Structural equation modelling for digital soil mapping

Marcos E. Angelini

Structural equation modelling for digital soil mapping

Marcos E. Angelini

INVITATION

You are cordially invited to attend the public defence of

Structural equation modelling for digital soil

Tuesday, 6 March 2018
at 1:30 pm in the Aula of
Wageningen University,
Generaal Foulkesweg 1a,
Wageningen

Marcos E. Angelini
angelini75@gmail.com

PARANYMPHS

Cecilia C. Contessotto
englisht@gmail.com

Federico G. Olmedo
guillermo.olmedo@gmail.com

Structural equation modelling for digital soil mapping

Marcos E. Angelini

Thesis committee

Promotor
Prof. Dr Gerard B.M. Heuvelink
Special Professor Pedometrics and Digital Soil Mapping
Wageningen University & Research

Co-promotor
Dr Bas Kempen
Researcher, ISRIC – World Soil Information, Wageningen

Other members
Prof. Dr Rachel E. Creamer, Wageningen University & Research
Dr Peter A. Finke, Ghent University, Belgium
Dr Laura Poggio, The James Hutton Institute, Aberdeen, United Kingdom
Dr Willem T. Kruijer, Wageningen University & Research

This research was conducted under the auspices of the C.T. de Wit Graduate School for Production Ecology and Resource Conservation.

Structural equation modelling for digital soil mapping

Marcos E. Angelini

Thesis
submitted in fulfilment of the requirements for the degree of doctor
at Wageningen University
by the authority of the Rector Magnificus
Prof. Dr A.P.J. Mol,
in the presence of the
Thesis Committee appointed by the Academic Board
to be defended in public
on Tuesday 6 March 2018
at 1:30 p.m. in the Aula.

Marcos E. Angelini
Structural equation modelling for digital soil mapping, 186 pages.

PhD thesis, Wageningen University, Wageningen, the Netherlands (2018)
With references, with summaries in English and Spanish

ISBN 978-94-6343-734-9
DOI https://doi.org/10.18174/430659

Dedicado a esos maravillosos seres que formaron parte de esta experiencia
Sofía, Bruno, Amelie e Indira

“Truth is a pathless land”
Jiddu Krishnamurty, 1929

Index

Chapter 1

General introduction

1.1. Background

Climate change and land degradation are of increasing societal and governmental concern. For this reason several international programs have been initiated in recent years, such as the Land Degradation Neutrality concept[1] (SDG 15.3) launched by the UN Convention to Combat Desertification (UNCCD) and the UN Environment Programme (UNEP) in Rio+20 in 2012, the Sustainable Development Goals (SDG) of the United Nations[2] (UN) defined in 2014 and the *4 per 1000* initiative (Conference of the Parties to the United Nations Framework Convention on Climate Change in Paris in 2015[3]). In order to measure, monitor and predict the impact of climate change and the magnitude of land degradation, soil information in space and time at national, regional and global scale is essential. The soil science community and national institutions are actively working under the Global Soil Partnership[4] (Food and Agriculture Organization of the UN) and the *GlobalSoilMap*[5] consortium, among other organisations and networks, to provide regional and global soil information. This means that soil observations collected at sampling sites need to be upscaled to soil maps at different spatial scales.

In the past, soil surveyors described the soil-landscape system based on a conceptual mental model, supported by spatial information, such as topographic maps and aerial photographs, field descriptions and laboratory analysis of soil samples. Soil characteristics were summarised through classification systems, which also helped communication between soil scientists. The main products of these methods were polygon maps and associated reports, and the main goals were to value land and support agricultural planning (Brevik and Hartemink, 2010). With the expansion of technology in many fields, more and more tools became available, such as remote and proximal sensors, internet, GPS, database infrastructures, as well as (geo)statistical methodologies (Brevik et al., 2016). These developments inspired researchers to investigate ways of updating soil mapping methodologies, which finally led to the emergence of a new sub-discipline in soil science named "Digital Soil Mapping (DSM)" (McBratney et al., 2003) and the establishment of an international working group on digital soil mapping under the International Union of Soil Sciences. DSM is defined as "*the creation and population of geographically referenced soil databases generated at a given resolution by using field and laboratory observation methods coupled with environmental data through quantitative relationships.*"[6].

[1] http://www2.unccd.int/land-degradation-neutrality

[2] https://www.un.org/sustainabledevelopment/

[3] http://4p1000.org/

[4] http://www.fao.org/global-soil-partnership/en/

[5] http://globalsoilmap.net/

[6] http://digitalsoilmapping.org/

According to Minasny and McBratney (2016), DSM has three components: (1) input, which refers to the data and methods used to observe and collect soil data and correlated environmental data; (2) methods used to process the input data, that can be any soil inference system; and (3) output, or resulting soil information that may include rasters of the predicted soil variables along with their associated uncertainty. Most studies in DSM spatially predict soil properties or classes from (either new or legacy) laboratory data and spatially exhaustive environmental covariates using empirical models, such as regression kriging, artificial neural networks, (boosted) regression trees and random forest (e.g. Hengl et al., 2004; Were et al., 2015; Yang et al., 2016). During the last decade, DSM matured and became an accepted mapping method in the soil science community. For example, it has been included in the Soil Survey Manual of the USDA (Soil Science Division Staff, 2017) and has been proposed as a standard methodology for the Global Soil Organic Carbon map (FAO, 2017). Also, it has been used for national soil mapping in many countries (e.g. Kempen et al., 2015; Viscarra Rossel et al., 2015; Mulder et al., 2016; Padarian et al., 2017).

However, Brevik et al. (2016) notes that “linking all of this new information to soil properties and processes can still be a challenge and enhanced pedologic models are needed”. The large volume and high dimensionality of the new data complicates the efficient and effective linkage of the data to soil properties and processes. One approach is to rely on highly empirical “machine-learning” approaches (Fitzpatrick et al., 2016; Hengl et al., 2017), but this runs the risk of ending up with models that are difficult to interpret and do not truly advance our understanding of the soil and soil functioning. An alternative approach is to make better use of pedological knowledge in extracting information from new data sources for use in DSM. The latter route is taken in this thesis.

1.2. Problem definition and opportunities

1.2.1. Digital soil mapping

The standard procedure of most digital soil mapping techniques is to build statistical predictive models that relate field and lab data from soil observations to proxies of soil-forming factors, which are spatially exhaustive and typically referred to as environmental covariates (Minasny and McBratney, 2016). Next the model is applied to unvisited locations, typically the nodes of a regular grid covering the mapping area, to predict the soil type and/or soil properties. Even though some studies apply three-dimensional modelling of soil properties (e.g. Adhikari et al., 2013; Mulder et al., 2016; Poggio and Gimona, 2017), most DSM studies tend to represent the spatial variation of soil properties individually and for individual depth layers (Grunwald, 2009). Among the statistical models, regression kriging (Hengl et al., 2004), and its variants, has been one of the most commonly applied methods. It models the relation between soil observations and covariates with a linear regression approach and the spatial autocorrelation of the regression residuals with a variogram (Hengl et al., 2017). The fitted linear regression model is then applied to a stack of gridded covariate layers to predict the soil property of interest across a mapping area, while the regression residuals are kriged to the prediction grid. The linear regression predictions and kriged residuals are added together to form the regression kriging predictor. However, the number of covariates to represent soil-forming processes, generally derived from remote sensing data, has increased exponentially in the last decade (Kuenzer et al., 2015). To analyse this multi-dimensionality, a family of methods known as machine learning, are now replacing the linear regression approach used in regression kriging. Classification and regression trees, random forest regression, support vector machines, artificial neural networks and many other variants are some of the methods used for this purpose (Kuhn and Johnson, 2013).

The empirical methods have shown to be able to produce accurate maps at various spatial scales (Kempen et al., 2015; Heuvelink et al., 2016; Hengl et al., 2017; Malone et al., 2017), but they do not provide knowledge about the interrelationships between the soil components and their functioning. In traditional soil survey, the complex interrelations among several soil properties are captured by the pedon concept and described in soil survey reports. In DSM, methods that incorporate soil forming process knowledge for spatial prediction are limited (Kempen et al., 2009). Incorporating process knowledge in DSM would advance our understanding of soil and soil formation, it would guide us in deciding which new data (from remote sensing products) should be used in DSM from a pedological point of view, instead of relying on highly empirical machine-learning approaches that have their own limitations, and presumably, it would also lead to models that have better extrapolation performance.

1.2.2. Including process knowledge in DSM

The development of theories on soil formation has a rich history in pedology. Dokuchaev (1883) and later Jenny (1941) developed the concept of soil-forming factors and the theoretical *CLORPT* model. Simonson (1959), based on Glinka (1927), defined the concept of pedon (the soil body) and developed the "Generalized Theory of Soil Genesis", which organises the soil forming processes in additions, removals, transfers and transformation of soil constituents. These theories support the understanding of the soil system, how soils behave in terms of functioning and how soil properties connect to one another. Furthermore, these theories have been used by surveyors as a keystone in soil mapping, since they were fundamental for developing a mental model of spatial soil distribution. However, mental models are only conceptual and mainly descriptive, rather than concrete models or functions that can be used for (quantitative) prediction.

Based on Glinka and Simonson's theories of soil genesis, behaviour and functioning, researchers began developing mechanistic models of soil-forming processes (e.g. Boast, 1973; Runge, 1973; Kirkby, 1977; Phillips, 1993; Huggett, 1998). Later, the development of landscape models (e.g. Moore and Burch, 1986; Moore et al., 1991), allowed further development of soil-landscape evolution models (e.g. Schoorl et al., 2002; Temme et al., 2006; Vanwalleghem et al., 2010; Finke, 2012a; Vanwalleghem et al., 2013; Stockmann et al., 2014; Opolot et al., 2015; Temme and Vanwalleghem, 2016). These models are attractive because they mimic the physical, chemical and biological processes that shape the soil. However, these processes are very complex and not all fully understood. Moreover, mechanistic soil-landscape models are governed by numerous initial and boundary conditions that are often poorly known. Small variations in the initial and boundary conditions may produce large differences in the model output. These models are also computationally challenging because they operate in three-dimensional space and time. Thus, the development of mechanistic soil models has not yet reached a stage in which these models can produce accurate soil maps that can compete with empirical DSM approaches.

In summary, conventional DSM techniques are empirical and do not include soil process knowledge very well. Mechanistic soil-landscape models include process-knowledge but cannot be applied easily for soil mapping, because of their high complexity and large uncertainty. One solution to incorporate soil process knowledge in DSM could be to develop a hybrid approach that combines elements of empirical and mechanistic models.

The use of pedological expert knowledge in DSM, as done in conventional mapping, is scarce and limited to only a few publications. The *CLORPT* concept is softly implemented through purely empirical methods. There are some examples however, that combine expert knowledge with statistical approaches for soil mapping more explicitly. For example, Bui (2004) concluded that the mental model of a soil surveyor could be structured in a reproducible hierarchical framework, although this idea was not implemented in practice. Zhu et al. (2001) tried to reproduce the mental process of soil surveyors using a GIS-based model and fuzzy logic techniques. Kempen et al. (2009) combined expert knowledge through a conceptual model and logistic regression to update a soil map. These examples, however, targeted to predict soil classes instead of continuous soil properties. McKenzie and Gallant (2006) developed an approach in which pedological rules connect environmental covariates to soil classes. They also measured covariates in the field to set limits for the pedological rules and predicted some soil properties for each soil class. This is a step forward, but it still does not include interrelations between soil properties. More recently, Taalab et al. (2015) implemented a Bayesian network approach to model soil-landscape relations and predict soil classes and bulk density. Although the authors acknowledge the limitations of the approach to model continuous soil variables, they stressed that the model is easily interpretable because it is based on cause–effect relationships.

