European Journal of Soil Science

European Journal of Soil Science, September 2017, **68**, 575–591 doi: 10.1111/ejss.12446

Multivariate mapping of soil with structural equation modelling

M. E. Angelini[a,b,c], G. B. M. Heuvelink[a,b] & B. Kempen[b]

[a]*Soil Geography and Landscape Group, Wageningen University, Droevendaalsesteeg 3 (Building 101), PO Box 47, 6708 PB, Wageningen, the Netherlands,* [b]*ISRIC – World Soil Information, Droevendaalsesteeg 3 (Building 101), PO Box 353, 6708 PB, Wageningen, the Netherlands, and* [c]*INTA-CIRN, Instituto de Suelos, N. Repetto y Los Reseros s/n, 1686, Hurlingham, Argentina*

Summary

In a previous study we introduced structural equation modelling (SEM) for digital soil mapping in the Argentine Pampas. An attractive property of SEM is that it incorporates pedological knowledge explicitly through a mathematical implementation of a conceptual model. Many soil processes operate within the soil profile; therefore, SEM might be suitable for simultaneous prediction of soil properties for multiple soil layers. In this way, relations between soil properties in different horizons can be included that might result in more consistent predictions. The objectives of this study were therefore to apply SEM to multi-layer and multivariate soil mapping, and to test SEM functionality for suggestions to improve the modelling. We applied SEM to model and predict the lateral and vertical distribution of the cation exchange capacity (CEC), organic carbon (OC) and clay content of three major soil horizons, A, B and C, for a 23 000-km^2 region in the Argentine Pampas. We developed a conceptual model based on pedological hypotheses. Next, we derived a mathematical model and calibrated it with environmental covariates and soil data from 320 soil profiles. Cross-validation of predicted soil properties showed that SEM explained only marginally more of the variance than a linear regression model. However, assessment of the covariation showed that SEM reproduces the covariance between variables much more accurately than linear regression. We concluded that SEM can be used to predict several soil properties in multiple layers by considering the interrelations between soil properties and layers.

Highlights

- We tested structural equation modelling (SEM) for multi-layer and multivariate soil mapping.
- SEM models soil property covariation better than multiple linear regression.
- The SEM re-specification step improves prediction accuracy.
- SEM supports learning about soil processes from data.

Introduction

Many environmental and agro-economic activities require accurate information about the spatial distribution of soil types and properties. This information is being generated increasingly through digital soil mapping (DSM) techniques (Minasny & McBratney, 2016). They are largely data driven and make use of empirically established relations between soil and landscape properties and exploit spatial correlation in soil properties. Soil properties are typically modelled and predicted individually, and for different horizons or depth layers separately. This might result in unrealistic or inconsistent predictions because interrelations between soil properties are not taken into account. For example, if soil organic carbon (SOC) is predicted layer by layer, the resulting predicted SOC profiles might be physically unrealistic. If SOC and soil organic nitrogen are predicted separately, the resulting maps might produce implausible C:N ratios (Heuvelink *et al.*, 2016). Although the accuracy of the individual maps might be acceptable, the consistency of the

Correspondence: M. E. Angelini. E-mail: angelini75@gmail.com
Received 22 July 2016; revised version accepted 12 May 2017

predictions between several soil properties and between layers might fail to meet required standards and possibly impair subsequent analyses.

The problem of inconsistency between multiple spatial predictions is not new to soil science or to other fields. There are many techniques that can deal with the simultaneous prediction of several dependent variables, such as cokriging (Webster & Oliver, 2007), factorial kriging (Goovaerts, 1992) and regression-cokriging (Orton *et al*., 2014; Heuvelink *et al*., 2016). These geostatistical methods model the spatial interrelations explicitly among several soil properties, but the modelling becomes cumbersome as the number of variables increases. Multivariate linear regression, partial least squares regression and multivariate machine-learning algorithms have also been used to predict multiple dependent variables simultaneously (e.g. Viscarra Rossel *et al*., 2006; Xu *et al*., 2013). These methods are useful for predicting many dependent variables simultaneously, but they are empirical and lead to complex models that are difficult to interpret. As a result they cannot be used easily for extrapolation and provide little insight into cause and effect relations.

Mechanistic models also predict multiple soil and landscape properties simultaneously (Opolot *et al*., 2015; Temme & Vanwalleghem, 2015). Their advantage is that they are based on mechanistic principles, which fosters extrapolation and aids understanding of physical, chemical and biological processes. These dynamic models are unfortunately often very complex. Apart from large uncertainties in the model inputs and parameters, model structural uncertainty can also be large.

Recently, we proposed structural equation modelling (SEM) as a compromise between empirical and mechanistic approaches for soil spatial prediction (Angelini *et al*., 2016). It is designed specifically for modelling cause and effect interrelations and can include dependencies between dependent variables (Bollen, 1989). It has been applied extensively in ecology (Grace *et al*., 2012). It can be considered a semi-mechanistic approach because the starting point of model formulation is a mechanistic conceptual model, although calibration relies predominately on empirical approaches and the model cannot describe dynamic processes explicitly (Grace *et al*., 2012). In our previous study (Angelini *et al*., 2016), we demonstrated that it is possible to include interrelations between soil properties in the modelling process. In a case study we made 2-D predictions for an area in the Argentine Pampas with SEM. In addition, SEM also seems suitable for multiple layer soil prediction because it can represent vertical processes through implementation of a conceptual model, and relations between soil properties at different depths or horizons can be included. Angelini *et al.* (2016) did not explore more advanced SEM techniques that can improve model performance, one of which is that SEM can be used in an exploratory way to detect additional relations that could be included in the conceptual model (Grace *et al*., 2012). This might improve the predictive power and help to increase understanding of the system and develop new theories.

The objectives of this study were to apply SEM for multi-layer and multivariate soil mapping and test the functionality of SEM for suggested model improvement. We apply SEM to model and predict the cation exchange capacity, organic carbon and clay content of three major soil horizons, A, B and C, in an area of the Argentine Pampas. We validate the resulting maps with cross-validation of the prediction accuracy and the accuracy with which the covariation among different soil properties and among the same soil property for different layers is represented.

Materials and methods

Study area

The study area covers about 23 000 km^2 in the Argentine Pampas between 35°00′–33°17′W and 58°55′–61°21′S (Figure 1). Before cultivation this was a grassland plains region formed by aeolian sediments consisting of loess and loess-like materials. The main soil types are Typic and Vertic Argiudolls (Phaeozems in WRB classification, IUSS Working Group WRB, 2015) in association with soil that has natric horizons (Solonetz in Soil Taxonomy and WRB) (Morrás & Moretti, 2016). In spite of its apparent homogeneity, the loess is derived from several sources that affect the soil chemical and physical properties (Morrás & Moretti, 2016).

Annual precipitation ranges between 900 and 1000 mm. Rain is deficient in the summer and in excess in winter. The average summer temperature is 23°C and the average temperature in winter is 10°C. Under this climate, land use has changed from native grassland to mainly arable land in the past century.

Soil data

The region was surveyed during the 1960s and 1970s. Data were extracted from 344 profiles of the soil information system of the Argentine National Institute of Agricultural Technology (INTA). Figure 1 shows the sampling locations.

We selected three soil properties, percentage of soil organic carbon (OC mass percentage), clay content (mass percentage) and cation exchange capacity (CEC in $cmol_c\ kg^{-1}$ soil), which we model for three major soil horizons: A, B and C. The original soil horizons were grouped as follows.

- A horizon: A1 and Ap or any subdivision of these (e.g. Ap1, Ap2).
- B horizon: B2, Bt, Bn or any subdivision of these.
- C horizon: usually represented as C, C2, R or X.

We did not include transitional horizons, such as AB, BA or BC. Figure 2 shows the frequency of occurrence of the horizons and the distribution of the soil properties down the profile. Note that most A horizons occur above 50-cm depth, whereas the C horizon generally starts at 100-cm depth or deeper. Figure 3 shows the correlations among soil properties and horizons. More detailed information about the soil data is provided in Angelini *et al.* (2016).

External factors

Table 1 summarizes the external factors used in the modelling process. The main sources of information included the following. The

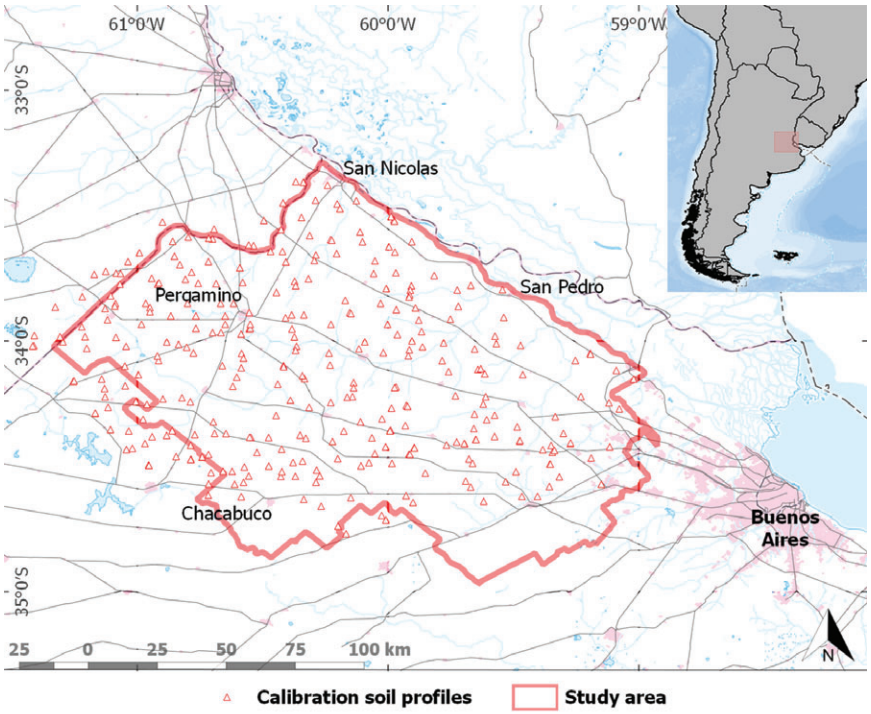

Figure 1 Extent of the study area and locations of soil profiles used for calibration and cross-validation.

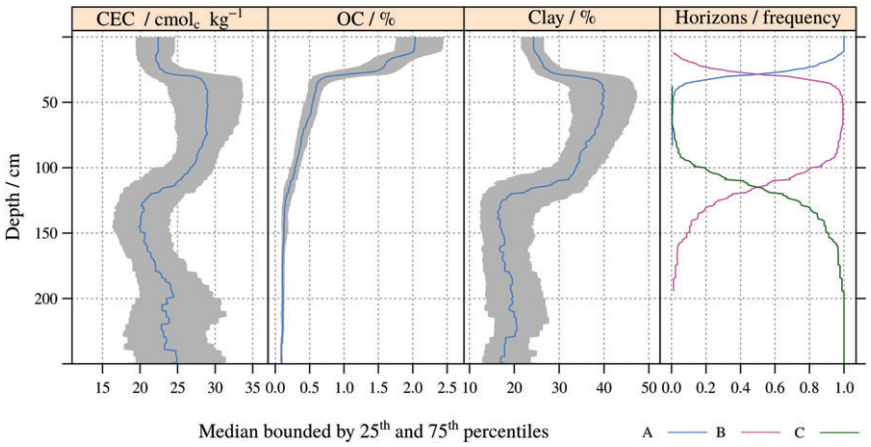

Figure 2 Graphs of the median of cation exchange capacity (CEC), organic carbon (OC) and clay (Clay), as a function of depth; the grey area represents the 50% envelope between the 25th and 75th quantiles. Frequency of occurrence of each horizon type as a function of depth (Horizons).

