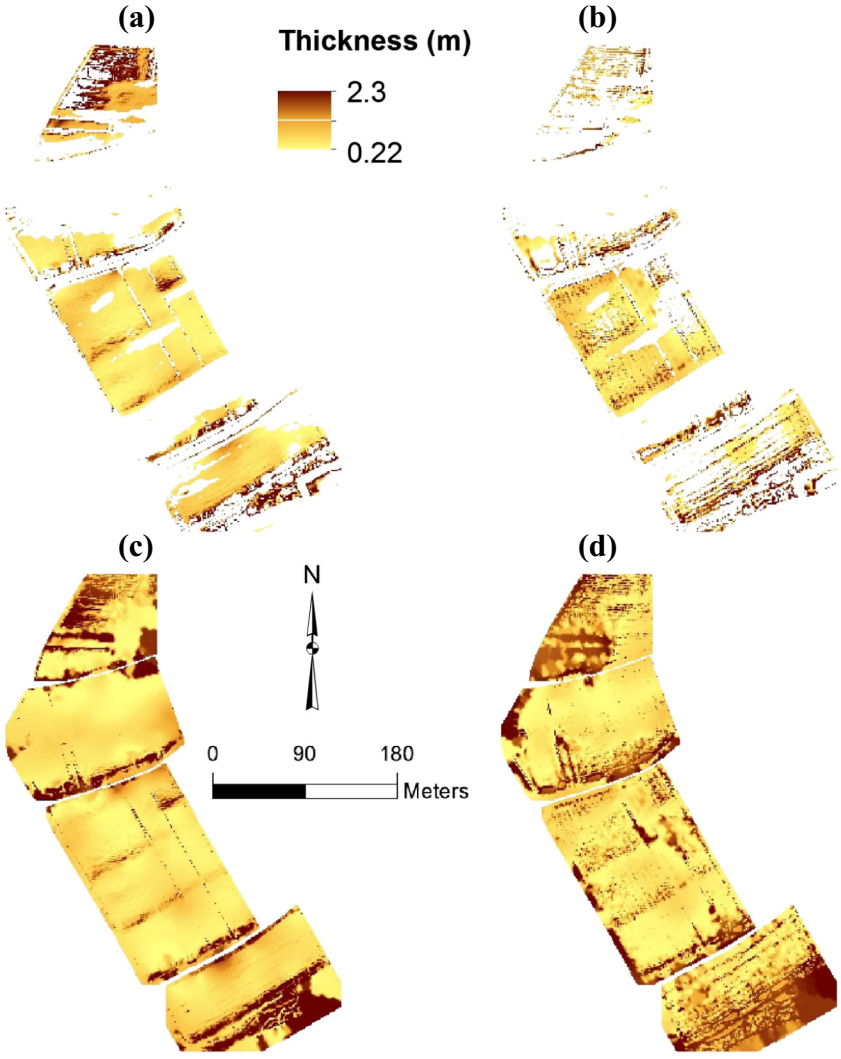

Fig. 5. The *ST* values inferred by Model-S and Model-L by using explanatory variables that were derived from LiDAR according to two scenarios as follows: (a) Model-S-LiDAR-Strict; (b) Model-L-LiDAR-Strict; (c) Model-S-LiDAR-Large; and (d) Model-L-LiDAR-Large.