Up to now, most studies that incorporate soil process knowledge in a statistical mapping framework are limited to modelling and mapping soil classes and not soil properties. None of the studies modelled the multivariate dimension of soil using structural equation modelling (SEM). SEM is built on the basis of process knowledge, can model continuous soil properties, and can take soil property interrelationships into account. These characteristics of SEM suggest that it is a very useful technique to bridge the gap between empirical and mechanistic approaches for DSM.

1.2.3. Causal analysis and SEM

Before explaining SEM and its potential application in DSM, it is useful to address an aspect of SEM related to its classical use. Judea Pearl, one of the most renowned researchers in causal analysis, stated that SEM is the "*primary language of causal modelling*." Causal analysis, or causal modelling, is the probabilistic analysis of causality, which refers to cases where one fact will likely cause another fact. The use of SEM in causal analysis has been controversial (Pearl, 1998), since the word *cause* is not part of the vocabulary of most probabilistic statisticians. In their view, observational studies can only reveal correlations, not cause–effect relationships. To clarify his position, Pearl (1998) described the causal interpretation of SEM and defined the proper use of this framework. One of its main requirements in SEM is to have previous knowledge of a causal relation. This is the cornerstone of SEM that makes it appealing for pedometricians. The idea to propose a causal structure behind the probabilistic distribution of the data is to cope with possible interventions to manipulate system variables.

1.3. Structural equation modelling

SEM has been developed to test hypotheses that explain the relationships in a system (Grace et al., 2012). It uses a conceptual model to define how the system variables connect to one another and uses empirical data to check whether the assumed relationships represent reality. The roots of SEM are diverse, since the need to link hidden processes with observations has a long history in many research fields (Hägglund, 2001). The most significant developments in mathematics that led to the development of SEM however, took place during the first part of the 20th century. SEM has adopted features from factor analysis developed by Spearman (1904) and Lawley (1940) in psychology. Wright (1921) provided a framework to study cause–effect relationships in biology, and although he did not refer to it as SEM, his methodology used several features that SEM has nowadays, such as graph models and path analysis. Years later, Haavelmo (1943) developed in econometrics the "system of simultaneous equations", which was similar to Wright's work. During the following years there was a vast development in these fields that produced a robust framework for SEM, which was well compiled by Bollen (1989). The number of publications in ecology using SEM has increased rapidly since 2000 (Grace et al., 2010). What makes SEM an attractive modelling method for ecologist is its capability to connect data to theories by constructing latent variables that represent concepts (Grace et al., 2010).

The main components of SEM are a conceptual model, a graphical model, a system of equations and observational data. The *conceptual model* describes the processes that occur in a system, explaining the relationships between system variables in such a manner that the expected correlation between these becomes evident. Next these relationships are translated to a *graphical model*, which is composed of boxes that represent variables and arrows that represent the relations between them. An exceptional feature of SEM is that it can include *latent variables* (generally represented with ellipses instead of boxes), which are abstract conceptualizations of the real world that cannot be measured directly, but only indirectly through observed variables. In a soil science context, examples of latent variables might be soil health, soil fertility and soil maturity. Latent variables may also be used to represent the true values of soil properties, which are unknown because of measurement errors. There are also different types of relations that are represented in the graphical model with different arrows. For example, cause–effect relationships are indicated with one-headed arrows indicating the direction of the effect, whereas double-headed arrows are used to show correlations in residuals, useful for instance when an unavailable external factor affects two system variables. Finally, the graphical model also includes model parameters, which can be fixed or free. Fixed parameters are those that are known, such as the *measurement error* variance of a lab instrument, while free parameters are estimated in the calibration process. The system of equations is the mathematical representation of the graphical model. It has a *measurement model* and a *structural model*. The first specifies how the latent variables are measured. The second defines the cause–effect relationships between latent variables. The free parameters are estimated by optimising a function, usually the likelihood function, to make the model-implied variance–covariance matrix agree with the sample variance–covariance matrix. When the difference between these matrices is large, it is possible to obtain suggestions from the model for improvement of the model structure. This means that it is also possible to consider relationships that were previously unknown by the researcher. Once the model is calibrated, the system of equations can be used to predict *endogenous system variables* (i.e., soil properties) from *exogenous variables* (i.e., environmental covariates).

We, as pedologists, are interested in understanding processes, just as ecologists do, but we also aim to make soil maps, in other words (spatially) predict soil properties. So far, prediction has not been done using SEM, but it is not a very difficult step after the model has been set-up and calibrated. The reasons that SEM might be a useful tool for DSM are: (1) we know the conceptual relationships between soil properties, and between soil properties and environmental covariates, so we can define a conceptual and a graphical model; (2) we have soil observations and covariate data needed to calibrate the model; (3) we can use the calibrated model to make spatial predictions (soil maps); and (4) we can interpret the SEM model to advance our understanding of soil processes in a region. Therefore, I think that SEM has several features to bridge the gap between purely empirical approaches and process-based modelling. I think that it should allow us to explicitly include pedological knowledge in the mapping process, to model several soil properties and soil layer depths simultaneously, and to represent our mental model in a graphical way, which is easy to understand. Moreover, if a model is created on the basis of pedological knowledge, it should be easier to extrapolate from one region to another. But I also acknowledge that the spatial correlation in the observations is not taken into account in SEM, which is an important feature for mapping and would require a modification of the current SEM approach.

1.4. Objectives

The overall objective of this PhD thesis is to extend DSM with soil process information through the *development*, *calibration*, *application* and *validation* of a structural equation (SE) model.

To achieve this objective, I define four specific objectives with their respective research questions, as follows:

1. **To develop and apply SEM in DSM** (Chapter 2).

    - How can soil process knowledge and soil property interrelations be incorporated in a SE model?
    - Are knowledge-based interrelations more efficient than data-driven interrelations for prediction?
    - Which are the implications of including measurement error in SEM?
    - What does the graphical model contribute to DSM?

2. **To test SEM for multiple layers and multivariate mapping of soil properties, as well as to implement tools that can improve model performance** (Chapter 3).

    - How suitable is SEM for multiple layer and multivariate mapping of soil properties?
    - What can we learn from the data that we did not know *a priori*?
    - Does SEM represent the covariation between predicted soil properties better than empirical statistical approaches?

3. **To explore the capabilities of SEM for model extrapolation** (Chapter 4).

    - Is SEM more capable than purely empirical methods to extrapolate from a region to another region with similar soil-landscape conditions?
    - To what extent does the improvement of a model for a specific region affect its extrapolation capability?
    - How different are SE models between areas with similar soil-landscape conditions?

4. **To include the geostatistical component in SEM** (Chapter 5).

    - How can spatial correlation be accounted for when calibrating a SE model?
    - What is the impact of residual spatial correlation on the estimated model parameters?
    - Does the spatial SEM approach result in more accurate soil maps?

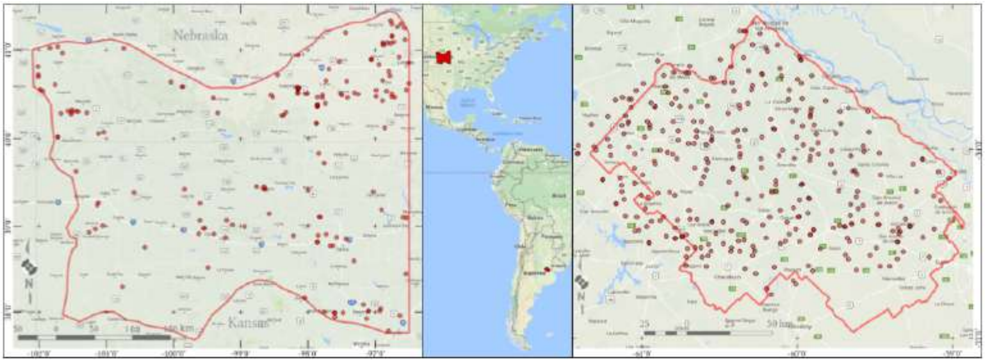


Figure 1.1: *Study areas used in this research. Argentinian study area to the right, US study area to the left, reference map in the centre. Red circles are locations of soil profiles used in the different studies.*

1.5. Study areas

To illustrate the use of SEM for DSM I use two study areas (Fig. 1.1), one in Argentina and one in the United States. The Argentinian study area is located in the Rolling Pampas, a sub-region of the Argentinian Pampas, and encompasses ca. 23 000 km^2 in the north-eastern corner of the Buenos Aires Province. The dominant soil classes are Phaeozems (Argiudolls and Argiaquolls in Soil Taxonomy), associated with Solonetz (Natracuolls and Natruacualfs in Soil Taxonomy) that developed in aeolian sediments. Precipitation ranges between 900 and 1000 mm per year.

The study area in the United States is a ca. 150 000 km^2 area located in the Great Plains, covering parts of Kansas and Nebraska. The Platte and Arkansas rivers form the northern and southern boundary. The eastern boundary runs north-south through Manhattan, KS, while the west boundary is defined by the foot slope of the Rocky mountains. Precipitation varies from about 800 mm in the east to about 500 mm in the west of the study area.

The Argentinian study area was selected based on the availability of soil data for a relatively large geographical area that has a relatively small variation in biophysical conditions, which simplified the consideration of soil-forming processes. Similar criteria were used to select the study area in the United States, with the addition that the area should have similar soil types and parent material as the Argentinian study area. The selected US study area is much larger than the Argentinian one. The main reason for this is that we needed a larger area to have sufficient soil profile data available to calibrate the SE models. Both study areas were originally covered by prairie grasslands. The Argentinian area has been largely converted to cropland, while the US area has only been partially converted to cropland. This is because the US area has a sub-humid to semi-arid climate, while the Argentinian one is humid. Even though the study areas were chosen because of their similarities, the differences in specific factors ans processes of soil formation will likely influence the SE models developed for both areas.

The Argentinian study area was used for the first, second and third specific objectives (Chapters 2 to 4), while the US study area was used to examine the extrapolation potential (Chapter 4) and application of the spatial SEM approach (Chapter 5).

1.6. Expected contributions

The aim of this research is to fill the gap between purely empirical and mechanistic DSM models by analysing the usefulness of a hybrid (empirical-mechanistic) approach for DSM. This is not only attractive because it may enrich the DSM toolbox and outperform existing approaches for soil mapping, but also because it helps soil mappers and soil scientists getting a better understanding of local soil-forming processes. Models of local and regional processes can be hypothesised and tested with empirical data, while they can also be used for mapping. Studying causal relations is very common in soil genesis, but much less so in soil mapping, and so this research might also contribute to bridging these two soil science disciplines.