Shuttle Radar Topography Mission (SRTM) digital elevation model (DEM) was pre-processed to reduce artefacts and striping noise, and then used to derive the external terrain factors listed in Table 1.

Enhanced vegetation index (EVI [MOD13Q1]) and land-surface temperature and emissivity (LST [MOD11A2]) were taken from MODIS (moderate-resolution imaging spectroradiometer; the MOD13Q1, MCD43A4 and MOD11A2 were retrieved from the online Reverb/ECHO tool [http://reverb.echo.nasa.gov/reverb/], courtesy of the NASA EOSDIS Land Processes Distributed Active Archive Center [LP DAAC], USGS/Earth Resources Observation and Science [EROS] Center, Sioux Falls, South Dakota, USA. https://lpdaac.usgs.gov/citing_our_data#sthash.yGKPuOqi.dpuf). The standard deviation of a 15-year monthly time series from March 2000 to December 2014 was calculated per pixel for EVI, which represents land cover dynamics. The mean value of LST was computed for the same period as an indicator of mean soil temperature, which depends on soil texture, among other factors. We also computed the normalized difference of water index (NDWI) from MODIS MCD43A4 (Poggio *et al.*, 2013) by averaging time-series imagery for the periods 17 January to 26 February (late summer) and 8 October to 11 November (mid-spring) 2000–2015. These two periods were selected because of the large contrast in vegetation intensity between them. The NDWI represents seasonal vegetation dynamics of arable land and lowland. Finally, we generated an image of distance to the Paraná River, which can be considered to represent parent material (Morrás & Moretti, 2016).

All variables were standardized by subtracting their mean and dividing by their standard deviation.

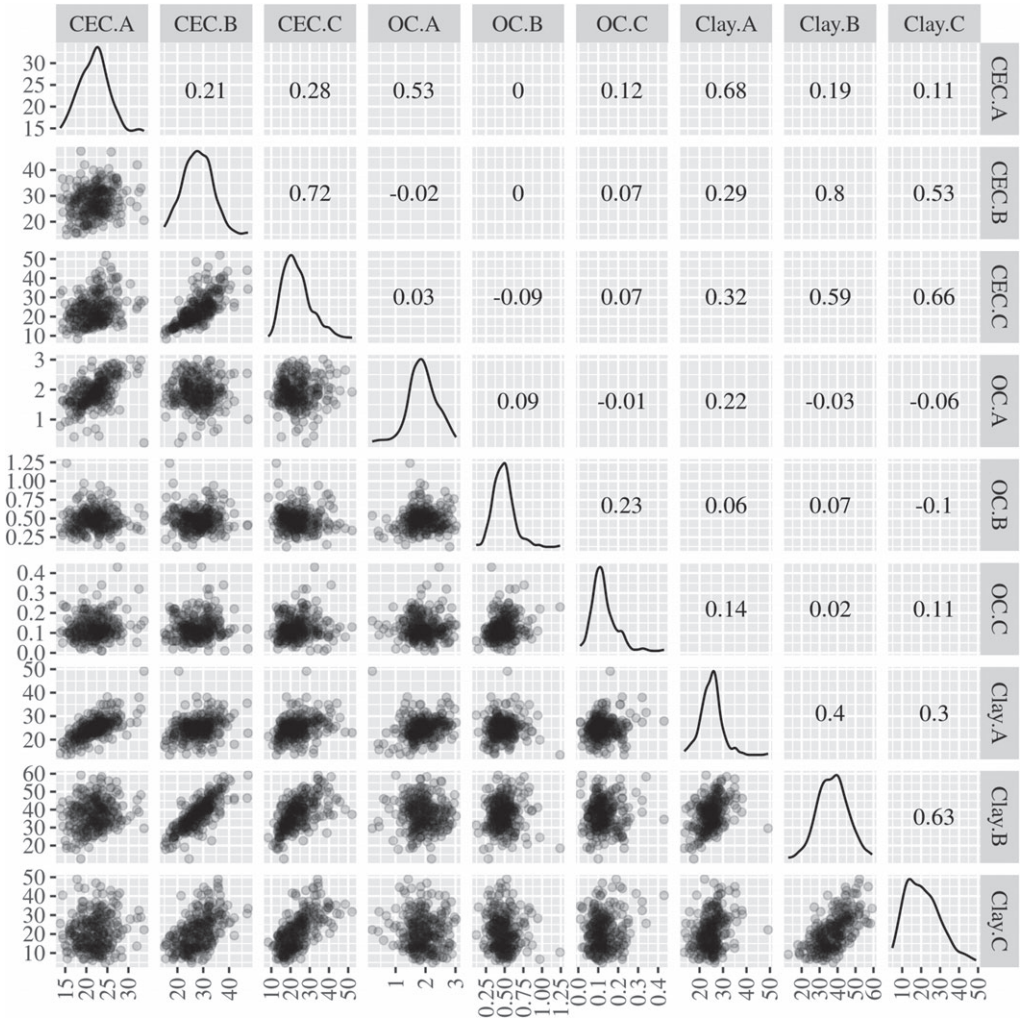

Figure 3 Correlation graph of soil properties by horizons. The upper right triangle shows the correlation between properties, the diagonal presents the histogram of the properties and the lower left triangle the scatter plots. Soil properties are abbreviated such that the name of the soil property is followed by the horizon name and separated by a dot, so that Clay.A represents the clay percentage in horizon A, Clay.B is the clay percentage in horizon B, and so on. OC is organic carbon and CEC is cation exchange capacity.

Modelling framework for structural equation

To formulate, apply and evaluate an SE model we divided the modelling process into seven steps (Figure 4).

1 *Conceptual model*: a conceptual model identifies the mechanistic processes that explain the functioning of a system. Its development means it is necessary to consider the (hypothesized) physical, chemical and biological laws that define the system. One has to link concepts to system variables and explain the main relations among these.
2 *Graphical model*: the conceptual model becomes more specific in a graphical model that defines the type of variables included, such as observed, latent or composite variables (Grace *et al.,* 2012). Arrows have to be identified that represent cause and effect relations between the variables.
3 *Mathematical model*: the mathematical model automatically follows from the graphical model. It includes three basic equations (Bollen, 1989):

$$\mathbf{x} = \mathbf{\Lambda}\boldsymbol{\xi} + \boldsymbol{\delta}, \tag{1}$$

$$\mathbf{y} = \mathbf{K}\boldsymbol{\eta} + \boldsymbol{\epsilon}, \tag{2}$$

$$\boldsymbol{\eta} = \mathbf{B}\boldsymbol{\eta} + \mathbf{\Gamma}\boldsymbol{\xi} + \boldsymbol{\zeta}, \tag{3}$$

where $\mathbf{x}$ is a vector of q observed exogenous variables (i.e. external factors), $\mathbf{y}$ is a vector of p observed endogenous variables (i.e. soil properties), $\boldsymbol{\xi}$ and $\boldsymbol{\eta}$ are vectors of n latent exogenous and m endogenous variables, $\mathbf{\Lambda}$ and $\mathbf{K}$ are $q \times n$ and $p \times m$ coefficient matrices that link observed to latent variables, $\boldsymbol{\delta}$ and $\boldsymbol{\epsilon}$ are vectors of measurement errors of length q and p, respectively (mutually independent and zero-mean normal deviates), $\mathbf{B}$ and $\mathbf{\Gamma}$ are $m \times m$ and $m \times n$ coefficient matrices of

Table 1 External factors

Factor	Description	Source	Resolution
LSTM	Mean of 14 years of daytime 8-day land-surface temperature	Terra/MODIS, product MOD11A2	1 km
EVISD	Standard deviation of 14 years of enhanced vegetation index (EVI) for 16 days	Terra/MODIS, product MOD13Q1	250 m
NDWI.A	Normalized difference water index (NDWI) bands NIR (∼850 nm) and SWIR (∼1240 nm). Summer season	MODIS product MCD43A4	500 m
NDWI.B	Normalized difference water index (NDWI) bands NIR (∼850 nm) and SWIR (∼1240 nm). Spring season	MODIS product MCD43A4	500 m
DEM	Altitude (metres)	SRTM	30 m
VDCHN	Vertical distance to channel network (metres)	SRTM	30 m
TWI	Terrain wetness index	SRTM	30 m
RIVER	Distance to Paraná River (metres)	–	30 m
LAT	Latitude of plain coordinates (metres)	–	30 m
LON	Longitude of plain coordinates (metres)	–	30 m

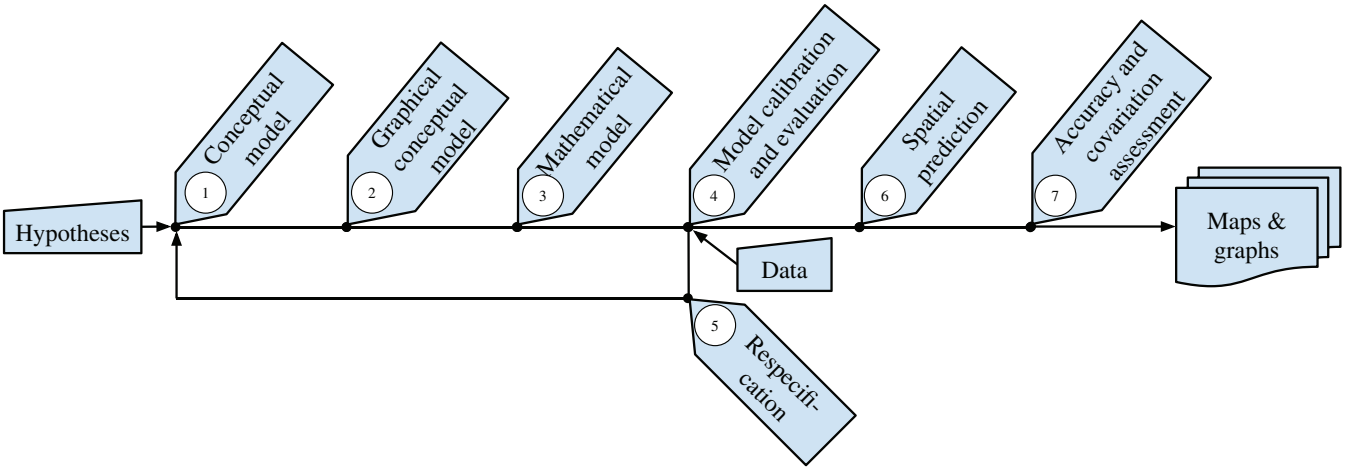

Figure 4 Steps in structural equation modelling (SEM) for spatial prediction of soil properties.

endogenous and exogenous relations and $\boldsymbol{\zeta}$ is vector of length m of model error for variable $\boldsymbol{\eta}$ (Grace *et al.,* 2012). Note that the diagonal elements of $\mathbf{B}$ are forced to zero so that soil properties cannot depend on themselves. Equations (1) and (2) define the measurement model, whereas Equation (3) corresponds to the structural model. Three more terms complete the mathematical model, $\boldsymbol{\Psi}$ is the $m \times m$ variance–covariance matrix of $\boldsymbol{\zeta}$, the off-diagonal elements of which represent relations between latent endogenous variables that cannot be explained by other means. The terms $\boldsymbol{\Theta}_{\boldsymbol{\delta}}$ and $\boldsymbol{\Theta}_{\boldsymbol{\epsilon}}$ are $q \times q$ and $p \times p$ variance–covariance matrices of $\boldsymbol{\delta}$ and $\boldsymbol{\epsilon}$.