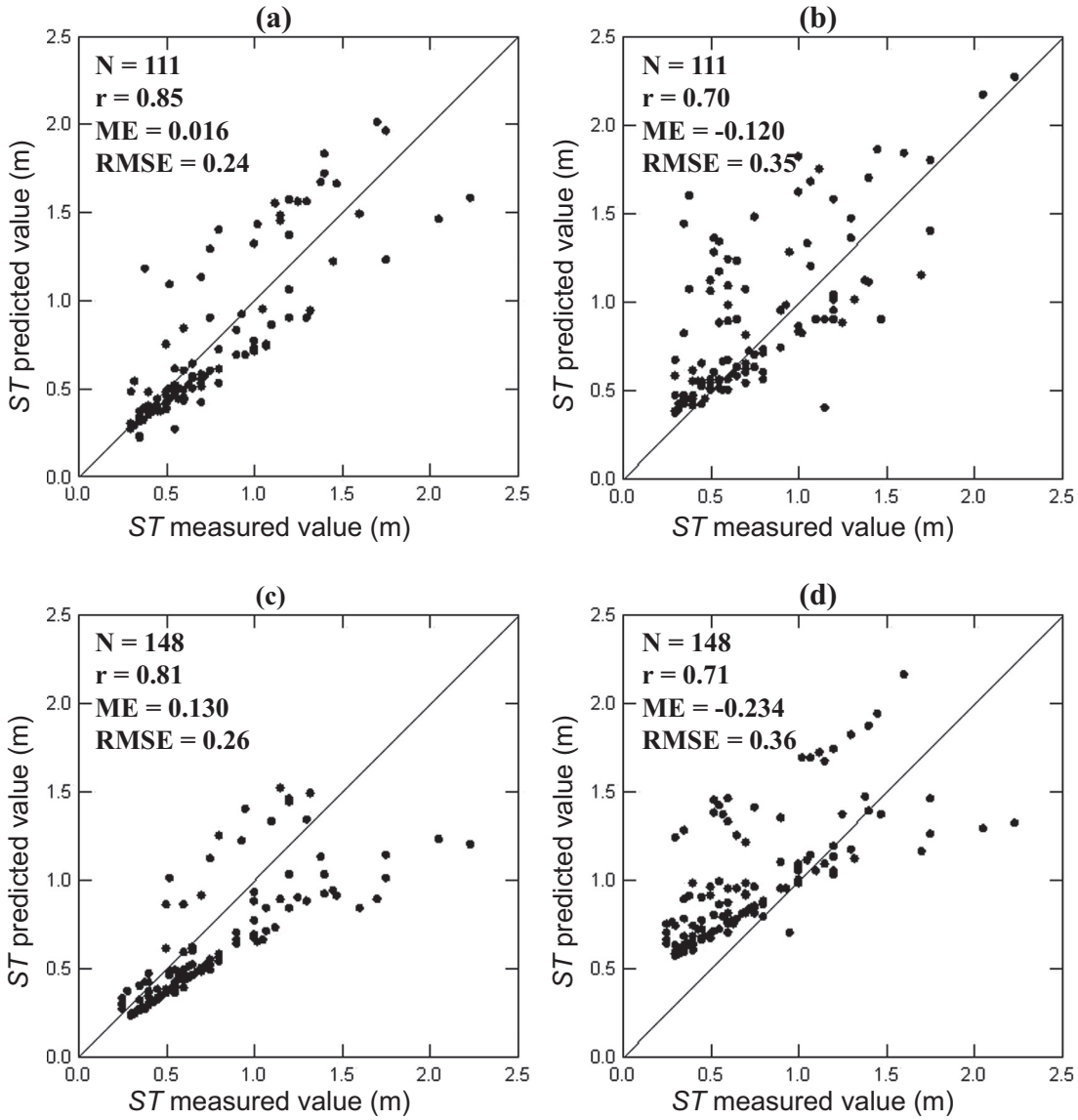

Fig. 6. Soil thickness (*ST*) values for the validation set plotted against the *ST* values that were inferred from Model-S and Model-L by using the explanatory variables that were derived from LiDAR and according to two scenarios as follows: (a) Model-S-LiDAR-Strict; (b) Model-L-LiDAR-Strict; (c) Model-S-LiDAR-Large; and (d) Model-L-LiDAR-Large.

were unable to determine which groups differed based on the ANOVA results. To examine specific group differences, we used the Dunnett pairwise mean comparison method. The results of this comparison (Table 7) allowed us to conclude that the residuals obtained with Model-S-LiDAR-Strict (Fig. 6a) were significantly lower than the residuals that were obtained from the other three situations. Furthermore, Fig. 6 and the associated statistical criteria indicated that Model-S performed better relative to Model-L in all situations.

Model-S was extrapolated by using predictor variables that were derived from the DEM acquired by DGPS over the 17 ha of the study area. In addition, Model-S was used to map *ST* (Fig. 7a) by applying it to topographic attributes that were derived from the LiDAR over areas of 3 km^2. The validation results (Fig. 7b) that used a data set of 231 individuals (Fig. 7c) indicated that the extrapolation of Model-S using topographic attributes derived from LiDAR provided comparable results relative to those presented in Fig. 6c. For Fig. 6c, the model was applied using similar conditions but over a restricted area. These results are important regarding the extrapolation of a model that was established over a small area to a large area where the response variable is sparse and the predictor variables are exhaustively sampled, but potentially tainted by measurement errors or lower accuracy.

Table 6
ANOVA of the residuals according to the model used to predict soil thickness.

Source of variation	Sum of squares	Degrees of freedom	Mean square	*F*-ratio	*Pr* (*F*)
Model	10.805	3	3.602	49.07	<0.0001
Error	37.727	514	0.073		
Total	48.532	517			

Table 7
The Dunnett pairwise mean comparison of residuals that resulted from the models that were used to predict soil thickness. (1) Model-S-LiDAR-Strict; (2) Model-L-LiDAR-Strict; (3) Model-S-LiDAR-Large; and (4) Model-L-LiDAR-Large.