Chapter 2

Mapping the soils of an Argentine Pampas region using structural equation modelling

Current digital soil mapping (DSM) methods have limitations. For instance, it is difficult to predict a large number of soil properties simultaneously, while preserving the relationships between them. Another problem is that prevalent prediction models use pedological knowledge in a very crude way only. To tackle these problems, we investigated the use of structural equation modelling (SEM). SEM has its roots in the social sciences and is recently also being used in other scientific disciplines, such as ecology. SEM integrates empirical information with mechanistic knowledge by deriving the model equations from known causal relationships, while estimating the model parameters using the available data. It distinguishes between endogenous and exogenous variables, where, in this application, the first are soil properties and the latter are external soil-forming factors (i.e. climate, relief, organisms). We introduce SEM theory and present a case study in which we applied SEM to a 22 900 km^2 region in the Argentinian Pampas to map seven key soil properties. In this case study, we started with identifying the main soil-forming processes in the study area and assigned for each process the main soil properties affected. Based on this analysis we defined a conceptual soil-landscape model, which was subsequently converted to a SEM graphical model. Finally, we derived the SEM equations and implemented these in the statistical software R using the latent variable analysis (`lavaan` package). The model was calibrated using a soil dataset of 320 soil profile data and 12 environmental covariate layers. The outcomes of the model were maps of seven soil properties and a SEM graph that shows the strength of the relationships. Although the accuracy of the maps, based on cross-validation and independent validation, was poor, this paper demonstrates that SEM can be used to explicitly include pedological knowledge in prediction of soil properties and modelling of their interrelationships. It bridges the gap between empirical and mechanistic methods for soil-landscape modelling, and is a tool that can help produce pedologically sound soil maps.

Based on:
Angelini, M. E., Heuvelink, G. B. M., Kempen, B. and Morrás, H. J. M., 2016. Mapping the soils of an Argentine Pampas region using structural equation modelling. *Geoderma*, **281**, pp.102-118.

2.1. Introduction

Numerous environmental and agro-economic activities require accurate information on the spatial distribution of soil types and properties. Conventional soil maps, which supply this information, are based on a conceptual (mental) model of soil spatial variation supported by field observations and laboratory data. Although conventional soil survey can produce accurate soil maps, it has several drawbacks (Hartemink et al., 2010). First, map units are represented as homogeneous units so that spatial variation within these units is not made explicit, as it is only explained in a narrative way through reports and map legends. Second, although validation can be done as an independent procedure, the conventional soil mapping process does not assess map accuracy, and thus, maps typically lack uncertainty information. Third, the qualitative mental models and soil mapping rules used to generate the maps are often not documented and, therefore, not easily reproducible (Hewitt, 1993). In this context, new techniques to produce soil maps were introduced in the past decades and summarised under the name "Digital Soil Mapping" (DSM) (McBratney et al., 2003).

Most DSM approaches are data-driven and rely heavily on empirically established relationships between the soil and the environment (Dobos et al., 2006; Kempen et al., 2009; Cambule et al., 2013; Adhikari et al., 2014). Methods that incorporate soil-forming process knowledge for spatial prediction using DSM, such as Kempen et al. (2009), are limited, partly because the complex interrelationships between soil-forming factors, soil-forming processes and soil properties are not easy to capture and quantify. However, incorporation of soil process knowledge in DSM approaches is attractive, because it provides an important source of information and should yield maps that better represent the physical, chemical and biological processes that shape the soil. Including soil process knowledge could bring benefits both for DSM and soil process modelling (Heuvelink and Webster, 2001; Stoorvogel et al., 2009; Vanwalleghem et al., 2010; Finke, 2012b).

Most DSM studies describe the spatial variation of a single soil property (Grunwald, 2009). Separate mapping of multiple soil properties might yield predictions that make sense individually, but whose combination may be unrealistic. For instance, if the carbon and nitrogen content of the soil are mapped independently, this may result in implausible C:N ratios that may affect agronomic evaluations. The complex interrelationships among soil properties are captured in conventional soil survey by soil survey reports and pedon descriptions. However, soil properties interdependencies have not yet been included efficiently in DSM. Some techniques, such as cokriging, can deal with multiple, correlated soil properties in a unified approach, but fitting a valid model of coregionalization becomes very problematic when the number of soil properties is large (Knotters et al., 1995; Heuvelink, 2006) and it does not add information about process knowledge.

In this study, we attempt to integrate knowledge about soil-forming processes and interrelationships between soil properties in DSM by applying a statistical technique known as structural equation modelling (SEM) (Grace and Keeley, 2006). The roots of SEM are in the social sciences (Sobel, 1982; Pearl, 1988), but recently it has also been applied in environmental science (Grace et al., 2010; Brahim et al., 2011; Grace et al., 2014; Lamb et al., 2014). To the best of our knowledge, it has not yet been applied in soil mapping. SEM could be used not only to predict soil properties spatially, but also to help understand and explain the complex interrelationships behind the development and evolution of soil properties. The aim of this study is to explain the principles of SEM and describe how it may be used for soil mapping. We illustrate the methodology with a case study in the Argentine Pampas.

2.2. Structural equation modelling

2.2.1. SEM overview

SEM is a methodology for developing and testing hypotheses about relationships in a system, and encompasses different statistical tools for causal analysis. Although the focus is on testing hypotheses, it can also be used for prediction. It includes both graphical and equational forms. Graphs are not only a representation of the cause–effect network, but also a tool to identify requirements for model fitting (Grace and Keeley, 2006). According to Rosseel (2012), SEM includes elements from three different statistical techniques: (1) factor analysis, developed by Spearman (1904) and Lawley (1940) in psychology, (2) path analysis, popularised by Duncan (1966) in social science, and (3) simultaneous equation systems, developed by econometricians (Haavelmo, 1943; Koopmans, 1945).

Incorporating cause–effect relationships in statistical models induces correlations between system state variables and driving factors, but it is important to be aware that correlation is not the same as causality. This common issue in multivariate analysis was discussed in-depth in the SEM literature by various authors (Pearl, 1988, 2009; Shipley, 2000; Grace, 2006), who established basic procedures to apply SEM and note that it cannot be used to prove causality, but merely to test models that represent causal hypotheses (Grace, 2006).

Although SEM builds on conventional regression techniques, its application is more

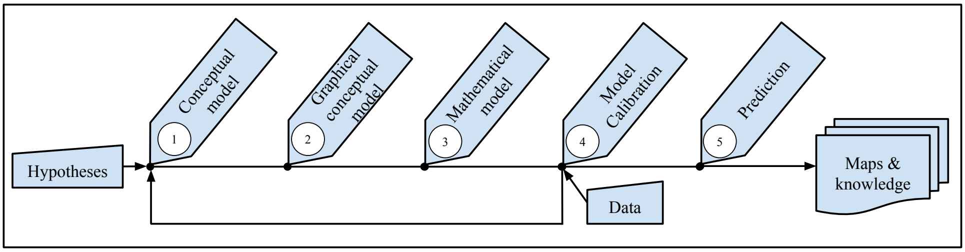


Figure 2.1: *Main steps of SEM application.*

flexible. The extra options in SEM are that i) measurement error of observed variables can be explicitly incorporated, ii) direct and indirect effects of predictors are jointly analysed, iii) a variable can be a dependent and independent variable in the system at the same time, and iv) multiple dependent variables are modelled simultaneously (Arhonditsis et al., 2006; Grace and Keeley, 2006). These are also important advantages for DSM because we know that many soil properties are cross-correlated, are influenced by environmental variables and influence each other in various ways, and cannot be measured without error.

Based on Grace et al. (2012), Fig. 2.1 shows a diagram of the principal SEM components when the main goal is to use SEM to predict target variables. SEM begins with the development of general hypotheses which are adapted to a specific conceptual model (1) applicable to a case study. The conceptual model is converted, first, into a graphical model (2), and later, into a mathematical model (3). Next, making use of a dataset that can come from either experimental or observational data, the model is calibrated and assessed (4). Poor model fitting could suggest changes to the conceptual model (1) and re-specification of graphical and mathematical models (2 and 3). When model fitting is satisfactory, (spatial) prediction of target variables (5) takes place. These five steps are explained in detail in the following sections.

2.2.2. Conceptual and graphical model

To begin with, it is important to analyse and summarise the general scientific concepts behind the objective. For example, if we are interested in predicting soil properties, we first have to analyse the main soil genetic models, such as those of Jenny (1941) and Simonson (1959), that state the theoretical relationships between soil properties. Conversion from theories to a conceptual model is a process where the researcher has to consider how these theories can be linked with observational data (Grace et al., 2012). After the conceptual model has been derived, dominant processes and causal relationships between system state variables and external drivers are described and portrayed graphically.

Graphs in SEM, also called causal diagrams, are meant to summarise causal connections among variables (Pearl, 1995), and specify the variables involved in the system. It is important to be aware that specification of variables is an exhaustive process where the nature, distribution and theoretical meaning of every variable, as well as its interrelationships with the other variables, have to be analysed in-depth. Moreover, SEM graphs include different types of variables, and in accordance with this, both external drivers and system state variables are distinguished.

To illustrate the above, Fig. 2.2 presents a theoretical SEM graph with all different types of variables, parameters and connections that a SE model may include. In practice, as we shall see in Sections 3.3 and 2.3.4, SEM can be much more complex. In the scheme shown in Fig. 2.2, some of the variables are measured (boxes), while others are not (ovals). The latter category are generally called latent variables. Latent variables are measured indirectly through observed variables. A latent variable could represent the "real" value of an observed variable. For instance, the soil organic carbon stock at some location may be the latent variable, which is measured through a carbon stock observation affected by measurement error. A latent variable could also be an abstract construct, such as "soil fertility", which can be assessed by means of several observed variables, such as measured organic carbon, nitrogen and phosphorous. Arrows between latent and observed variables mean "is measured through". These define the measurement model. Arrows that connect latent variables mean "affect" and represent the structural model. Thus, a SE model is defined by a measurement model that specifies how latent variables are measured, and a structural model that describes the interrelationships between the latent variables. The rest of the components of Fig. 2.2 will be explained in the next section.

2.2.3. Mathematical model

As was explained in the previous section and illustrated in Fig. 2.2, the graphical SE model is composed of latent and observed variables. A latent variable can be exogenous (i.e., not depending on other latent variables, $\boldsymbol{\xi}$), which we call external drivers, or endogenous (i.e., depending on other latent variables, $\boldsymbol{\eta}$), which we call state variables. Measurements of exogenous variables are denoted by $\mathbf{x}$, measurements of endogenous variables by $\mathbf{y}$.