4 *Model calibration and evaluation*: these comprise a comparison of the variance–covariance matrix of the data, denoted by $\mathbf{S}$, with the model-implied variance–covariance matrix $\boldsymbol{\Sigma}(\boldsymbol{\theta})$, which is written as a function of $\boldsymbol{\theta}$, where $\boldsymbol{\theta}$ represents all model parameters ($\mathbf{B}$, $\boldsymbol{\Gamma}$, $\mathbf{K}$, $\boldsymbol{\Lambda}$, $\boldsymbol{\Psi}$, $\boldsymbol{\Theta}_{\boldsymbol{\delta}}$ and $\boldsymbol{\Theta}_{\boldsymbol{\epsilon}}$). The model parameters are generally estimated by maximum likelihood (ML). Model evaluation also includes a close examination of estimated coefficients to determine whether their signs are coherent with the conceptual model and their magnitude agrees with what might rationally be expected (Bollen, 1989).

5 *Model respecification*: conceptual models typically do not take into account all relations of complex systems such as the soil system. Models are kept deliberately simple, and knowledge about system functioning is often limited. There could also be alternative conceptual models. For these reasons, conceptual models might be misspecified. Misspecification might be detected partly by SEM, requiring a modification of the model.

6 *Spatial prediction*: prediction in classical SEM applications refers to predicting the scores of the latent variables (Rosseel, 2012). Here we are interested in using the calibrated equations to predict the dependent variables from the measured independent variables. The solution is derived from Equations (1) and (3) (Angelini *et al.,* 2016):

$$\hat{\boldsymbol{\eta}} = (\mathbf{I} - \mathbf{B})^{-1} \boldsymbol{\Gamma} \boldsymbol{\Lambda}^{-1} \mathbf{x}. \quad (4)$$

Note that the dependent variables are predicted from independent variables only, even though they depend on other dependent

variables. The prediction error variance can also be computed (Angelini *et al*., 2016).

7 *Model accuracy and covariation assessment*: in this final step the prediction maps are evaluated in terms of their accuracy and covariation among predicted soil properties.

In this study, we applied the seven steps above to model and predict the cation exchange capacity (CEC) and its two main controlling factors, soil organic carbon (OC) and clay content. Most of the steps above have been explained in detail in Angelini *et al.* (2016), except for steps 4, 5 and 7. These are given in more detail below.

Model calibration and evaluation

Measures of overall fit aim to assess the validity of the calibrated model. There is not a single measure, however, that can assess the model-fitting completely and for this reason several statistics have been developed (Kline, 2015). Most overall fitting measures are based on a comparison of the sample variance–covariance matrix $\mathbf{S}$ and the model-implied variance–covariance matrix $\boldsymbol{\Sigma}(\boldsymbol{\theta})$. Matrix $\mathbf{S}$ is computed directly from the observations of the endogenous variables, whereas $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ follows from Equations (1) to (4):

$$\boldsymbol{\Sigma}(\boldsymbol{\theta}) = (\mathbf{I}-\mathbf{B})^{-1}\left((\boldsymbol{\Gamma}\boldsymbol{\Phi}\boldsymbol{\Gamma}^{\mathrm{T}}+\boldsymbol{\Psi})\left((\mathbf{I}-\mathbf{B})^{-1}\right)^{\mathrm{T}}+\boldsymbol{\Theta}_{\boldsymbol{\epsilon}}\right., \qquad (5)$$

where $\boldsymbol{\Phi}$ is the $n \times n$ variance–covariance matrix of $\boldsymbol{\xi}$, computed from the observations of exogenous variables. Note that use of $\boldsymbol{\Phi}$ effectively means that the exogenous variables are treated as random effects in Equation (3). This is required because variation in the exogenous variables is also incorporated in the calculation of $\mathbf{S}$. It must then also be included in $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ to make the comparison valid. Note also that we made the simplifying assumptions $\boldsymbol{\Lambda}=\mathbf{K}=\mathbf{I}$ and $\boldsymbol{\Theta}_{\boldsymbol{\delta}}=0$. Note that the latter assumption implies that the vector of covariates $\mathbf{x}$ becomes deterministic. These assumptions apply to our soil mapping example, but the methodology also applies more generally (e.g. Bollen, 1989, Chapter 4).

The simplest way to assess overall model performance would be by computing the difference between $\mathbf{S}$ and $\boldsymbol{\Sigma}(\boldsymbol{\theta})$. The standardized root mean-square residual (SRMR) is the standardized average of the absolute differences between $\mathbf{S}$ and $\boldsymbol{\Sigma}(\boldsymbol{\theta})$, which operates on the correlation matrices instead of the covariance matrices (Kline, 2015). Another measure that is frequently used is goodness of fit (GFI), which is analogous to the coefficient of determination used in linear regression. It measures the amount of variance and covariance in the data that is explained by the model (Jöreskog & Sörbom, 1981; Bollen, 1989).

Model validity measures are also often used in SEM, such as the comparative fit index (CFI), among others. The CFI was developed by Bentler (1990) to estimate the overall model fit when the sample size is small. This index compares the chi-square (χ^2) value of the model with the χ^2 value of a so-called baseline model. The baseline is the simplest model, where $\mathbf{B}$ and $\boldsymbol{\Gamma}$ are zero (no cause and effect relations), there are no latent variables and correlation between observed variables is zero. The diagonal matrix $\boldsymbol{\Phi}$ (variance of $\mathbf{x}$) contains free parameters only. The CFI measures how much better the selected model is than the baseline model, where zero means no improvement and one means a perfect fit. The SEM literature suggests a CFI cut-off value of 0.95, although it is case dependent (Marsh *et al.,* 2004). In addition to these measures, we computed the model R^2.

Model respecification

Often, our knowledge about system functioning is limited, or the variables that we wish to observe are difficult to measure, such as soil-forming process variables for which we often have only proxies. Lack of knowledge on soil-forming processes means that we might not know which cause and effect relations to include in the graphical model. Misspecification of a model might result from inclusion or exclusion of relations in a model. Respecification, or modification of the model, might solve this problem by a knowledge-based and or empirical approach (Bollen, 1989). The first develops alternative approaches that conform to our knowledge, whereas the second uses algorithms to obtain 'suggestions' that may help to improve the model. Here we focus on the empirical approach, also referred to as exploratory analysis in SEM literature.

Exploratory analysis involves adding or removing a new parameter (new relation between two properties), and subsequently checking whether this improves test statistics for model fitting. This stage has been automated in SEM modelling using different tests such as the Lagrange multiplier (Bentler, 1990), a χ^2-test with one degree of freedom. This test estimates how much χ^2 decreases if one of the model restrictions is released (i.e. if a relation not yet part of the model is included) (Kline, 2015). The test reports a modification index (MI) for every possible parameter (arrow in the graphical model) that can be added to the model, analogous to the approach used in stepwise regression. In this study we checked for modifications in $\mathbf{B}$, $\boldsymbol{\Gamma}$ and $\boldsymbol{\Psi}$ only (i.e. which endogenous variables depend on other endogenous and exogenous variables) and on the covariance of system noise between endogenous variables.

Model accuracy and assessment of covariation

In Angelini *et al.* (2016), we determined the accuracy of the individual soil maps through common measures. Covariation among predicted variables, which measures how correlations between dependent variables are reproduced by the model, is not taken into consideration by these conventional accuracy metrics. Although some studies have addressed the issue (e.g. Orton *et al*., 2014), models with multivariate outcomes in DSM have not used covariation in this way.

We assess accuracy by leave-one-out cross-validation, in which the model parameters were re-estimated each time. We quantified prediction bias with the mean error (ME) and overall accuracy with the root mean squared error (RMSE). The prediction power was estimated by the amount of variance explained (AVE), also

known as the Nash–Sutcliffe efficiency (Krause *et al.,* 2005). It is defined as:

$$\mathrm{AVE} = 1 - \frac{\sum_i^n \left(y_i - \hat{y}_i\right)^2}{\sum_i^n \left(y_i - \bar{y}\right)^2}, \tag{6}$$

where y_i is the i-th measurement of the target variable, $\hat{y}_i$ is the corresponding predicted value, $\bar{y}$ is the mean and n is the number of observations.

We compute the mean and median standardized squared prediction error proposed by Lark (2000) as an indicator of correct assessment of map uncertainty. Apart from these measures, we computed a measure for the preservation of the relations among soil properties. Following the rationale of SEM, we compare the correlation matrix of measured soil properties with the predicted correlation matrix. These matrices are standardized versions of the observation covariance matrix $\mathbf{S}$ and the model-induced covariance matrix $\boldsymbol{\Sigma}(\boldsymbol{\theta})$. From their difference, a correlation difference matrix can be obtained. The SRMR measure may then be used as a summary measure of how well covariation is reproduced in the model predictions.

For comparison, we also fitted multiple linear regression (MLR) models to predict OC, clay content and CEC for the three horizons individually with the same covariates as used in SEM. For these models we computed the cross-validation statistics and assessed the preservation of covariation through the standardized $\boldsymbol{\Sigma}(\boldsymbol{\theta})_{\mathrm{MLR}}$. We compared this with the correlation matrix of the observations and computed the $\mathrm{SRMR}_{\mathrm{MLR}}$.

Results

Conceptual model

Cation exchange capacity is determined by the sum of the CEC of each individual colloid in the soil. Sources of colloids in the soil are clay and humus particles. The smaller is the particle, the larger is its surface to adsorb cations (Brady & Weil, 2013).

The soil of the study area has small amounts of OC: 1–3% in A horizons and typically less than 1% in B and C horizons (Figure 2). The amount of OC in the C horizon can be considered negligible and therefore we assume that it does not affect the CEC in this horizon.