Modality	Difference	Standardized difference	Critical value	Critical difference	*Pr* > *Diff*	Significant
1 vs 3	−0.114	−3.346	2.340	0.080	0.003	Yes
1 vs 4	0.250	7.342	2.340	0.080	0.000	Yes
1 vs 2	0.136	3.733	2.340	0.085	0.001	Yes

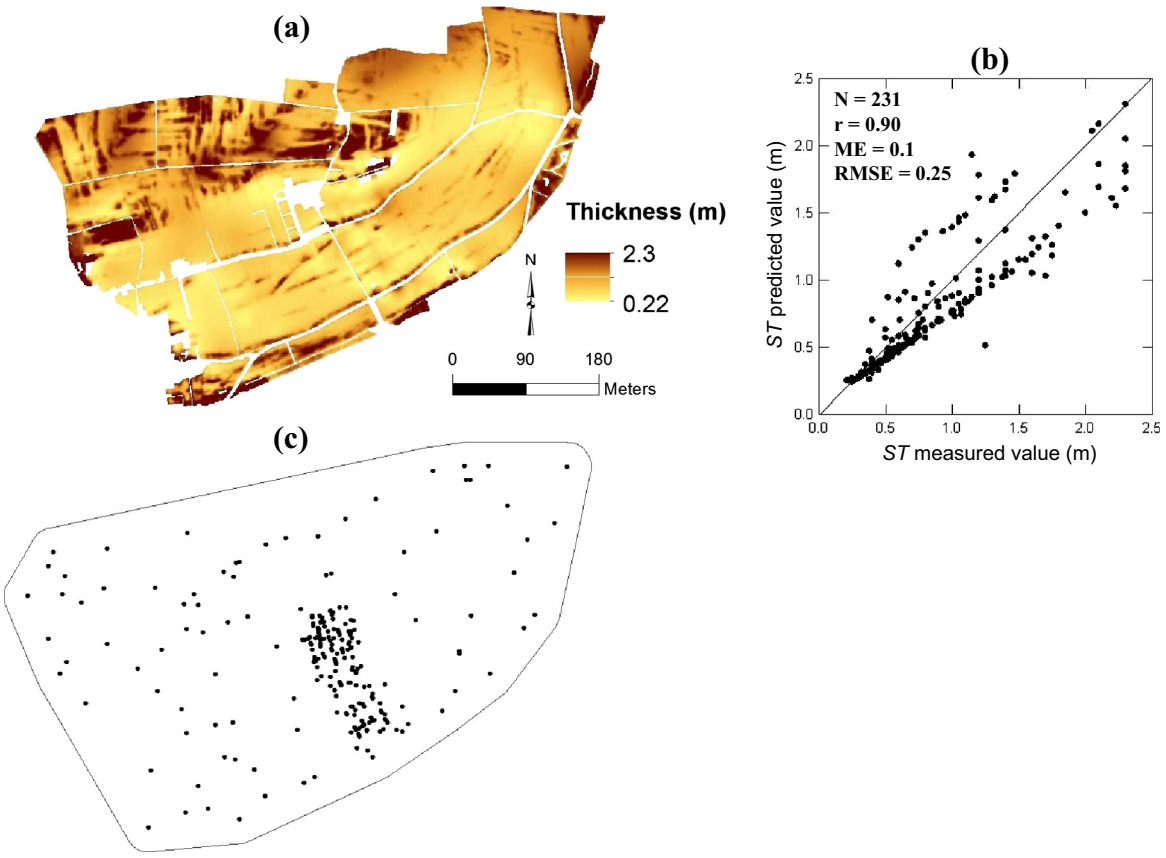

Fig. 7. Extrapolation of the model over a large area: (a) Mapping of the *ST* by extrapolation of Model-S by using the topographic attributes derived from LIDAR; (b) Mapping validation results based on an independent data set; and (c) Validation set: global sampling pattern.

5. Conclusions

The geostatistical approach developed in this study based on factorial kriging analysis allowed us to identify relevant relationships among scale, the variable of interest and the auxiliary information. Regardless of scale, *ST* was related to the small-scale structure of the primary topographic attributes such as slope, and to the large-scale structure of the secondary topographic attributes such as profile curvature. Our findings indicate that the nugget variability of the considered variables must be removed. This decomposition process is associated with a modeling approach, such as PLSR, which accounts for collinearity between the predictor variables and leads to an efficient prediction model. Both of the developed models explained a large proportion of the soil thickness variation as a function of the topographic attributes. Nevertheless, the validation results indicated that the model based on small-scale structure (Model-S) was better at predicting soil thickness. In addition, these results stressed the sensitivity of the developed models to variations in the secondary topographic attributes derived from the DEM, such as curvature. This finding is particularly true when the weight of the secondary topographic attribute is more important or equivalent to the weight of the primary topographic attribute (e.g., *S*) in the model. Thus, our results indicate that a high coefficient of determination value does not necessarily guarantee better prediction when the model is tested using an external validation data set.