The mathematical model that underlies the SEM graph consists of structural and measurement equations. Following the approach of Jöreskog and Sörbom (1982), all observed variables are standardized prior to modelling, by subtracting the mean

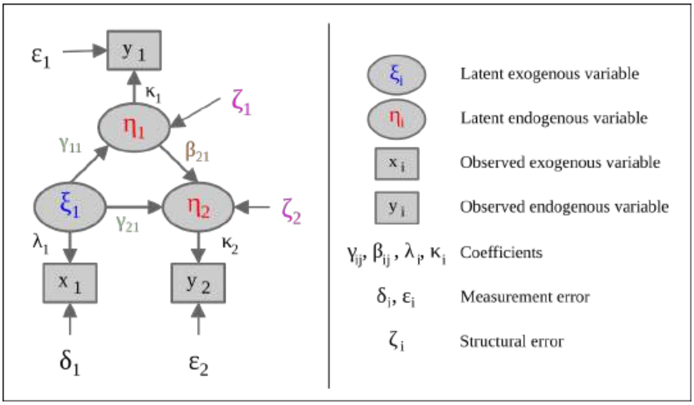


Figure 2.2: *Graphical example of a structural equation model —adapted from Jöreskog and Sörbom (1982).*

and dividing by the standard deviation. The structural model is given by:

$$\boldsymbol{\eta} = \mathbf{B}\boldsymbol{\eta} + \boldsymbol{\Gamma}\boldsymbol{\xi} + \boldsymbol{\zeta} \tag{2.1}$$

where $\boldsymbol{\eta}$ is an $m \times 1$ vector of state variables, $\boldsymbol{\xi}$ is an $n \times 1$ vector of external drivers, $\mathbf{B}$ is an $m \times m$ coefficient matrix, $\boldsymbol{\Gamma}$ is an $m \times n$ coefficient matrix, and $\boldsymbol{\zeta}$ is an $m \times 1$ vector of normally distributed residuals with zero mean (Jöreskog and Sörbom, 1982). Note that the diagonal elements of $\mathbf{B}$ will be zero, because an endogenous variable can depend on other variables but not on itself. The variance–covariance matrix of $\boldsymbol{\zeta}$ is denoted by $\boldsymbol{\Psi}$, which is not shown in Fig. 2.2.

The measurement model is defined by:

$$\mathbf{y} = \mathbf{K}\boldsymbol{\eta} + \boldsymbol{\varepsilon} \tag{2.2}$$

$$\mathbf{x} = \boldsymbol{\Lambda}\boldsymbol{\xi} + \boldsymbol{\delta} \tag{2.3}$$

where $\mathbf{y}$ is a $p \times 1$ vector of endogenous observed variables with measurement error vector $\boldsymbol{\varepsilon}$, and $\mathbf{x}$ a $q \times 1$ vector of exogenous observed variables with measurement error $\boldsymbol{\varepsilon}$ and $\boldsymbol{\delta}$ are taken as mutually independent (also from $\boldsymbol{\zeta}$), zero-mean normal deviates, with covariance matrices $\boldsymbol{\theta}_{\varepsilon}$ and $\boldsymbol{\theta}_{\delta}$, respectively. $\mathbf{K}$ is a $p \times m$ coefficient matrix and $\boldsymbol{\Lambda}$ a $q \times n$ coefficient matrix (Iacobucci, 2009).

2.2.4. Model calibration

Parameters of the SE model that need to be estimated from the available data or fixed in some other way are $\mathbf{B}$, $\mathbf{\Gamma}$, $\mathbf{K}$, $\mathbf{\Lambda}$, $\mathbf{\Theta}_\varepsilon$, $\mathbf{\Theta}_\delta$, and $\mathbf{\Psi}$. $\mathbf{\Theta}_\varepsilon$ and $\mathbf{\Theta}_\delta$ are typically taken as diagonal matrices because measurement errors of different variables are usually uncorrelated. In addition, the diagonal elements of $\mathbf{\Theta}_\varepsilon$ and $\mathbf{\Theta}_\delta$ are also often externally derived, from lab precision information or through expert judgement. For instance, the lab measurement error variance of soil organic carbon may be specified by the manufacturer of the laboratory equipment or derived from differences between duplicates. Measurement errors in field assessment of the thickness of the A horizon may be derived from comparison of independent observations of multiple experts, or by letting experts quantify the accuracy of their interpretations through expert elicitation.

Matrices $\mathbf{K}$ and $\mathbf{\Lambda}$ must be provided by the user. In this case study, these are all taken as identity matrices, because we only use direct measurements of the various soil properties and because we only have a single measurement of each soil and external variable at each measurement location. For instance, the "measured soil thickness" (measurement variable) is the "true soil thickness" (latent variable) augmented with a measurement error.

The elements of matrices $\mathbf{B}$ and $\mathbf{\Gamma}$ will be zero except when these correspond to an arrow in the graphical model. For instance, if the graphical model has no arrow pointing from latent variable η_1 to latent variable η_2, then β_{21} will be zero. Thus, the number of parameters included in $\mathbf{B}$ and $\mathbf{\Gamma}$ that need to be inferred equals the number of arrows in the graphical model. These parameters must be estimated with calibration data. Different estimation methods can be used and the mathematics behind it is somewhat involved. Below we give a brief description, while we refer to Bollen (1989); Iacobucci (2009, 2010); Grace et al. (2012) for detailed explanations.

The estimation of model parameters can be done in two different ways in SEM. Local estimation, which consists of analysis of each system state variable and its predictors separately, allows complex model specification and prevents propagation of misspecification errors Grace et al. (2012). Global estimation, which is implemented in most current software packages and used in this paper, is based on a comparison of observed variance–covariance matrices with model-implied variance–covariance matrices. It can use different estimation approaches, such as maximum likelihood and weighted least squares, to match the model-implied and observed matrices. It is important to be aware that the number of parameters to be estimated cannot be too large in case of a small calibration dataset, and indeed in some configurations the estimation algorithm might not converge due to under-determination. Lee and Song

(2004) compared two methods of parameter estimation and found that in their case a Bayesian approach needed a minimum of two to five observations per estimated parameter, while maximum likelihood required a larger sample size.

Finally, matrix $\mathbf{\Psi}$ is also estimated as part of the calibration procedure, but estimation can be simplified by fixing some of its elements to zero. This will obviously not be the case for the diagonal elements, but it may make sense to assume that the structural noise associated with one latent variable is uncorrelated with that of another. In the simplest case, all these correlations will be assumed zero and $\mathbf{\Psi}$ will be taken as a diagonal matrix. This may be a good way to start, and next it may be examined if relaxing the assumptions still yields an estimable model that in addition has an improved performance.

Several software tools have been developed for implementation of SE models. Many of these are commercial, such as LISREL (Jöreskog and Yang, 1996), Mplus (Muthén and Muthén, 1998) and AMOS (Arbuckle, 1997), while others have been developed under the GNU General Public License, such as `lavaan` (Rosseel, 2012), a package in R (R Core Team, 2017). We used `lavaan` in the case study and explain its syntax and use in Section 2.2.6.

2.2.5. Prediction

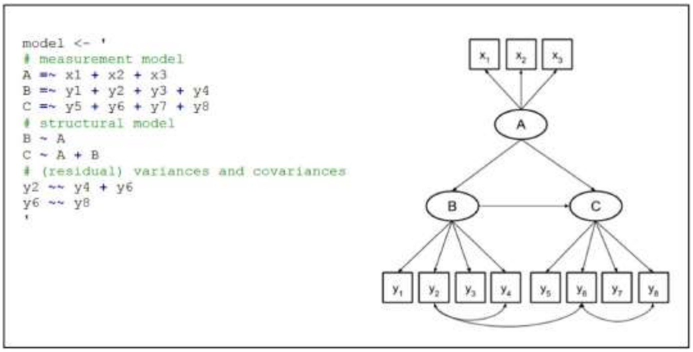
Once the SE model has been calibrated, prediction is relatively straightforward. Consider the case where at a prediction location the observed exogenous variables $\mathbf{x}$ are available (i.e., environmental factors), while we wish to predict the endogenous variables $\boldsymbol{\eta}$ (i.e., the soil properties). From Eqns. 2.1 and 2.3 we easily derive:

$$\hat{\boldsymbol{\eta}} = (\mathbf{I} - \mathbf{B})^{-1}\mathbf{\Gamma}\mathbf{\Lambda}^{-1}\mathbf{x} \tag{2.4}$$

while the variance–covariance matrix of the prediction error is given by:

$$\begin{aligned} Var(\boldsymbol{\eta} - \hat{\boldsymbol{\eta}}) = Var\left((\mathbf{I} - \mathbf{B})^{-1}\left(\mathbf{\Gamma}\boldsymbol{\xi} + \boldsymbol{\zeta} - \mathbf{\Gamma}\mathbf{\Lambda}^{-1}\mathbf{x}\right)\right) = \\ (\mathbf{I} - \mathbf{B})^{-1}\left(\mathbf{\Gamma}\mathbf{\Lambda}^{-1}\mathbf{\Theta}_{\delta}(\mathbf{\Gamma}\mathbf{\Lambda}^{-1})^{T} + \mathbf{\Psi}\right)\left((\mathbf{I} - \mathbf{B})^{-1}\right)^{T} \end{aligned} \tag{2.5}$$

Note that estimation errors of the model parameters are not taken into account. We discuss the implications of this in the Discussion Section.



```
model <- '
# measurement model
A =~ x1 + x2 + x3
B =~ y1 + y2 + y3 + y4
C =~ y5 + y6 + y7 + y8
# structural model
B ~ A
C ~ A + B
# (residual) variances and covariances
y2 ~~ y4 + y6
y6 ~~ y8
'
```

Figure 2.3: *Example of lavaan syntax.*

2.2.6. Implementation with `lavaan`

We used the `lavaan` package (Rosseel, 2012) to implement the SE model for the case study. Rosseel (2012, 2013, 2017) gives a good introduction, including model syntax, interpretation of output, and instructive examples. The basic components of a fully specified SE model in `lavaan` are: definition of the measurement model; definition of the structural model; specification of (residual) (co)variances.

An example of `lavaan` syntax that specifies these components is given in Fig. 2.3. The operator =~ means "is manifested by", ~ means "is regressed on" and ~~ means "is correlated with". After the model has been defined, next it must be calibrated with the available data. This is done with the `sem()` function. Prediction and calculation of prediction error variance is not included in `lavaan`, but it is achieved by applying Eqns. 2.4 and 2.5 in R.