One of the main causes of soil spatial variation in the study area is parent material. Particle-size distribution shows a coarse to fine gradient from southwest to northeast. The loess deposits have been reworked by aeolian and fluvial processes (Morrás & Moretti, 2016). Rain and subsequent water infiltration caused argilluviation, which is considered one of the dominant and most extensive soil-forming processes in the area. Consequently, the B horizons generally have more clay than A and C horizons (Figure 2). Areas with different patterns of water flow might have different redistributions of clay in the soil profile. Therefore, the spatial and vertical distribution of clay content depends mainly on the initial amount and type of clay in the parent material, the climate and the relief.

The accumulation of organic matter is another predominant process in the area; organic carbon accumulates mainly in the top layer and can be redistributed to deeper layers by eluviation and pedoturbation. Organic matter accumulation depends on climate and relief, which control temperature and availability of water, land cover, which determines organic matter supply, water infiltration, time and other soil conditions, such as texture and pH (Brady & Weil, 2013).

Another factor that controls CEC is pH. For reasons of simplicity we did not consider pH in the conceptual model.

Graphical and mathematical model

The conceptual model, which characterizes the main forces and processes that control the distribution of CEC, clay and OC, was transformed into a graphical model (Figure 5). Figure 6 shows the variables and model coefficients that have to be estimated from this model. All coefficients are elements of the matrices involved in the definition of the mathematical model. Let us first consider the measurement model (Equations (1) and (2)), which comprises the matrices $\boldsymbol{\Lambda}$, $\mathbf{K}$, $\boldsymbol{\Theta}_{\boldsymbol{\delta}}$ and $\boldsymbol{\Theta}_{\boldsymbol{\epsilon}}$. We assumed that the external factors are observed deterministic variables; therefore, $\boldsymbol{\Lambda}$ is an identity matrix and $\boldsymbol{\Theta}_{\boldsymbol{\delta}}$ is zero. As a result, $\boldsymbol{\xi}$ is equal to $\mathbf{x}$. The matrix $\mathbf{K}$ is also an identity matrix because we assume direct measurement of each soil property, involving only random measurement errors characterized by $\boldsymbol{\Theta}_{\boldsymbol{\epsilon}}$. The diagonal elements of $\boldsymbol{\Theta}_{\boldsymbol{\epsilon}}$ comprise the (known) measurement error variances of each soil property determined with data from an inter-laboratory comparison study (WEPAL, 2015).

Second, the structural model (Equation (3)) is defined by $\boldsymbol{\Gamma}$, $\mathbf{B}$ and $\boldsymbol{\Psi}$. The elements of these matrices have a non-zero value only if there are corresponding arrows in the graphical model. Thus, we obtain:

$$\boldsymbol{\Gamma} = \begin{bmatrix} \gamma_{11} & \gamma_{12} & 0 & \gamma_{14} & 0 & 0 & 0 & 0 & 0 & 0 \\ \gamma_{21} & \gamma_{22} & \gamma_{23} & 0 & 0 & \gamma_{26} & 0 & 0 & 0 & 0 \\ 0 & \gamma_{32} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \gamma_{41} & 0 & 0 & \gamma_{44} & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \gamma_{54} & \gamma_{55} & \gamma_{56} & \gamma_{57} & 0 & \gamma_{59} & 0 \\ 0 & 0 & 0 & 0 & 0 & \gamma_{66} & \gamma_{67} & \gamma_{68} & \gamma_{69} & \gamma_{610} \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \tag{7}$$

$$\mathbf{B} = \begin{bmatrix} 0 & 0 & 0 & \beta_{14} & 0 & 0 & 0 & 0 & 0 \\ \beta_{21} & 0 & 0 & 0 & \beta_{25} & 0 & 0 & 0 & 0 \\ 0 & \beta_{32} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & \beta_{46} & 0 & 0 & 0 \\ 0 & 0 & 0 & \beta_{54} & 0 & \beta_{56} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \beta_{71} & 0 & 0 & \beta_{74} & 0 & 0 & 0 & 0 & 0 \\ 0 & \beta_{82} & 0 & 0 & \beta_{85} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & \beta_{96} & 0 & 0 & 0 \end{bmatrix} \tag{8}$$

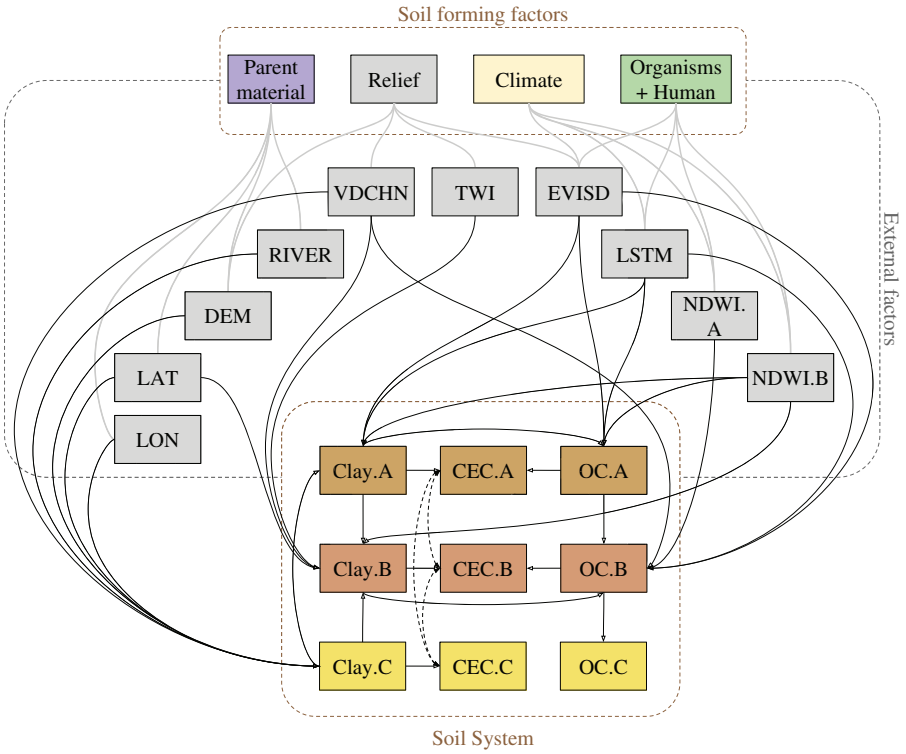

Figure 5 Graphical model. Grey continuous lines represent the theoretical relation between soil-forming factors and external factors. Black continuous arrows are cause and effect links. Black dashed arrows are expected correlations between system errors. External factors are described in Table 1. Soil system variables are abbreviated such that the name of the soil property is followed by the horizon name and separated by a dot, so that Clay.A represents the clay percentage in horizon A, Clay.B is the clay percentage in horizon B, and so on. OC is organic carbon and CEC is cation exchange capacity.

$$\mathbf{\Psi} = \begin{bmatrix} \psi_{11} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & \psi_{22} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \psi_{33} & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \psi_{44} & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \psi_{55} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & \psi_{66} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \psi_{77} & \psi_{78} & \psi_{79} \\ 0 & 0 & 0 & 0 & 0 & 0 & \psi_{87} & \psi_{88} & \psi_{89} \\ 0 & 0 & 0 & 0 & 0 & 0 & \psi_{97} & \psi_{98} & \psi_{99} \end{bmatrix} \quad (9)$$

For example, γ_{12} refers to the arrow in Figure 6 that models the effect of external factor ξ_2 (the standard deviation of the enhanced vegetation index, EVISD) to η_1 (the organic carbon of horizon A, OC.Ar), and β_{54} represents the effect of η_4 (the clay percentage of horizon A, Clay.Ar) to η_5 (the clay percentage of horizon B, Clay.Br). (Letter 'r' at the end of variable names refers to the true value of soil properties (e.g. OC.A is the observed organic carbon of the A horizon, OC.Ar is the true ('real') OC of the A horizon).) Matrix $\mathbf{\Psi}$ has the variances of the structural errors on its diagonal, and allows for non-zero covariance between the CEC structural errors. It is a symmetric matrix (i.e. $\Psi_{ij} = \Psi_{ji}$ for all i and j).

Model calibration and evaluation

The model was fitted with the lavaan package (Rosseel, 2012). After calibration, the measures of model fit were CFI = 0.92, SRMR = 0.043 and GFI = 0.93 (Table 2, step 0). The CFI and P values suggest that there might be some important relations that have not been considered in the model specification. Therefore, we analysed the coefficients and carried out an exploratory respecification analysis that provides suggestions of what can be included in the model.

Model respecification

The first modification of the original model was based on the analysis of its parameters. The coefficient γ_{82} (which linked OC.Br to CEC.Br in Figure 6) was negative. We forced it to be positive, but because this caused convergence problems we decided to remove this link. Next, Clay.Cr and Clay.Br were affected by LAT (γ_{69}, γ_{59}). We expected a positive effect of LAT (latitude) on both soil properties, but because of interaction between LON (longitude) and RIVER (distance to the Paraná River), the coefficients were positive in one link and negative in another. We decided to remove these also (even though they were significant) and replace them with an effect of RIVER on soil properties (γ_{67}, γ_{57}). After these modifications, we obtained new measures of model fit (Table 2, step 1).

Next, we applied an exploratory analysis to respecify the model. We checked suggestions for additional links between external factors and both clay and OC (γ coefficients) with MI, which is a univariate test, and new links have to be included one by one. Table 3 lists the first group of suggestions that were included (step 2). These modifications improved all measures (Table 2, step 2). There were additional relations between soil properties and also several proposed links between CEC and external factors (of all three horizons). Although we know that these are not direct cause and

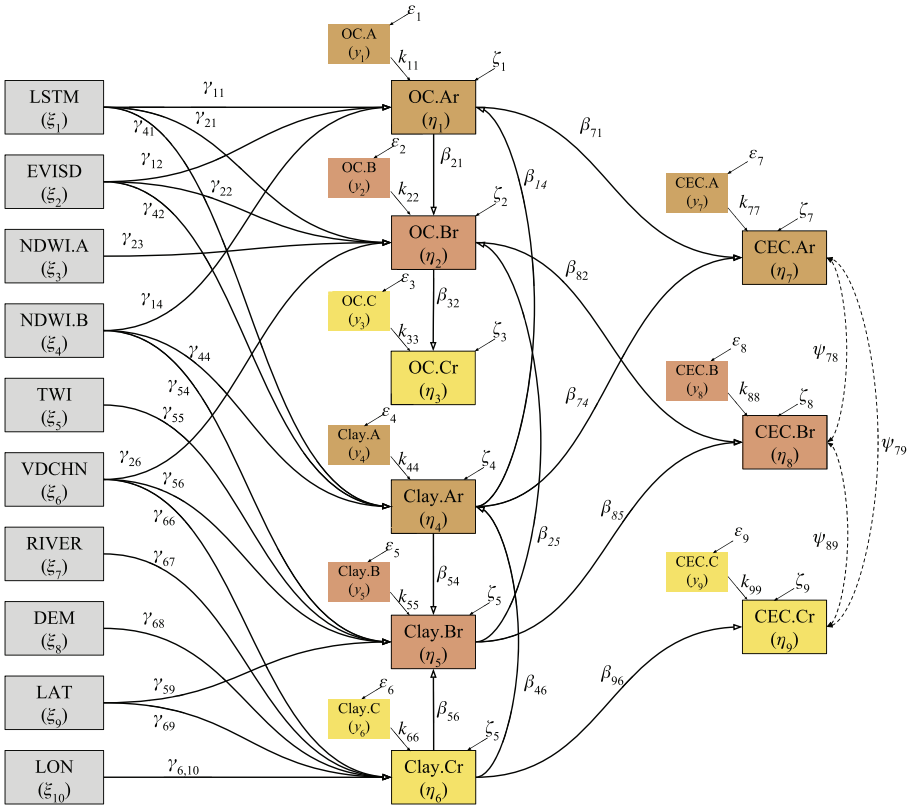

Figure 6 Graphical model with parameters. Thick continuous arrows represent **B** and **Γ** matrices, thin continuous arrows represent **K**, **Ψ** and $\boldsymbol{\Theta}_{\boldsymbol{\epsilon}}$ matrices, and dashed double-headed arrows represent the model error correlations. External factors (grey boxes) are described in Table 1. Soil system variables (coloured boxes) are abbreviated such that the name of the soil property is followed by the horizon name and separated by a dot, so that Clay.A represents the clay percentage in horizon A, Clay.B is the clay percentage in horizon B, and so on. OC is organic carbon and CEC is cation exchange capacity. Letter 'r' at the end of variable names refers to the true value of soil properties (e.g. OC.A is the observed organic carbon of the A horizon; OC.Ar is the true ('real') OC of the A horizon).