The ability of Model-S to map soil thickness was confirmed by using predictor variables that were derived from LiDAR. In addition, the extrapolation of Model-S beyond the area where it was generated appeared relevant. This finding has important implications for modeling soil properties according to topographic attributes and the spatial generalization of a model established over a small area where all data are precise, to a large area where the target variable is sparse and the predictor variables are exhaustive and less precise.

Acknowledgments

The financial support provided by the ANR (Agence Nationale de la Recherche) (ANR-08-VULN-006 -04/LANDSOIL) VMCS project of LANDSOIL is gratefully acknowledged.

References

Bell, J.C., Cunningham, R.L., Havens, M.W., 1994. Soil drainage class probability mapping using a soil-landscape model. Soil Sci. Soc. Am. J. 58, 464.470.
Bellemlih, S., 1999. Stocks particulaires holocènes et bilans de matières dans un bassin fluviatile en domaine sédimentaire — Le bassin du Négron, Sud-ouest du Bassin Parisien, France. Université de Tours, France, (Ph.D. thesis).
Bolline, A., 1971. Les rideaux en Hesbaye gembloutoise — Etude morphologique et sédimentologique. Bull. Soc. Géogr. de Liège 7, 61–67.
Bourennane, H., King, D., Chéry, P., Bruand, A., 1996. Improving the kriging of a soil variable using slope gradient as external drift. Eur. J. Soil Sci. 47, 473–483.
Bourennane, H., Salvador-Blanes, S., Cornu, S., King, D., 2003. Scale of spatial dependence between chemical properties of topsoil and subsoil over a geologically contrasted area (Massif Central, France). Geoderma 112, 235–251.
Bourennane, H., Nicoullaud, B., Couturier, A., Pasquier, C., Mary, B., King, D., 2012. Geostatistical filtering for improved soil water content estimation from electrical resistivity data. Geoderma 183–184, 32–40.
Castrignano, A., Wong, M.T.F., Stelluti, M., De Benedetto, D., Sollitto, D., 2012. Use of EMI, gamma-ray emission and GPS height as multi-sensor data for soil characterization. Geoderma 175–176, 78–89.
Chartin, C., Bourennane, H., Salvador-Blanes, S., Hinschberger, F., Macaire, J.-J., 2011. Classification and mapping of anthropogenic landforms on cultivated hillslopes using DEMs and soil thickness data — example from the SW Parisian Basin, France. Geomorphology 135, 8–20.
Chilès, J.P., Delfiner, P., 1999. Geostatistics: Modeling Spatial Uncertainty. Wiley, New York, (695 pp.).
Claessens, L., Heuvelink, G.B.M., Schoorl, J.M., Veldkamp, A., 2005. DEM resolution effects on shallow landslide hazard and soil redistribution modelling. Earth Surf. Process. Landf. 30, 461–477.
Debella-Gilo, M., Etzelmüller, B., 2009. Spatial prediction of soil classes using digital terrain analysis and multinomial logistic regression modeling integrated in GIS: examples from Vestfold County, Norway. Catena 77, 8–18.