2.3. Case study

2.3.1. Study area

The study area is located in the Rolling Pampas sub-region between latitudes 35° 00' S and 33° 17' S, and longitudes 58° 55' W and 61° 21' W, and encompasses about 23 000 km^2 in the north-eastern corner of the Buenos Aires Province (Fig. 2.4). The dominant soil classes are Phaeozems (Argiudolls and Argiaquolls in Soil Taxonomy), associated with Solonetz (Natracuolls and Natruacualfs in Soil Taxonomy) developed on aeolian sediments (loess and loess-like materials) that have different sources and compositions (Morrás, 1999; Zárate, 2003). These characteristics give cause to rich soil mineralogical differences that affect the spatial soil properties and

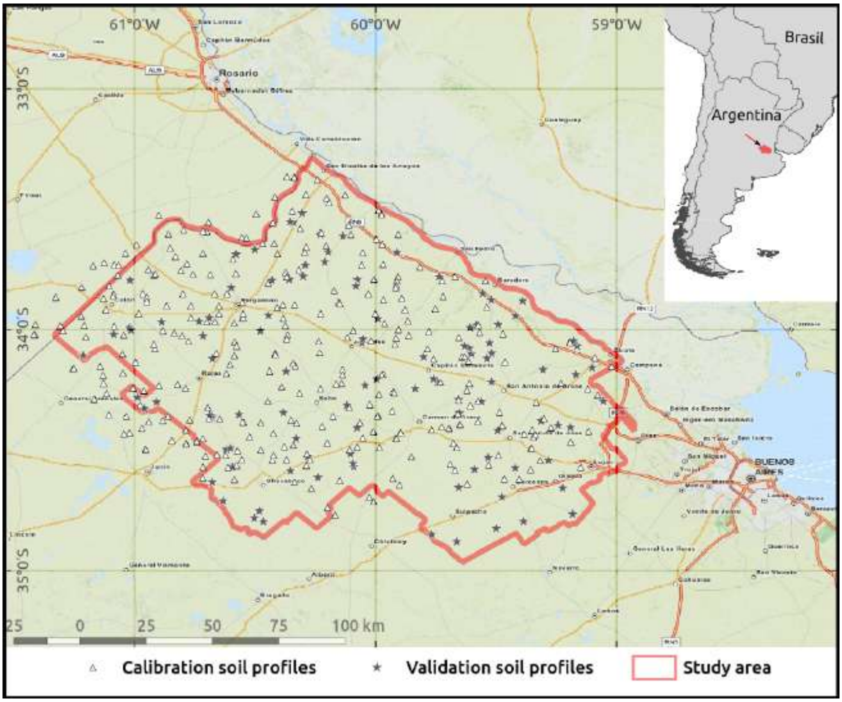


Figure 2.4: *Extent of the study area, and locations of soil profiles used for calibration and validation.*

dynamics (Cruzate, 2001; Morrás et al., 2002). The vast plains of the study area are covered by small and shallow depressions, which form the Pampean wetland (Quirós, 2005) and are endowed with rich soils. The plains were transformed from native grasslands to cropland during the last century (Viglizzo et al., 2004).

Annual precipitation ranges between 900 and 1000 mm. Summer months are characterised by rainfall deficits, while there is a rainfall excess during winter months. The average minimum temperature of the coldest month is 10℃, and the average maximum temperature of the hottest month is 23℃ (Cabrini and Calcaterra, 2008).

2.3.2. Data

Calibration data

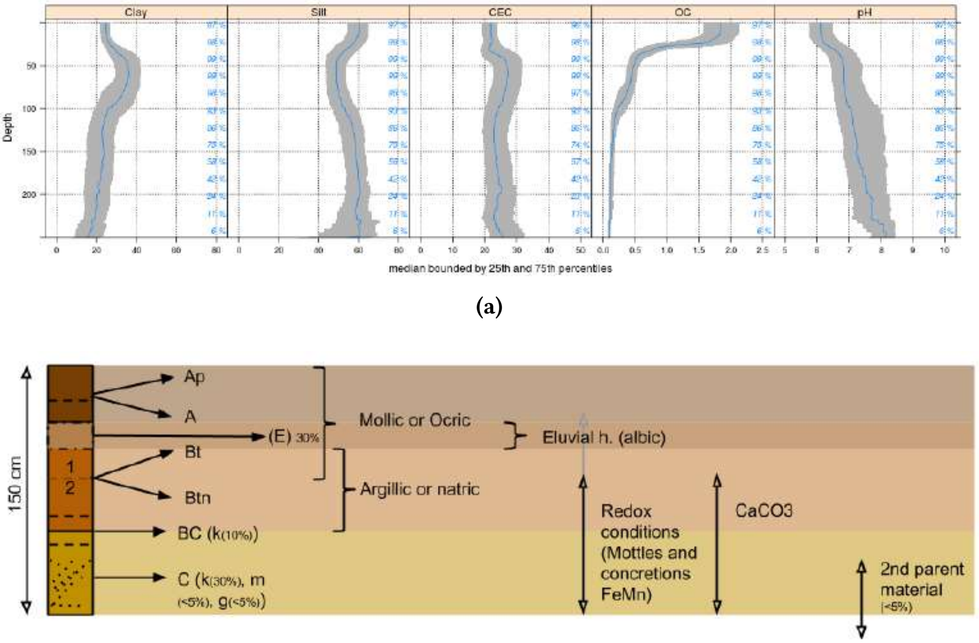
The soils of the study area were surveyed at scale 1:50 000 during the 1960s and 1970s under the Soil Map Plan that covered 500 000 km^2 of the Argentinian Pampas region (INTA, 1964). Soil sampling, description and classification were done following the Soil Survey Manual (Soil Survey Staff, 1951) and 7th Approximation (Soil Survey

***Table 2.1:** Summary statistics of calibration and validation data sets. Statistics are minimum value (Min), median, mean, maximum value (Max), standard deviation (SD), coefficient of variation (CV) and sample size (n).*

Soil properties	**Unit**	**Min**	**Median**	**Mean**	**Max**	**SD**	**CV**	**n**
	Calibration data set							
Thickness of A horizon	cm	6.0	26.0	25.1	60.0	7.47	0.30	320
Organic carbon A horizon	g/100 g soil	0.22	1.88	1.90	3.02	0.47	0.25	320
Total bases A horizon	cmol+/kg	9.5	18.8	18.7	28.6	3.2	0.17	320
Base saturation A horizon	% of CEC	67.0	84.0	85.0	100.0	6.8	0.08	320
ESP A horizon	% of Total Bases	0.4	2.3	4.3	50.2	6.5	1.49	320
ESP B horizon	% of Total Bases	0.3	2.5	8.2	63.4	12.0	1.46	320
Clay ratio B/A	%/%	0.35	1.49	1.53	2.94	0.37	0.25	320
	Validation data set							
Thickness of A horizon	cm	8.0	20.0	22.4	53.0	7.5	0.33	93
Organic carbon A horizon	g/100 g soil	1.14	1.76	1.82	2.94	0.35	0.19	92
Total bases A horizon	cmol+/kg	11.8	16.0	15.8	22.6	2.2	0.14	92
Base saturation A horizon	% of CEC	63.1	72.8	73.4	99.4	6.2	0.08	92
ESP A horizon	% of Total Bases	0.6	2.2	3.2	25.2	3.6	1.12	92
ESP B horizon	% of Total Bases	1.4	2.9	5.6	32.9	6.6	1.19	100
Clay ratio B/A	%/%	0.78	1.28	1.26	1.70	0.16	0.13	92

Staff, 1960). The current soil database of the study area contains 342 soil profiles (Fig. 2.4) with soil morphology descriptions and laboratory data (INTA, 2015). Of these, 320 could be used for modelling. Several profile descriptions had missing information for the target properties: nine for total bases, four for base saturation, nine for exchangeable sodium percentage (ESP) of the A horizon and 55 for ESP of the B horizon. Since values were missing not randomly but preferentially, SEM cannot handle missing values, and we preferred to make use of the full dataset, we replaced the missing values with average values calculated on observations in the neighbourhood.

Fig. 2.5 and Table 2.1 show the main soil features of the study area and summary statistics of the calibration data. ESP (of A and B horizons) data are skewed. This skew is caused by the presence of lowlands where large Na^+ concentrations are found that result in large ESP values. These lowlands only cover a small portion of the study area. Fig. 2.5a shows that soils in the study area typically have a clay illuviation horizon between 25 and 100 cm. The CEC more or less follows the clay profile. pH generally increases with depth resulting in alkaline conditions in the subsoil. Fig. 2.5b shows that most soils have a mollic epipedon, except in lowlands, where it is usually classified as Ochric. More than 90% of soils in this region have Argillic horizons with two or three Bt sub-horizons per soil profile. Natric subsurface horizons are present in many lowlands, with heterogeneous spatial distribution



(a)

(b)

Figure 2.5: ***a)*** *Graphs of average of clay and silt percentage, CEC in cmol+/kg soil, percentage of OC, and pH, as function of soil depth. Blue lines represent average values, blue figures are the percentage of soil profiles used for estimation, while the shadow area is the interquartile range.* ***b)*** *Schematic representation of main soil features: in the left, an indicative representation of the most common soil profile. Continuous lines are horizon boundaries, dashed lines represent possible boundaries. Percentages indicate the proportion of soil profiles with this feature. Lower case letters in brackets indicate horizon suffix from Soil Taxonomy.*

of sodium concentration and presence of Albic (E) horizons. Also, B horizons have a relatively high proportion of interstratified illite-smectite clays, which creates impediments for water percolation and swelling features. The presence of discontinuous duripans and fragipans in some C horizons (about 30% of calibration data), cemented with calcium carbonates, also increases hydromorphic conditions in the area, particularly in lowlands. Finally, some profiles show a second parent material (or buried paleo-soil), which in a few cases are present within the solum.

Validation data

Field soil data and soil samples were collected from 100 locations based on a stratified simple random sampling design (Brus et al., 2011). The stratification criteria were accessibility and soil moisture regime. Three accessibility strata were derived from a map of the distance to the nearest road. Four moisture regime strata were derived from legacy soil maps at scale 1:50 000 (INTA, 1964). Intersecting the accessibility and moisture regime strata gave twelve sampling strata. Sampling locations were allocated to the strata proportional to their surface area but weighed by distance to roads (decreasing the inclusion probability for less accessible strata by a factor two for accessibility stratum 2 and a factor four for stratum 3 compared to stratum 1 that has the highest accessibility), with a minimum allocation of two points per stratum.

At each location, a sample was taken from the A, B, and the top 30 cm of the C horizon from two pits ten meters apart. The sample material of each set of two samples was combined in a composite sample. The soil samples were analysed for texture, organic carbon, pH, electrical conductivity, exchangeable bases (Ca^{2+}, Mg^{2+}, K^{+}, and Na^{+}), and cation exchange capacity in the laboratory of INTA's Soil Institute in Buenos Aires. Table 2.1 shows summary statistics of the validation data set. Most soil properties show differences in the mean with respect to calibration data. This could be explained by differences in age, laboratory and sampling methods. For total bases of the A horizon and base saturation, and to a lesser extent the Clay B/A ratio, the differences in means are large compared to the standard deviations. This indicates that there is limited overlap between the two data distributions and indicates that there are systematic differences between the calibration and validation data.

In order to assess laboratory accuracy, 36 samples from thirteen locations that were selected randomly from the sampling pool, were taken in duplicate. All soil samples (including duplicates) were recoded in random order before shipment to the laboratory.

Covariates

Table 2.2 shows the twelve covariates that were used in the SE model. The covariates were resampled to a common grid of 250 m resolution, which is the resolution of the MOD13Q1 MODIS product.

Digital elevation model (SRTM) The SRTM (Farr and Kobrick, 2000) digital elevation model (DEM) at 30 m resolution was used to derive a set of topographic covariates. The DEM was first processed to remove striping and random phase noise through filtering and to reduce artefacts generated by riparian forest. Next, SAGA GIS (Conrad et al., 2015) was used to derive maps of the vertical distance to the channel network (vdchn) at 30 m resolution, multiresolution index of valley bottom

Table 2.2: *Environmental covariates used in the structural equation model.*

Code	Description	Source	Resolution
lstm	Mean of 14 years of Daytime 8-day Land-surface Temperature	Terra/MODIS, product MOD11A2.	1 km
lstsd	Standard deviation of 14 years of Daytime 8-day Land-surface Temperature	Terra/MODIS, product MOD11A2.	1 km
wdist	Distance to water bodies	Landsat 8 images	30 m
dem	Altitude	SRTM	30 m
vdchn	Vertical distance to channel network	SRTM	30 m
mrvbf	Index of multiresolution valley bottom	SRTM	250 m
wti	Wetness terrain index	SRTM	250 m
maxc	Maximum curvature	SRTM	250 m
slope	Slope	SRTM	250 m
river	Distance to Parana River	-	30 m
evim	Mean of 14 years of Enhanced Vegetation Index (EVI) 16-days	Terra/MODIS, product MOD13Q1	250 m
evisd	Standard deviation of 14 years of Enhanced Vegetation Index (EVI) 16-days	Terra/MODIS, product MOD13Q1	250 m

flatness (mrvbf) at 250 m resolution, wetness terrain index (wti) at 250 m resolution, maximum curvature (maxc) at 250 m resolution and slope (slope) at 250 m resolution from the processed DEM.