Table 2 Changes in model-fitting measures after every respecification step

Step	χ^2	d.f.	*P*-value	CFI	GFI	SRMR
0	228.4	91	<0.000	0.916	0.926	0.043
1	239.5	94	<0.000	0.911	0.924	0.040
2	183.1	86	<0.000	0.941	0.942	0.035
3	127.3	81	0.001	0.972	0.960	0.030
4	90.9	77	0.133	0.992	0.971	0.024

d.f., degrees of freedom; CFI, comparative fit index; GFI, goodness of fit; SRMR, standardized root mean-square residual.

effect relations, they might be caused by intermediate soil properties that were not included in the system, such as pH. Therefore, we decided to include these suggestions (step 3, Table 3). The measures of fit show a large improvement with CFI and GFI close to one (Table 2, step 3).

Finally, we included suggestions for the residual variance–covariance (Table 3, step 4, operator '∼∼') between soil properties because we know that there may be correlation among these that was not identified by the cause and effect relations. Note that the CEC of the A horizon has a positive residual covariance with clay of the B and C horizons, which means that large (small) residuals in CEC.Ar also tend to have large (small) residuals in Clay.Br and Clay.Cr. This might be caused by hidden factors, such as pH and parent material. A similar effect occurs between OC and clay of the C horizon. In this case, depth of the C horizon could account for correlations between the residual errors because it has larger (smaller) clay and OC contents when the upper boundary is closer to (further from) the soil surface. The last modification of the respecification step is to include the residual covariance between CEC of the C horizon and clay of the A horizon, which could also be related to parent material. After this, the measures of fit were acceptable, and we continued with this model (Table 2, step 4).

The respecified model was fitted by maximum likelihood estimation. The resulting graphical model with parameter estimates is shown in Figure 7. Note that NDWI.B and TWI have a small effect only on soil properties, whereas other external factors such as latitude, longitude, distance to the river and the digital elevation model have a strong effect. It is notable that the relations between clay at different horizons, although significant, are not very strong. The relation between OC of the A and B horizons is also very weak, which does not conform to the conceptual model. The main contributors to CEC of the A horizon are clay and OC, whereas CEC of the B and C horizons is primarily governed by clay.

Spatial prediction

Figure 8 shows maps of all soil properties for all horizons. The CEC maps of the B and C horizons have a similar pattern that is affected by proximity to the Paraná River (northeast boundary), which was used to represent parent material. The same pattern also occurs in the maps of clay, which was expected because of the strong relation

Table 3 List of suggestions given by lavaan package

Step	Variable	Operator	Variable	MI
2	OC.Ar	~	LAT	9.09
	Clay.Ar	~	LON	7.66
	OC.Ar	~	DEM	6.89
	Clay.Br	~	LON	5.39
	Clay.Br	~	DEM	9.65
	Clay.Br	~	LSTM	5.58
	OC.Br	~	LON	5.26
	OC.Ar	~	RIVER	5.55
3	CEC.Cr	~	RIVER	30.25
	CEC.Br	~	NDWI.A	9.85
	CEC.Cr	~	LON	4.81
	Clay.Cr	~	NDWI.A	3.50
	Clay.Cr	~	EVISD	7.50
4	CEC.Ar	~~	Clay.Br	9.47
	CEC.Ar	~~	Clay.Cr	8.07
	OC.Cr	~~	Clay.Cr	10.49
	CEC.Cr	~~	Clay.Ar	5.87

Step refers to the steps followed in the respecification process (Results Section, Model respecification). Variable can be either a soil property or an external factor. Operator refers to which kind of relation links the variables (~ 'regressed on', ~~ 'correlated with'). MI is the modification index provided by lavaan. Soil system variables are abbreviated such that the name of the soil property is followed by the horizon name and separated by a dot, so that Clay.A represents the observed clay percentage in horizon A, Clay.B is the observed clay percentage in horizon B, and so on. OC is organic carbon and CEC is cation exchange capacity. Letter 'r' at the end of variable names refers to the true value of soil properties (e.g. OC.A is the observed organic carbon of the A horizon, OC.Ar is the true ('real') OC of the A horizon).

between clay and CEC expressed in the SE model. Figure 8 shows clearly that the vertical variation in OC is much greater than the lateral variation. The OC contents in B and C horizons are very small and almost constant.

Model accuracy and assessment of covariation

Table 4 shows the measures of accuracy derived with cross-validation, and R^2 of the fit of the SEM model. The AVE values show that the model explains a large proportion of the lateral and vertical variation in soil properties. For OC the AVE is 91%, for clay it is 72% and for CEC it is 53%. The AVE decreases when it is calculated per horizon. The AVE for OC is small for all horizons. Clay of the A horizon also has a small AVE value, which explains the poor prediction of the CEC. The AVE for clay of the B and C horizons is relatively large, and so is that for CEC. Figure 9 shows scatter plots of predicted against observed values for the three soil properties, by horizon and for the joint horizons. Results confirm the AVE statistics in Table 4. The MLR gives cross-validation statistics that are similar to those of SEM. The model R^2 of MLR is slightly larger than that of SEM, whereas AVE, which is based on cross-validation, is slightly larger for SEM.

The ME (Table 4) shows that SEM and MLR predictions are unbiased. Prediction error variances of both models give an adequate measure of the uncertainty for most soil properties; the standardized squared prediction error has a mean close to 1, although their medians have slightly smaller values than the theoretical value 0.455. The RMSE shows that prediction accuracy decreases with depth for CEC and clay, which have maximum values of 5.5 $cmol_c\ kg^{-1}$ for CEC.C and almost 7% for Clay.C.

Figure 10 shows the $\mathbf{S}$, $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ and $\boldsymbol{\Sigma}(\boldsymbol{\theta})_{\mathrm{MLR}}$ matrices, which are the standardized variance–covariance matrices of the data, SEM and MLR. Darker colours represent stronger correlations between pairs of soil properties, or between the same soil property at different horizons. It shows clearly that SEM reproduces interrelations more accurately than MLR because similarities are larger between $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ and $\mathbf{S}$ than between $\boldsymbol{\Sigma}(\boldsymbol{\theta})_{\mathrm{MLR}}$ and $\mathbf{S}$. Figure 11 shows the absolute values of $\mathbf{S} - \boldsymbol{\Sigma}(\boldsymbol{\theta})$ and $\mathbf{S} - \boldsymbol{\Sigma}(\boldsymbol{\theta})_{\mathrm{MLR}}$, which confirms this result. Improved performance of SEM is also confirmed by the SRMR, which is 0.024 for SEM (Table 2), whereas $\mathrm{SRMR}_{\mathrm{MLR}}$ is 0.065. All values of the SEM difference matrix are smaller than 0.1, whereas elements of the MLR difference matrix are up to four times larger. For example, covariation between CEC.A and OC.A is not represented adequately by MLR, whereas in SEM it matches the observed covariation much better.

Discussion

The conceptual soil-landscape model

The fitted graphical model in Figure 7 has several implications for the conceptual model. First, it confirms that CEC depends mainly on clay and OC. We also found, however, smaller effects from external factors. This might indicate that another soil property controls CEC that is affected by external factors. For example, Morrás & Moretti (2016) showed that the parent material of this study area varies in its granulometry and mineralogy; the clay mineralogy governs CEC and might be affected by other external factors. We can only assume this relation because we lack a map of soil mineralogy. Second, we decided to remove the relation between OC.B and CEC.B after examining the model parameters, although we know that there is a link between them. In this case, however, clay content of the B horizon is so large in parts of the study area that the effect of OC on CEC becomes negligible. Third, Figure 7 also shows that relations between the A and B horizons are not as strong as we would have expected because the coefficients of OC and clay that connect these two horizons are small. This corroborates Iriondo & Kröhling's (2004) hypothesis, which states that the top horizon of the soil in the study area has another parent material (San Guillermo Formation), namely an aeolian sediment layer of 15–35 cm.

Finally, we observe that there is no direct causality between the CEC of different horizons even when these may be strongly correlated. This is because CEC is a property of the colloidal fraction, which is not affected by the CEC of another layer. For example, CEC of horizon A could be correlated with that of horizon B because they share the same parent material; therefore, they have a similar colloidal fraction.