Desmet, P.J.J., Govers, G., 1996. Comparison of routing algorithms for digital elevation models and their implications for predicting ephemeral gullies. Int. J. Geogr. Inf. Syst. 10, 311–331.
Erskine, R.H., Green, T.R., Ramirez, J.A., MacDonald, L.H., 2006. Comparison of grid-based algorithms for computing upslope contributing area. Water Resour. Res. 42, W09416. http://dx.doi.org/10.1029/2005WR004648.
Gerlach, R.W., Kowalski, B.R., Wold, H., 1979. Partial least squares modelling with latent variables. Anal. Chim. Acta. 112, 417–421.
Goovaerts, P., 1997. Geostatistics for Natural Resources Evaluation. Oxford University Press, (483 pp.).
Goovaerts, P., Webster, R., 1994. Scale-dependent correlation between topsoil copper and cobalt concentrations in Scotland. Eur. J. Soil Sci. 45, 79–95.
Goulard, M., 1989. Inference in a co-regionalization model. In: Armstrong, M. (Ed.), Geostatistics. Kluwer Academic Publisher, Amsterdam, Holland, pp. 397–408.
Goulard, M., Voltz, M., 1992. Linear co-regionalization model: tool for estimation and choice of cross-variogram matrix. Math. Geol. 24, 269–286.
Höskuldsson, A., 1988. PLS regression methods. J. Chemom. 2, 211–228.
Kim, D., Zheng, Y., 2011. Scale-dependent predictability of DEM-based landform attributes for soil spatial variability in a coastal dune system. Geoderma 164, 181–194.
King, D., Bourennane, H., Isambert, M., Macaire, J.-J., 1999. Relationship of the presence of a non-calcareous clay-loam horizon to DEM attributes in a gently sloping area. Geoderma 89 (1–2), 95–111.
Knotters, M., Brus, D.J., Oude Voshaar, J.H., 1995. A comparison of kriging, co-kriging and kriging combined with regression for spatial interpolation of horizon depth with censored observations. Geoderma 67, 227–246.
Lagacherie, P., McBratney, A.B., Voltz, M. (Eds.), 2007. Digital Soil Mapping: an Introductory Perspective. Development in Soil Science, vol. 31. Elsevier Science & Technology, p. 600.
Li, S., MacMillan, R.A., Lobb, D.A., McConkey, B.G., Moulin, A., Fraser, W.R., 2011. Lidar DEM error analyses and topographic depression identification in a hummocky landscape in the prairie region of Canada. Geomorphology 129, 263–275.
Macaire, J.-J., Bellemlih, S., Di Giovanni, C., De Luca, P., Visset, L., Bernard, J., 2002. Sediment yield and storage variations in the Negron river catchment (South western Parisian Basin, France) during the Holocene period. Earth Surf. Process. Landf. 27, 991–1009.
McBratney, A.B., Mendoça, Santos, M.L., Minasny, B., 2003. On digital soil mapping. Geoderma 117, 3–52.
Moore, I.D., Grayson, R.B., Landson, A.R., et al., 1991. Digital terrain modelling: a review of hydrological, geomorphological, and biological applications. Hydrol. Process. 5, 3–30.
Moore, I.D., Gessler, P.E., Nielsen, G.A., Peterson, G.A., 1993. Soil attribute prediction using terrain analysis. Soil Sci. Soc. Am. J. 57, 443–452.
Muñoz, J.D., Kravchenko, A., 2012. Deriving the optimal scale for relating topographic attributes and cover crop plant biomass. Geomorphology 179, 197–207.
Salvador-Blanes, S., Cornu, S., Couturier, A., King, D., Macaire, J.-J., 2006. Morphological and geochemical properties of soil accumulated in hedge-induced terraces in the Massif Central, France. Soil Tillage Res. 85, 62–77.
Shi, X., Girod, L., Long, R., DeKett, R., Philippe, J., Burke, T., 2012. A comparison of LiDAR-based DEMs and USGS-sourced DEMs in terrain analysis for knowledge-based digital soil mapping. Geoderma 170, 217–226.
Smith, M.P., Zhu, A.X., Burt, J.E., Stiles, C., 2006. The effects of DEM resolution and neighbourhood size on digital soil survey. Geoderma 137, 58–69.
Stevenson, J.A., Sun, X., Mitchell, N.C., 2010. Despeckling SRTM and other topographic data with a denoising algorithm. Geomorphology 114, 238–252.
Tenenhaus, M., 1998. La régression PLS: Théorie et Pratique. Editions Technip, Paris, (252 pp.).
Thompson, J.A., Bell, J.C., Butler, C.A., 2001. Digital elevation model resolution: effects on terrain attribute calculation and quantitative soil-landscape modelling. Geoderma 100, 67–89.
Wackernagel, H., 1998. Multivariate Geostatistics. Springer-Verlag, Berlin, (256 pp.).
Wechsler, S.P., 2007. Uncertainties associated with digital elevation models for hydrologic applications: a review. Hydrol. Earth Syst. Sci. 11, 1481–1500.
Wilson, J.P., Repetto, P.L., Snyder, R.D., 2000. Effect of data source, grid resolution, and flow routing method on computed topographic attribute. In: Wilson, J.P., Gallant, J.C. (Eds.), Terrain Analysis: Principles and Application. John Wiley & Sons, New York, pp. 133–161.
Wold, S., Ruhe, A., Wold, H., Dunn III, W.J., 1984. The collinearity problem in linear regression. The partial least squares (PLS) approach to generalized inverses. SIAM J. Sci. Stat. Comput. 3, 735–743.
Zeverbergen, L.W., Thorne, C.R., 1987. Quantitative analysis of land surface topography. Earth Surf. Process. Landf. 12, 47–56.