Landsat 8 images Distance to water bodies can be considered a proxy of groundwater depth (Jenny, 1941). Five mosaiced Landsat 8 images were selected from 8, 15 and 17 November 2014. Bands from Operational Land Imager sensor (OLI): 4; 5; 6; 7 and Thermal Infrared sensor (TIRS): 11 were chosen to identify water bodies. In order to create a map of distance to water bodies, an empirical approach that included unsupervised classification (Conrad, 2001) and visual interpretation was implemented. First, the images were classified into 20 spectral classes, and next each class was assigned to water bodies, or not. To judge whether a class belongs to the water body class, visual interpretation was used. Next, a raster of the distance to the nearest water body pixel was created. All analyses were executed in SAGA-GIS (Conrad et al., 2015).

MODIS EVI 16-day composition (MOD13Q1) Enhanced Vegetation Index (EVI) at 16-days temporal and 250 m spatial resolution from the MODIS/Terra[1] remote sensor was used to characterise the land cover of the study area. A total of 342 images taken between March 2000 and December 2014 were processed using R. Standard deviation (evisd) and mean (evim) were computed per pixel for the whole period.

Terra land-surface temperature & emissivity (MOD11A2) Daytime 8-day 1 km grid Land-surface Temperature (LST) from the MODIS/Terra satellite was analysed as proxy of soil climate. Mean (lstm) and standard deviation (lstsd) were calculated from 689 images for the same period used to derive the EVI images.

Proxy of soil parent material Soil parent material of the study area has been discussed extensively (Scoppa, 1975; Gonzalez Bonorino, 1966; Zárate, 2003; Morrás and Moretti, 2016). The main sources of the aeolian sediments that cover the study area are The Cordillera de los Andes, Sierras Pampeanas and Paraná river. As a result, the major spatial differences in parent material are in the south-west–north-east direction. For this reason, and given that there was no parent material map available, we decided to use distance to the Paraná river as a proxy of the parent material distribution.

2.3.3. Conceptual model

In SEM, general hypotheses are converted into a conceptual model (Fig. 2.1), which identifies the main endogenous variables involved, and how these are inter-related and influenced by exogenous variables. The development of hypotheses of soil formation has a rich history (Dokuchaev, 1883; Jenny, 1941; Simonson, 1959; Runge, 1973; Phillips, 1993). To establish a conceptual model for the case study, an approach was adopted from Simonson (1959), that describes the main soil-forming processes, and Runge (1973), which presents an energy model for grasslands over unconsolidated parent materials. The model is a merge of the models of Jenny (1941) and Simonson (1959, 1978). First, an inventory was made of the main soil-forming processes in the area.

[1] MOD13Q1 and MOD11A2 were retrieved from the online Reberv/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. See more at: https://lpdaac.usgs.gov/citing_our_data#sthash.yGKPuOqi.dpuf

Table 2.3 summarises all processes and lists the key soil properties that could be used as process diagnostics.

Soil-forming processes

Table 2.3: *Main soil forming processes of the study area, their effects, involved horizons, key soil properties and external driving factors.*

Soil forming process	**Effect on soil formation**	**Soil horizons**	**Key soil properties**	**Driving factors**
Base cation cycle	Accumulation, recycling and depletion of base saturation and total bases	A	Base saturation, total bases	Water dynamics, parent material, climate, land cover
Argilluviation	Development of argillic horizon	B, A	Ratio of clay in B and A horizon	Water dynamics, climate
Pedoturbation	Mixing of soil material	A, B, C	Ratio of clay in B and A horizon	Water dynamics, parent material, land cover
Melanization	Colour darken by humic acid increase, development of mollic horizon	A, AB, BA	Soil organic matter	Climate, land cover, water dynamics
Soloni-/Solodisation	Increase of exchange sodium percentage (ESP)	B, A, E	Exchange sodium percentage (ESP)	Parent material, climate, water dynamics
Calcification	Loss, depletion and accumulation of calcium carbonate, development of calcic horizons	All horizons	Presence of calcium carbonate	Water dynamics, climate
Hydromorphism	Iron depletion, greenish colours, development of mottles and concretions	development of E horizon. E, B	Presence of E horizon, presence of mottles and concretions	Water dynamics
Water erosion	Change of epipedon thickness	A	A horizon thickness	Relief, water dynamics, land cover, parent material

Base cation cycle Base cation dynamics results in leaching or accumulation of base cations, depending on the equilibrium between these two processes. Base cation leaching involves eluviation of Ca^{2+}, Mg^{2+}, K^{+} and Na^{+}, usually referred to as the exchange complex bases. Although this process is much more prominent in other ecological regions with soils such as Acrisols (Bockheim and Gennadiyev, 2000), it also occurs in Phaeozems and Solonetz, simultaneously with argilluviation. Base cation leaching lowers the amount of bases per unit of clay and organic matter. At the same time it lowers the pH because leached bases are replaced by H^{+} and other acidifying cations, such as Al^{3+} and Fe^{3+}. Thus, acidification is more prominent in the epipedon than in the subsoil. A consequence of acidification is a decrease of cation exchange capacity (CEC), which is stable in most common clay types such as smectite and illite, but pH-dependent on organic matter and clays such as kaolinite and allophane. These physico-chemical processes are counteracted by accumulation of cations through biological activity and clay weathering. Plant roots and fauna can recycle nutrients upward and increase the concentration of the most important cations in the topsoil. This biocycling process is particularly intense in grasslands under temperate climates.

Melanization Organic residues are accumulated at the soil surface and transformed in humic substances through biochemical and abiotic processes. When humification is coupled with enrichment of base cations it creates a mollic horizon, which is characteristic for soils such as Phaeozems, Chernozems and Kastanozems (Bockheim and Gennadiyev, 2000). Melanization is the term for darkening of soil mineral particles by humus coatings and is a distinguishing process under grassland vegetation (Buol et al., 2011). Soil organic matter accumulation increases cation exchange and buffer capacity as does clay, although the CEC of organic matter is also controlled by pH. This is one of the most important processes in the study area and its magnitude varies spatially according to water dynamics, type of vegetation and soil temperature.

Argilluviation Argilluviation or lessivage (Duchaufour, 1998), is the mechanical migration of clay from surface to deeper soil horizons. This process is governed by the direction of water flow, which is predominantly vertical and downwards. The magnitude of argilluviation, usually assessed through the difference in clay content between the epipedon and the subsurface horizon, is related to bioclimatic conditions, basically to pluviometry, and is mainly conditioned by soil mineralogical and granulometric composition, and aggregate stability.

Argilluviation is one of the most important soil-forming processes in the study area, which has been affected at different degrees according to a combination of parent material granulometry and pluviometric spatial variations. In the Pampas region, soil textures are finer and pluviometry is higher from south-west to north-east; it is thus expected that the clay percentage ratio between horizon B and A increases in the same direction.

Solonization and solodization Solonization, or alkalinization, refers to the process of sodium accumulation that occurs in saline-sodic soils when soluble salts are leached. It promotes clay dispersion and rises soil pH above 8.5. Solodization is a result of elluviation of dispersed clays and the replacement of exchangeable sodium by other cations, mainly H^+, thus giving rise to a bleached and acid horizon above a natric one (Brady and Weil, 2014).

These processes result in soils with high exchangeable sodium percentage (ESP), a condition that is present in most lowlands of the study area (Quirós, 2005) where the sources of sodic salts are groundwater and water from saline streams. Groundwater can be found on the surface to a few meters depth in low landscape positions (Laurencena et al., 2002; Taboada et al., 2009). The poor drainage of these soils affects negatively vegetation growth and organic matter accumulation, which is lower than in upland soils.

Calcification In humid regions, calcification refers to the redistribution of calcium carbonates (Schaetzl et al., 1996). Calcium carbonate is generally inherited from parent material, and its distribution and concentration depend on climate, carbonate solubility, mineralization by microbes and vegetation, water flow direction, and soil texture. Accumulation of carbonates produces different morphogenetic features, from carbonate coatings to petrocalcic horizons (Schoeneberger et al., 2012). Petrocalcic horizons may be a restriction for root growth and water infiltration, so that a groundwater table could be established above this layer. Calcium carbonate increases soil pH, but not above 8.2 because it is not soluble above this value.

Within the study area, leaching of carbonates is a main process and in upland soils, secondary carbonates are present under different morphologies in the BC or C horizons, generally below 1.5 m from the surface. However, in lower landscape positions the groundwater level restrains the free vertical movement of water and concentrates carbonate salts at different depths (Durán et al., 2011; Imbellone et al., 2014).

Hydromorphism Hydromorphism or gleization is a process that takes place in anaerobic (or aquic) conditions. Reduction and oxidation of Fe and Mn create features such as mottles, concretions and a gleying colour pattern (Bockheim and Gennadiyev, 2000; Schoeneberger et al., 2012). These features are generally used to infer soil moisture regimes and involve removal, translocations and precipitation of Fe and Mn. In the study area, the upland soils may present few redoximorphic features in their argillic horizons, particularly those with higher contents of expandable clay. In lowlands different degrees of hydromorphism are observed; some soils are characterised by the development of albic horizons, small and weak mottles, and FeMn concretions at the contact with the Bt (INTA, 1964; Imbellone et al., 2010); other soils suffering longer waterlogging conditions show organic matter accumulation and gleying colour patterns close to the soil surface.

Pedoturbation Pedoturbation refers to different processes of soil homogenization by mixing. Depending on the vector that produces such effects, pedoturbation is referred to as bioturbation when the vector is the biota, argilliturbation when shrink-swell clay is the vector, or cryoturbation when ice is the vector (Blake et al., 2008). Bioturbation and argilliturbation are the main mixing processes in the study area.

Bioturbation is reported as one of the most influential processes of pedogenesis, although many of its aspects are still unknown (Wilkinson et al., 2009). These authors indicated that the process includes formation of surface mounds and soil burial, which are usually underestimated in field survey. Nevertheless, together with a particular process of humification, an intense mixing of soil components by flora and fauna is involved in the development of Pampean mollic epipedons. Argilliturbation is a process caused by type 2:1 clays with a high coefficient of linear extensibility, such as smectite. This group of clays changes its volume according to its water content. Depending on the proportion of these clays in the soil, it may develop Vertisols, soils that have deep and thick cracks in dry seasons, cuneiform peds and gilgai microrelief, among other characteristics (Blake et al., 2008; Buol et al., 2011). The soils of the study area do not belong to this soil order, although they show diverse proportions of smectite that confers vertic features to these (Imbellone et al., 2010; Durán et al., 2011; Morrás and Moretti, 2016).