Figure 7 shows that NDWI of spring (NDWI.B) and TWI have a small effect on soil properties, which means that either their

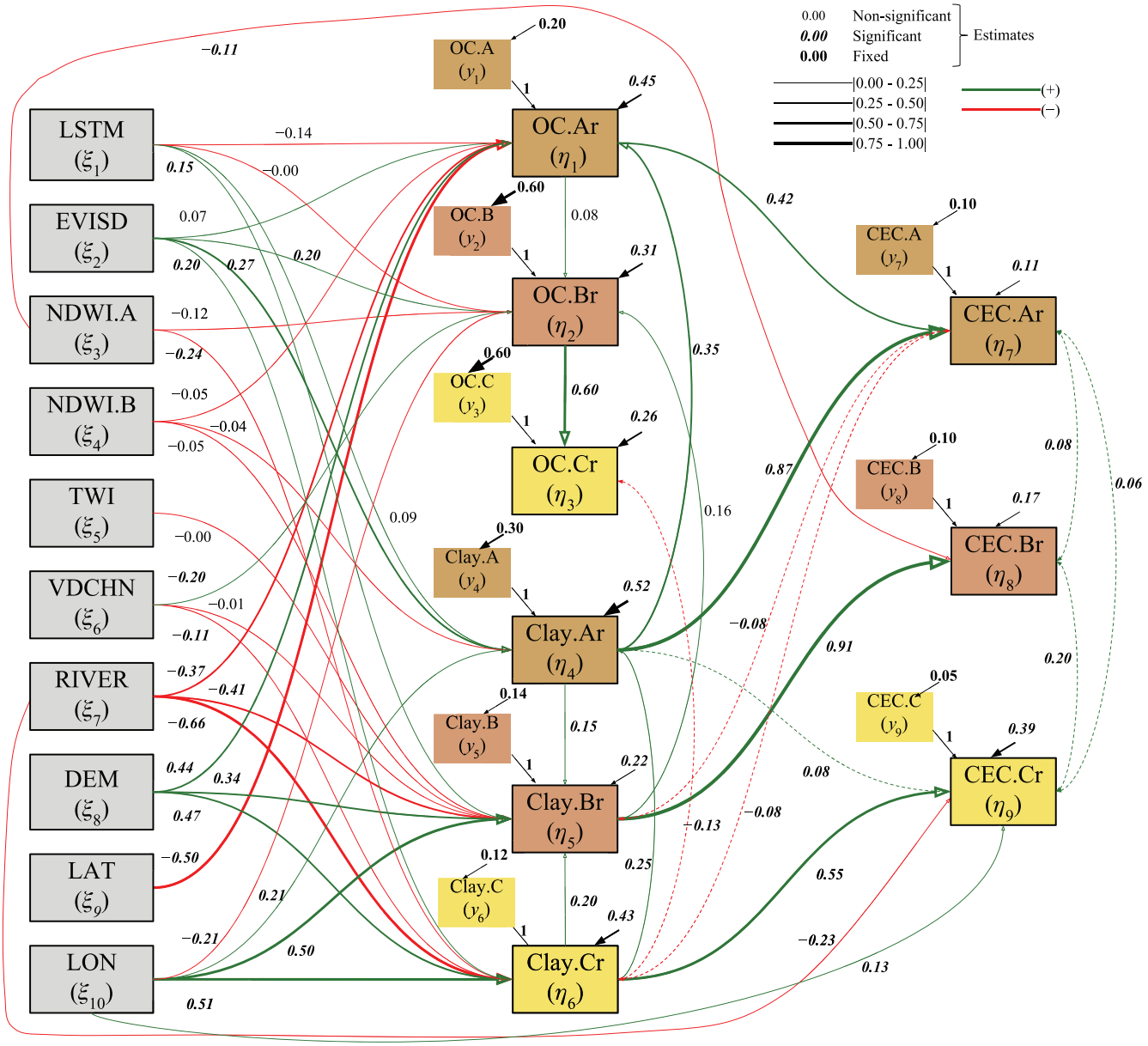

Figure 7 Final graphical fitted model. Arrow thickness represents the magnitude of coefficients and their colour is the sign. Black arrows represent elements of **K**, **Ψ** and $\boldsymbol{\Theta}_\epsilon$ matrices. Dashed arrows represent model error correlations. Bold italic numbers are significant estimates (P-value < 0.05), bold non-italic numbers are fixed coefficients and non-bold non-italic numbers are non-significant estimates. Note that all variables were standardized prior to modelling.

information is redundant or they do not represent the soil-forming factors accurately. This is in contrast to the results of Poggio *et al.* (2013) where NDWI predicted organic matter well. Figure 7 also helps to identify key external factors that have strong predictive power for several soil properties, such as DEM, distance to the Paraná River (representing parent material) and standard deviation of EVI. Incorporating the temporal variation of remote sensing data can increase the resolution of these factors and further increase their predictive power (Samuel-Rosa *et al.*, 2015).

The maps show that the spatial patterns of A-horizon properties differ from those of the B and C horizons. This can be explained by different SEM relations between soil properties and external factors for the A, B and C horizons. It confirms that factors that represent different soil-forming factors differ between horizons.

Model respecification

The model evaluation and respecification steps are the most subjective of an SEM procedure. The main criterion for deciding to modify the model is the lack of fit assessed by different measures (Grace *et al.*, 2012). There is, however, no complete agreement about the cut-off value of these measures because they are case dependent

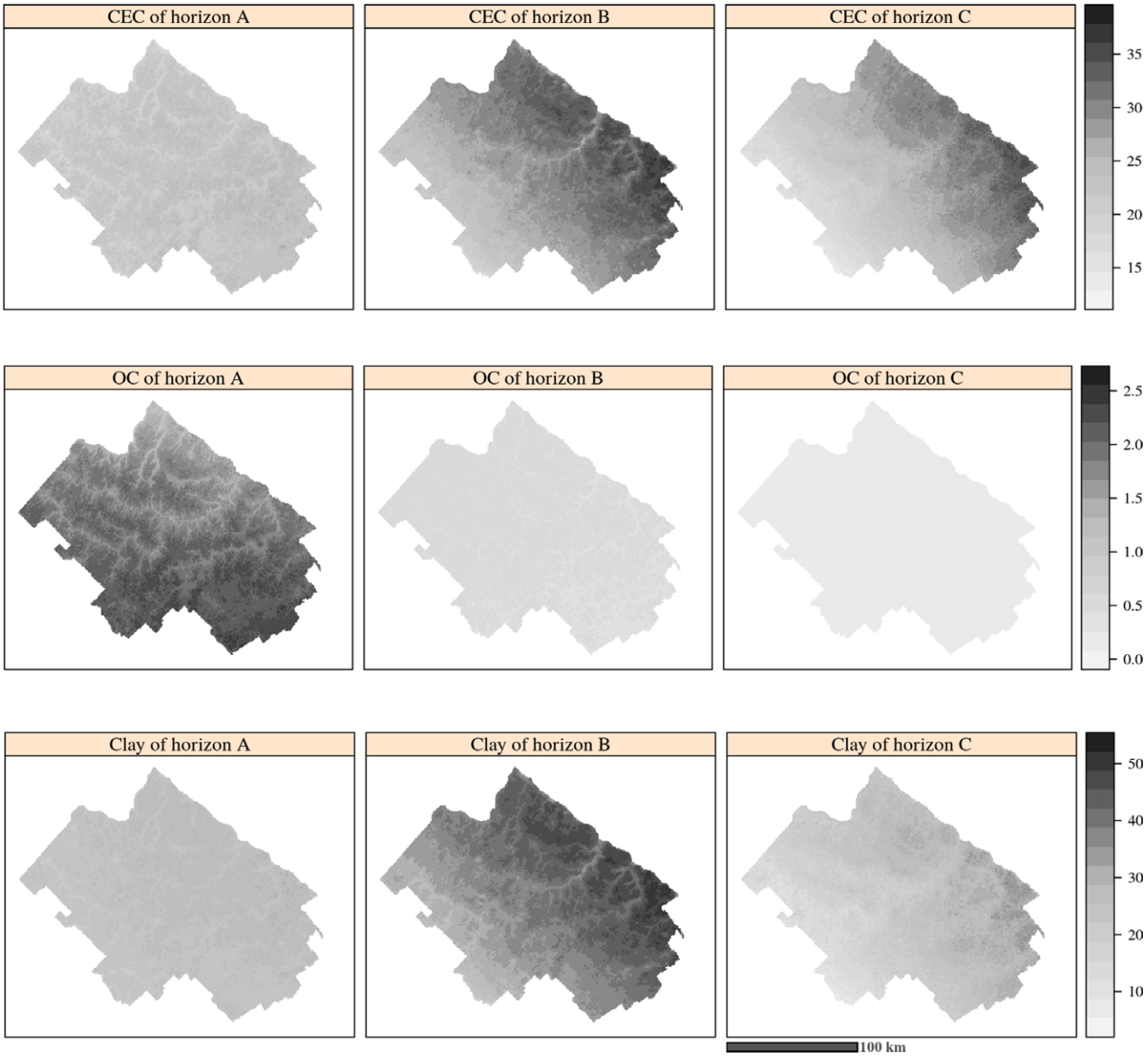

Figure 8 Maps of (a) cation exchange capacity (CEC) ($cmol_c\ kg^{-1}$), (b) organic carbon (OC) (mass %) and (c) clay (mass %) for the A, B and C horizons.

(Marsh *et al*., 2004). Kline (2015) remarked that exploratory analysis may mislead respecification or that it does not help to find the 'truth'. Most SEM applications rarely aim to predict dependent variables as we do in DSM. To achieve greater prediction accuracy, exploratory analysis might identify relevant relations between external factors and soil properties. Although prediction may be improved with exploratory analysis, it should be carried out prudently and with pedological mechanisms in mind.

The question arises as to how far one should go with model respecification. The exploratory analysis can include suggestions until the model fits the data (almost) perfectly, but this does not ensure an improvement in predictive power. It would require independent model validation, which for SEM means applying the fitted model to another independent dataset to prevent over-fitting of the model (Bollen, 1989). We used cross-validation for this without using the observation that was put aside.

Representing soil information with SEM

The resulting SEM graph (Figure 7) in combination with the maps (Figure 8) is a novel way to represent soil information. They show how soil properties and soil layers are interconnected and the effect on their spatial patterns. For example, the similarity in the spatial patterns of clay and CEC of the B horizon can be explained from

Table 4 Cross-validation and measures of model fit

Soil property	ME	RMSE	Mean SSPE	Median SSPE	AVE	R^2
	SEM					
CEC / $cmol_c\ kg^{-1}$	−0.004	4.30	–	–	0.53	–
OC / g 100 g^{-1}	0	0.25	–	–	0.91	–
Clay / g 100 g	−0.007	5.45	–	–	0.72	–
CEC A hor. / $cmol_c\ kg^{-1}$	0.002	3.16	1.03	0.40	0.18	0.21
CEC B hor. / $cmol_c\ kg^{-1}$	−0.009	3.00	1.05	0.39	0.50	0.52
CEC C hor. / $cmol_c\ kg^{-1}$	−0.008	5.46	0.97	0.23	0.45	0.47
OC A hor. / g 100 g^{-1}	0	0.40	1.07	0.38	0.24	0.27
OC B hor. / g 100 g^{-1}	0	0.14	1.06	0.33	0.03	0.06
OC C hor. / g 100 g^{-1}	0	0.06	1.02	0.37	0.02	0.03
Clay A hor. / g 100 g^{-1}	0	4.05	1.03	0.30	0.15	0.18
Clay B hor. / g 100 g^{-1}	−0.013	5.14	0.99	0.40	0.60	0.62
Clay C hor. / g 100 g^{-1}	−0.013	6.87	1.05	0.53	0.41	0.44
	MLR					
CEC A hor. / $cmol_c\ kg^{-1}$	0.006	3.23	1.05	0.38	0.14	0.21
CEC B hor. / $cmol_c\ kg^{-1}$	−0.009	4.04	1.06	0.36	0.49	0.53
CEC C hor. / $cmol_c\ kg^{-1}$	−0.003	5.48	1.03	0.24	0.45	0.49
OC A hor. / g 100 g^{-1}	0	0.41	1.05	0.42	0.22	0.28
OC B hor. / g 100 g^{-1}	0	0.14	1.04	0.35	0.00	0.08
OC C hor. / g 100 g^{-1}	0	0.06	1.05	0.35	−0.05	0.04
Clay A hor. / g 100 g^{-1}	0.007	4.17	1.07	0.31	0.10	0.19
Clay B hor. / g 100 g^{-1}	−0.010	5.21	1.05	0.41	0.59	0.63
Clay C hor. / g 100 g^{-1}	−0.011	6.90	1.04	0.52	0.40	0.45

Mean error (ME), root mean squared error (RMSE), mean and median of the standardized squared prediction error (SSPE) and amount of variance explained (AVE); R^2 is the coefficient of determination of the model fit. OC, organic carbon; CEC, cation exchange capacity; hor., horizon.

the fat arrow between these properties in Figure 7. This indicates that CEC depends strongly on clay content, even in the A horizon where clay (0.87) has twice as large an effect as that of OC (0.42) (recall that all variables were standardized prior to modelling, which means that coefficients can be compared directly).

Model accuracy

The maps of clay and consequently CEC from B and C horizons are reasonably accurate (Table 4 and Figure 9). The maps of OC of the B and C horizons show little spatial variation (Figure 8) and have poor accuracy (Table 4). The latter might be caused by the lack of spatial variation, the small amount of OC in these horizons and relatively large measurement error (Figure 7). Organic carbon and clay of horizon A are poorly predicted, which might be related to the hypothesis that the A horizon is a young sediment (Iriondo & Kröhling, 2004). In general, landscape properties can explain variation in soil properties of the top layers with greater accuracy than for deeper layers (e.g. Kempen *et al*., 2011). In our case it is the other way around for clay and CEC. This could be caused by either a lack of informative covariates or a parent material that is much younger than the subsurface horizons.

Cross-validation results in large AVE values when the three horizons are considered together (Table 4 and Figure 9). More than 91% of the variance in OC was explained by the SE model, 72% for clay and 53% for CEC. This may seem impressive, but this result must be put into perspective. If we used the horizon means only as predictors, about 88% of the variance in the OC data would be explained, 47% of the variance in clay and 15% of the variance in CEC. This confirms that lateral variation of these properties in the study area is much smaller than the vertical variation.

When SEM is compared with multiple linear regression (MLR), R^2 is slightly larger for MLR than SEM (Table 4). This was expected because SEM uses only relations (58 free parameters) that make sense from a pedological point of view, whereas MLR uses all the predictive power in covariates (99 free parameters), regardless of whether the predictive relations make sense pedologically. The AVE based on cross-validation shows that SEM performed slightly better than MLR, which might result from over-fitting of the MLR model. The differences between AVE and R^2 are smaller for SEM than MLR.

Spatial auto- and cross-correlation is not taken into account in SEM by default. The model error (ζ) is assumed independent. Residuals of spatial models, however, might have spatial correlation and taking this into account could help improve predictions (Lamb *et al*., 2014) for the same reason that regression kriging can outperform regression (Hengl *et al*., 2004). Lamb *et al.* (2014) developed a tool to incorporate the spatial autocorrelation among variables in SEM. To determine if our model results could be improved further by taking spatial autocorrelation into account, we fitted variograms (Webster & Oliver, 2007) to the SEM cross-validation residuals (Figure 12). They show that spatial correlation in the residuals of

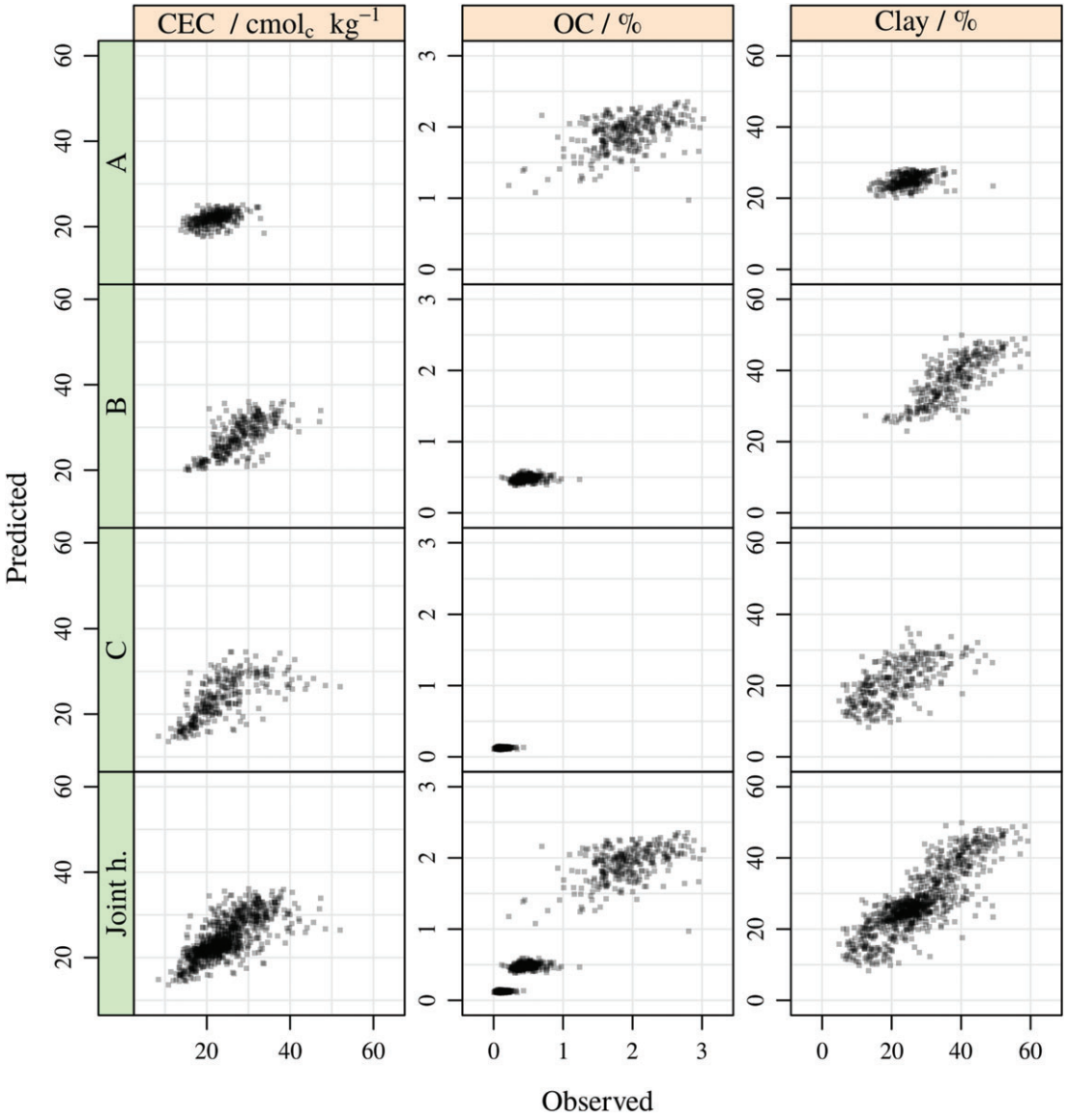

Figure 9 Scatter plots of measured against observed soil properties obtained by cross-validation. Columns of graphs are soil properties: cation exchange capacity (CEC), organic carbon (OC) and clay. Rows of graphs are horizons A, B and C, and 'Joint h.' represents the three horizons joined.

the C-horizon CEC and clay content is moderate and weak in the residuals of the A-horizon clay content. This suggests that there might be room for improvement; therefore, we intend to extend the application of SEM for DSM by taking spatial correlation into account in future.

The SE model reproduced the covariation between soil properties much better than MLR. We compared SEM with MLR because MLR combined with kriging (i.e. regression-kriging) is commonly used in DSM. However, the covariation can also be reproduced by multivariate linear regression (MvLR) (Fox & Weisberg, 2010), which quantifies the cross-correlations between residuals of the linear regressions for each soil property. We fitted an MvLR model to our data with the same covariates that were used for SEM. Assessment of covariation showed that MvLR reproduces the cross-correlations between soil properties perfectly, even better than SEM. This is not surprising because unlike SEM, MvLR puts no restrictions on the residual variance–covariance matrix. All elements can deviate from zero and a perfect reproduction of the cross-correlations can be achieved. An MvLR model is rarely fitted in practice; this approach adds many extra parameters that need to be estimated. In our case, with nine soil properties, the MvLR model would involve $9 \times (9-1)/2 = 36$ extra covariance parameters. With SEM we included only three extra covariance parameters and could reproduce the covariation well. Note that assessment of the covariation was based on the same data that were used to calibrate the models. This might have biased the results

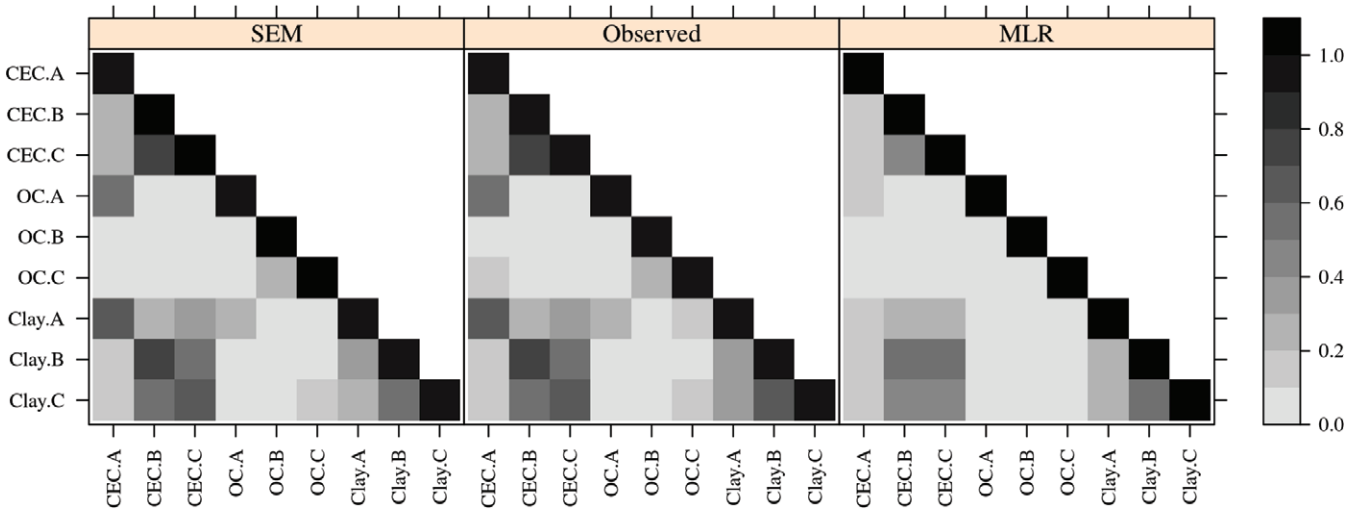

Figure 10 Correlation matrix of observations (Observed), derived from the structural equation (SE) model (SEM) and with multiple linear regression (MLR).