Although argilliturbation and bioturbation work together, their intensities may be different. The former affects mainly the epipedon, while the latter is stronger in the argillic horizon. They both affect soil particle distribution and base cation cycle, but because they are mostly mixing processes, there is no single soil property to assess the extent of these processes.

Water erosion Water erosion may be considered a destructive soil-forming process. The process takes place in mainly three steps: soil particles detachment, transportation and deposition. Land cover, rainfall characteristics and soil particle cohesion regulate the influence of raindrops in the first sub-process. The second sub-process is governed by running water, which is dependent on infiltration capacity and terrain slope. Run-off water passes from sheet flow (with little power to detach soil) to channelised flow (with high power to carry and detach soil particles). Finally, deposition takes place down-slope. The process mainly affects the topsoil thickness and consequently, soil physical properties and organic carbon distribution.

Water erosion is an important process in the study area, but presently it is mainly related to land degradation caused by land use, rather than resulting from a natural process. The relief of this area is from flat to gently undulating and rainstorms can be very intense. Gully erosion takes place on gentle slopes when natural vegetation is removed, and sedimentation occurs down-slope. Thus, thickness of the A horizon decreases on steeper slope positions and increases in the lowlands.

Interrelationship between soil-forming processes soil-forming processes interact with one another and some can only occur after another has taken place. For example, calcium carbonate has to be leached (calcification) from the solum before argilluviation can take place. This type of interaction does not mean that an increase in calcification intensifies argilluviation, but signifies that a process takes place when the other is ending. Thus, this relationship implies time as a factor. Another type of interaction occurs when clays are dispersed by sodium, which creates conditions for argilluviation. Under this condition, a linear relationship between solonization and argilluviation would be expected. A third type of interaction appears when a process simultaneously works against another process, thus affecting the same soil property. This happens, for instance, between argilluviation and pedoturbation. The conceptual model aims to take all these relationships into account.

External controls

Fig. 2.6 shows how soil-forming processes are affected by driving soil-forming factors. Here, soil-forming factors described by Runge (1973) were considered and modified based on the particular conditions of the study area. Runge (1973) states that soil genesis is a function of the organic matter production (o), the available water for leaching (w), and time (t). The w factor combines relief and climate factors into the “available water for leaching” factor. Runge (1973) observed that w is able to organise the profile, decrease the entropy and develop horizons by using

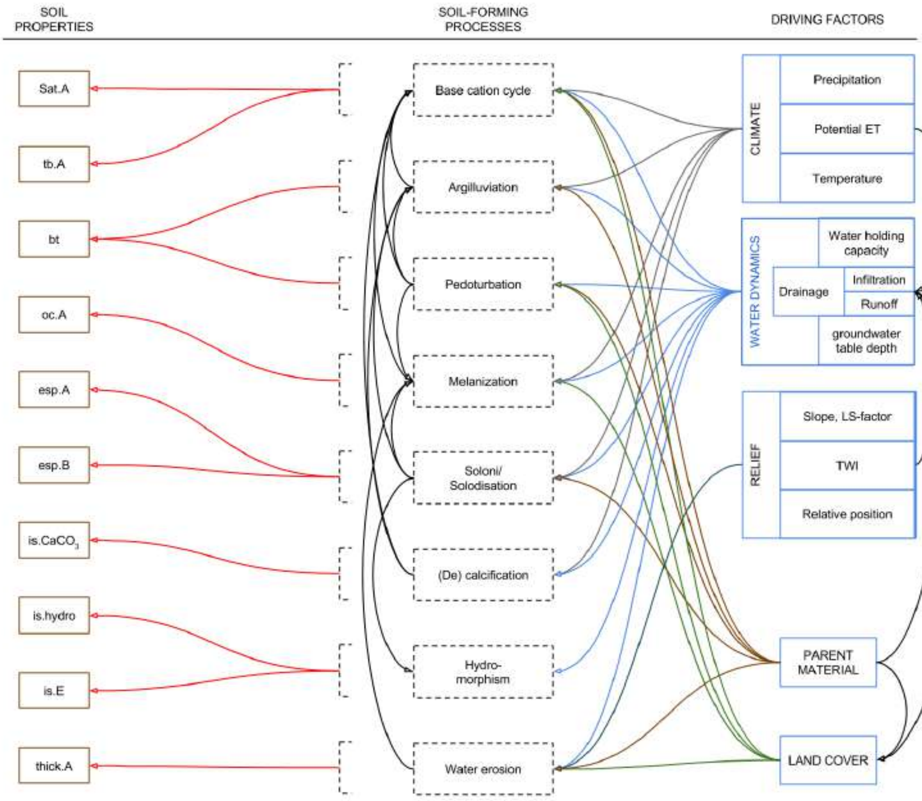


Figure 2.6: *Conceptual model depicting relationships between and across soil-forming and driving processes. Sat.A: base saturation of A horizon; tb.A: total bases of A horizon; bt: Clay ratio B/A horizons; oc.A: soil organic carbon of A horizon; esp.A and esp.B: exchangeable sodium percentage of A and B horizons, respectively; is.CaCO$_3$: presence of calcium carbonate; is.hydro: presence of hydromorphic conditions; is.E: presence of E horizon; thick.A: thickness of A horizon.*

gravitational energy, mainly in soil developed under grassland and unconsolidated materials Schaetzl and Anderson (2005). In this paper, this driving factor is called “water dynamics” (Fig. 2.6). Time, t in Runge’s model, was omitted as a driving factor for this paper, as the soil-developing time within the study area is relatively homogeneous and cannot explain spatial variation within the study area. A review of Zárate (2003) reports that the top 3–5 m loessial sediments are Late Pleistocene or Holocene in age. Finally, Runge’s model considers organic matter production as a third factor, which is mainly affected by available phosphorous (depending on parent material). However, this seems to be a simplistic assumption and thus, in this case we preferred to leave parent material and land cover (as organisms) as considered in Jenny’s model (Jenny, 1941).

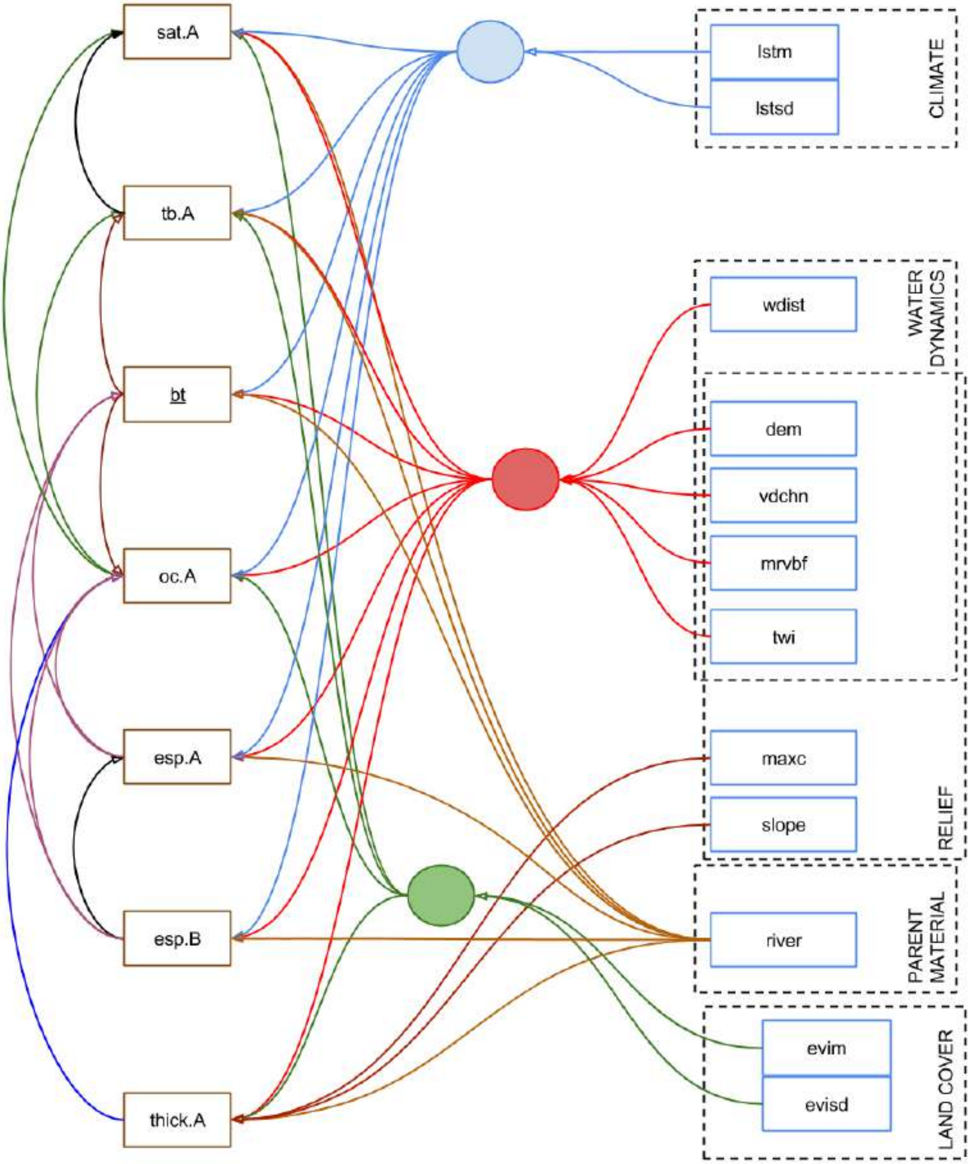


Figure 2.7: *Adaptation of the conceptual model shown in Fig. 2.6 to make it suitable for specification in lavaan. Circles are used to reduce the number of arrows between driving factors and soil properties: every arrow entering a circle connects to every soil property reached by the arrows leaving the circle.*

Graphical conceptual model

Fig. 2.6 shows the conceptual model that summarises the relationships between soil-forming processes, driving factors and affected soil properties. Black arrows that link soil-forming processes (dashed boxes) represent interrelationships between processes. These processes are controlled by driving factors, which also have interactions with each other. In other words, driving factors and soil-forming processes acted together during millennia to develop the current soils in the study area. In order to assess the intensity of the soil-forming processes within the area, ten key soil properties are linked to their respective process(es) through red arrows. These properties do not encompass all affected soil properties, but only the most relevant ones.

2.3.4. `lavaan` implementation

Before we could implement the conceptual model (Fig. 2.6) in `lavaan`, we had to simplify it. First, we removed the categorical soil properties (Presence of calcium carbonates, E horizon and hydromorphic conditions) because SE models that contain categorical latent variables do not meet the assumptions of the structural equation Eq. 2.1 and are more difficult to calibrate (Bollen, 1989). Second, we represented the driving factors with twelve environmental covariates. Interactions between the environmental covariates were not taken into account. Third, soil-forming processes were removed from the graphical model. The covariates were linked directly with the latent variables on basis of the relationships between the soil-forming processes and latent variables. The interrelations between soil-forming processes in Fig. 2.6 were represented through interrelationships between the latent variables. Fig. 2.7 shows the adapted, simplified graphical model for which a model in `lavaan` was derived. Note that ESP A horizon and ESP B horizon were log-transformed to satisfy the normal-distribution assumption. Next, all soil variables were standardized to mean zero and standard deviation one.