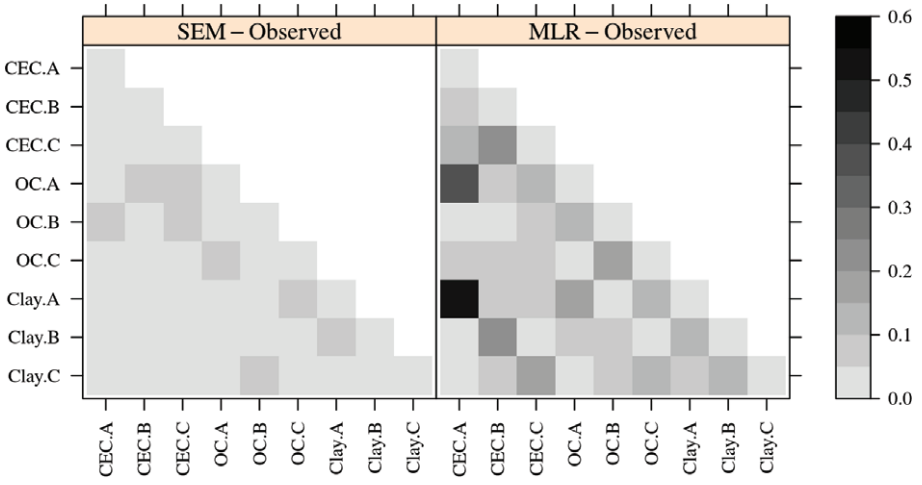

Figure 11 Absolute difference between correlation matrix of original data and structural equation modelling (SEM – Observed) and multiple linear regression (MLR – Observed).

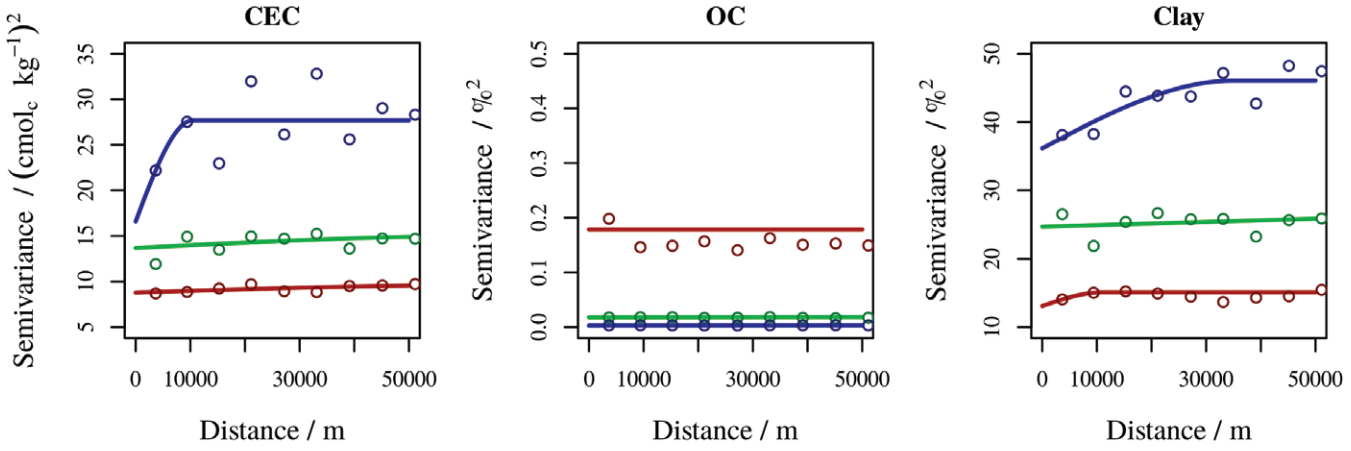

Figure 12 Experimental (dots) and fitted (solid line) variograms of cross-validation residuals of cation exchange capacity (CEC), organic carbon (OC) and clay. Red lines and dots represent the A horizon, green the B horizon and blue the C horizon.

and we should probably have split the dataset into calibration and validation datasets. Reproduction of covariation would probably deteriorate, but less so for SEM than for MvLR.

Conclusions

We have shown how to develop a conceptual model for several soil properties at multiple horizons and how to convert it into a graphical and mathematical model with SEM. We improved model fitting through model respecification and showed how to assess covariation of modelled soil properties.

We conclude that:

- SEM is a useful tool to predict several soil properties simultaneously for multiple horizons while maintaining covariation between soil properties and horizons. Model respecification helps to improve model accuracy and to learn from the data through suggestions that can improve the conceptual soil-landscape model.
- CEC depends largely on clay percentage and less on OC, and so does its prediction.
- SEM graphs in combination with soil maps provide insight into interrelations between soil properties and identify important sources of information that could be used to improve models in future studies.
- A simple method to assess covariation among soil properties could be applied to any DSM approach.
- Prediction of soil properties with separate multiple linear regression models causes inconsistencies between predictions of a soil property. Covariation assessment should be included in modelling that predicts several soil properties or properties at multiple depths.

Acknowledgements

This work was granted by the Argentine National Institute of Agricultural Technology (INTA), approved under resolution 743/12. We thank Dr Héctor J.M. Morrás for valuable comments and discussion.

References

Angelini, M.E., Heuvelink, G.B.M., Kempen, B. & Morrás, H.J.M. 2016. Mapping the soils of an Argentine Pampas region using structural equation modelling. *Geoderma*, **281**, 102–118.

Bentler, P.M. 1990. Comparative fit indexes in structural models. *Psychological Bulletin*, **107**, 238–246.

Bollen, K.A. 1989. *Structural Equations with Latent Variables*. John Wiley & Sons, New York.

Brady, N.C. & Weil, R.R. 2013. *The Nature and Properties of Soils*, 14th edn. Pearson, Harlow.

Fox, J. & Weisberg, S. 2010. *An R Companion to Applied Regression*, 2nd edn. Sage Publications, Thousand Oaks, CA.

Goovaerts, P. 1992. Factorial kriging analysis: a useful tool for exploring the structure of multivariate spatial soil information. *Journal of Soil Science*, **43**, 597–619.

Grace, J.B., Schoolmaster, D.R., Guntenspergen, G.R., Little, A.M., Mitchell, B.R., Miller, K.M. *et al.* 2012. Guidelines for a graph-theoretic implementation of structural equation modeling. *Ecosphere*, **3**, 1–44.

Hengl, T., Heuvelink, G.B.M. & Stein, A. 2004. A generic framework for spatial prediction of soil properties based on regression-kriging. *Geoderma*, **120**, 75–93.

Heuvelink, G.B.M., Kros, J., Reinds, G.J. & De Vries, W. 2016. Geostatistical prediction and simulation of European soil property maps. *Geoderma Regional*, **7**, 201–215.

Iriondo, M. & Kröhling, D. 2004. The parent material as the dominant factor in Holocene pedogenesis in the Uruguay River Basin. *Revista Mexicana de Ciencias Geologicas*, **21**, 175–184.

IUSS Working Group WRB 2015. *World Reference Base for Soil Resources 2014. Update 2015. International Soil Classification System for Naming Soils and Creating Legends for Soil Maps*. Food and Agriculture Organization of the United Nations, Rome.

Jöreskog, K.G. & Sörbom, D. 1981. *LISREL V: Analysis of Linear Structural Relationships by Maximum Likelihood and Least Squares Methods*. Department of Statistics, University of Uppsala, Uppsala.

Kempen, B., Brus, D.J. & Stoorvogel, J.J. 2011. Three-dimensional mapping of soil organic matter content using soil type-specific depth functions. *Geoderma*, **162**, 107–123.

Kline, R.B. 2015. *Principles and Practice of Structural Equation Modeling*. Guilford Publications, New York.

Krause, P., Boyle, D.P. & Bäse, F. 2005. Comparison of different efficiency criteria for hydrological model assessment. *Advances in Geosciences*, **5**, 89–97.

Lamb, E.G., Mengersen, K.L., Stewart, K.J., Attanayake, U. & Siciliano, S.D. 2014. Spatially explicit structural equation modeling. *Ecology*, **95**, 2434–2442.

Lark, R.M. 2000. A comparison of some robust estimators of the variogram for use in soil survey. *European Journal of Soil Science*, **51**, 137–157.

Marsh, H., Hau, K.-T. & Wen, Z. 2004. In search of golden rules: comment on hypothesis-testing approaches to setting cutoff values for fit indexes and dangers in overgeneralizing Hu and Bentler's (1999) findings. *Structural Equation Modeling: A Multidisciplinary Journal*, **11**, 320–341.

Minasny, B. & McBratney, A.B. 2016. Digital soil mapping: a brief history and some lessons. *Geoderma*, **264**, 301–311.

Morrás, H.J.M. & Moretti, L.M. 2016. A new soil-landscape approach to the genesis and distribution of Typic and Vertic Argiudolls in the Rolling Pampa of Argentina. In: *Geopedology – An Integration of Geomorphology and Pedology for Soil and Landscape Studies* (eds J. Zinck, G. Metternicht, G. Bocco & H. del Valle), pp. 193–209. Springer, Dordrecht.

Opolot, E., Yu, Y.Y. & Finke, P.A. 2015. Modeling soil genesis at pedon and landscape scales: achievements and problems. *Quaternary International*, **376**, 34–46.

Orton, T.G., Pringle, M.J., Page, K.L., Dalal, R.C. & Bishop, T.F.A. 2014. Spatial prediction of soil organic carbon stock using a linear model of coregionalisation. *Geoderma*, **230–231**, 119–130.

Poggio, L., Gimona, A. & Brewer, M.J. 2013. Regional scale mapping of soil properties and their uncertainty with a large number of satellite-derived covariates. *Geoderma*, **209–210**, 1–14.

Rosseel, Y. 2012. lavaan: an R package for structural equation modeling. *Journal of Statistical Software*, **48**, 1–36.

Samuel-Rosa, A., Heuvelink, G.B.M., Vasques, G.M. & Anjos, L.H.C. 2015. Do more detailed environmental covariates deliver more accurate soil maps? *Geoderma*, **243–244**, 214–227.

Temme, A.J. & Vanwalleghem, T. 2015. LORICA–a new model for linking landscape and soil profile evolution: development and sensitivity analysis. *Computers & Geosciences*, **90**, 131–143.

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

Webster, R. & Oliver, M. 2007. *Geostatistics for Environmental Scientists*. John Wiley & Sons, Chichester.

WEPAL 2015. *Certificate of Analysis. Reference Material ISE Sample 900*. International Soil-Analytical Exchange [WWW document]. URL http://www.wepal.nl/website/download_files/consensus/ISE/ISE900.pdf [accessed on 8 November 2016].

Xu, S., An, X., Qiao, X., Zhu, L. & Li, L. 2013. Multi-output least-squares support vector regression machines. *Pattern Recognition Letters*, **34**, 1078–1084.