The `lavaan` implementation of the graphical conceptual model is presented in the Appendix Section, which entails the specification of the measurement and structural models. WEPAL reports (WEPAL, 2015) and expert knowledge were used to define the laboratory measurement error, because this information was not provided by the INTA laboratory. Note that we could also have used the differences between replicates of the validation dataset to characterise the laboratory error, because both the calibration and validation samples were analysed by the same laboratory, but we decided otherwise because we preferred to keep the validation dataset completely independent from the modelling and because the calibration samples were analysed several decades before the validation samples.

The outcomes of the fitted model are a summary report with matrices of estimated parameters, which is also included as Appendix Section, and a graphical representation of the estimated coefficients (Fig. 2.8). High positive estimates are represented by dark green colours, while low negative estimates are shown in red. Both tone and thickness of the arrow symbolise the magnitude of the coefficient. Dash green arrows represent the measurement model parameters, which were fixed to one (Epskamp, 2015).

2.3.5. Prediction results

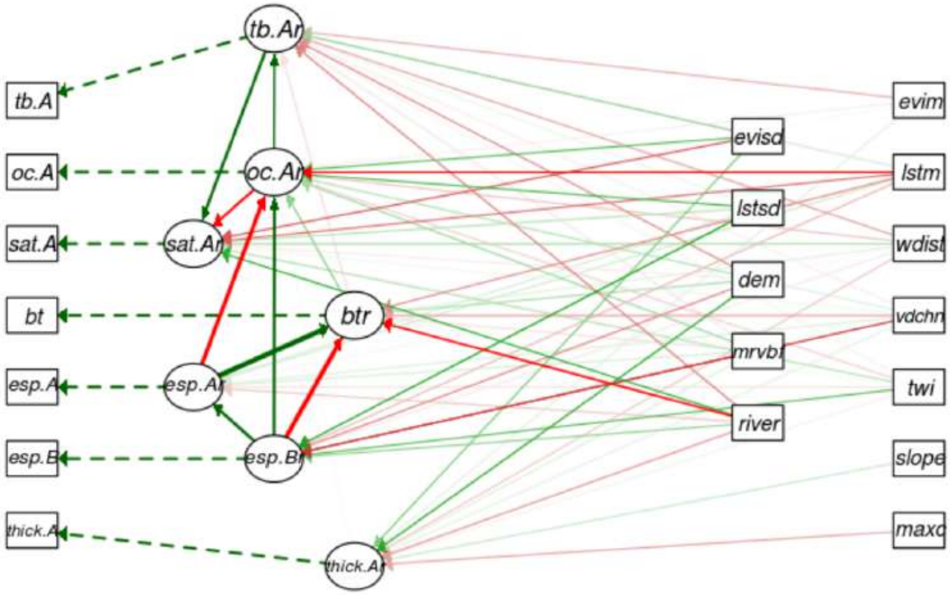


Figure 2.8: *Graphical representation of the calibrated SE model: ovals represent state variables (soil properties), which are connected to measured soil properties by dashed arrows. Boxes represent external drivers, which are connected to state variables by solid arrows. Red arrows mean negative correlation, green arrows positive correlation. The intensity of the colour represent the magnitude of the estimated parameter.*

The fitted model was used to predict the target soil properties using Eq. 2.4. The resulting soil property maps are shown in Fig. 2.9. The drainage pattern is a common factor in all maps, except for clay ratio of B/A horizons and total bases, where this pattern is weaker. It is interesting to analyse the prediction maps in combination with Fig. 2.8. For example, if one analyses the clay ratio B/A state variable in Fig. 2.8,

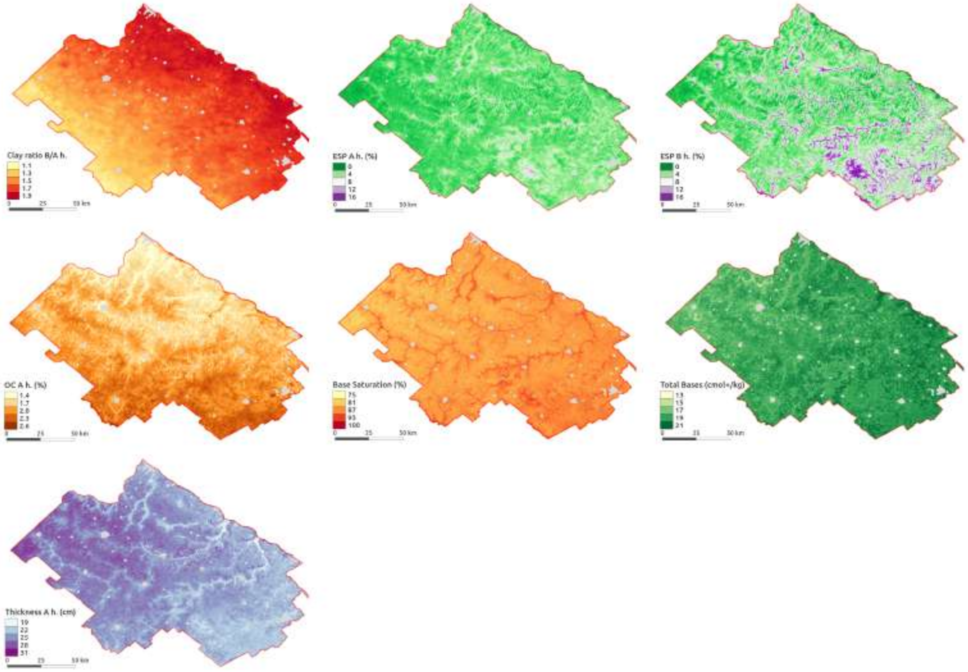


Figure 2.9: *Predicted soil property maps.*

one sees that the main external driver for this variable is the parent material proxy (distance to Paraná river). An increase of the distance to river leads to a decrease of the clay ratio, which is as expected. Also, the variation of land surface temperature (lstsd) and the vertical distance to channel network (vdchn) show a weak direct effect on clay ratio B/A. Moreover, exchangeable sodium percentage of A and B horizons also have a high impact on clay ratio B/A horizons, and thus there is an indirect effect from their predictors (lstsd and vdchan) on clay ratio, as well. Fig. 2.8 shows that ESP A horizon is mainly predicted from ESP B horizon, hence their spatial patterns are very similar. Similar interpretations can be done with the other predicted soil properties.

The estimated model parameters were also used to calculate the prediction error variance with Eq. 2.5. In this case, where the external drivers were taken as deterministic and known, the equation reduces to:

$$Var(\boldsymbol{\eta} - \hat{\boldsymbol{\eta}}) = (\mathbf{I} - \mathbf{B})^{-1} \boldsymbol{\Psi} \left((\mathbf{I} - \mathbf{B})^{-1} \right)^{T} \tag{2.6}$$

Note that the prediction error variance does not depend on location and hence is constant in space.

Table 2.4 shows the model R^2 and spatial mean for each soil property, as well as the 90% prediction interval width (PI width) estimated from the prediction error variance.

Table 2.4: *Model R^2, mean of the mapped values, and 90% prediction interval widths*

Soil Property	R^2	Spatial mean	90% PI width
Thickness of A horizon	0.09	24	20.2
Organic carbon A horizon	0.45	2.0	1.1
Total bases A horizon	0.34	19.0	8.7
Base saturation A horizon	0.70	86	21.0
ESP A horizon	0.92	3.4	6.8
ESP B horizon	0.29	6.7	11.1
Clay ratio B/A	0.67	1.5	0.84

2.3.6. Cross-validation and validation

In order to measure the quality of the laboratory we estimated the standard deviation of the measurement error (SD_{err}) by taking the square root of half the variance of the differences between duplicates. We also computed the ratio of the measurement error variance and sample variance (Var_{err}/Var_{tot}) to determine what proportion of the total variance in the validation dataset is accounted by measurement error variance. A third laboratory quality measure was the ratio of the measurement error variance and validation mean squared error (Var_{err}/MSE). This ratio indicates which part of the residual variation is explained by random measurement error. If this measure is close to one then it means that the model cannot perform better because all remaining variation is explained by measurement error in the data.

Table 2.5 shows the laboratory quality measures. The Var_{err}/Var_{tot} ratio shows that for total bases, 40% of the total variance in the validation data can be attributed to measurement error. For base saturation this is 19%, for ESP 17%. Comparison of the measurement error variance with the MSE, shows that for total bases the contribution of Var_{err} is 44%, while for ESP this is 25% for the A horizon and 22% for the B horizon.

The accuracy of the soil property maps was assessed using two approaches: leave-one-out cross-validation using the calibration dataset and independent validation using the validation dataset. In both cases, the quality measures were the Mean Error (ME), which measures prediction bias, the Root Mean Square Error (RMSE), which measures prediction accuracy, and amount of variance explained (AVE) (Samuel-Rosa et al., 2015). For the validation dataset, the ME and RMSE were estimated from weighted averaging of these measures per stratum, with weights proportional to the stratum surface areas (de Gruijter et al., 2006; Kempen et al., 2012). We estimated the upper and lower boundary of the 95% confidence interval of the ME to assess if the ME significantly differs from zero, which would indicate bias in the predictions.

Table 2.6 shows accuracy measures for predicted soil properties. Cross-validation shows that predictions are unbiased. The RMSEs and AVEs show low accuracy for all predicted soil properties. Thickness, total bases and base saturation of A horizon have the lowest AVEs (0–0.05), while the other properties only show small improvement (AVE 0.10–0.25) with respect to using the mean of the data as predictor. Validation with independent data shows biased predictions for total bases and base saturation of the A horizon. Suspected systematic differences between the distributions of the calibration and validation data sets can explain this bias (Table 2.1). Compared to cross-validation, RMSEs are generally higher, except for ESP of A and B horizons. Moreover, most soil properties present negative AVEs, which imply that the means of the validation data are better predictors than SE maps. Scatter plots of measured versus predicted values (Fig. 2.10) confirm these observations and show that there only is a small positive relation between measured and predicted values of organic carbon of the A horizon.

To verify if the SEM prediction error variance gives a proper measure of the map accuracy we computed the mean ($\bar{\theta}$) and median ($\tilde{\theta}$) standardized squared prediction error (θ) at each cross-validation location as proposed by Lark (2000). The θ is computed by taking the ratio of the squared cross-validation prediction error and prediction error variance at each cross-validation location. If the prediction errors are normally distributed and the prediction error variance is a correct assessment of the expected squared prediction error, this quantity has a Chi-square distribution with one degree of freedom. Hence, the $\bar{\theta}$ should be close to one. Lark (2000) proposed to compute also the $\tilde{\theta}$, as the median is less sensitive to outliers than the mean. Ideally the $\tilde{\theta}$ should be close to 0.455. If the median and/or mean are close to their ideal values, on average the prediction error variance is an unbiased quantifi-

Table 2.5: *Quality measures of lab measurement error.*

Soil Property	$\mathrm{SD_{err}}$	$\mathrm{Var_{err}/Var_{tot}}$	$\mathrm{Var_{err}/MSE}$
Organic carbon (g/100 g)	0.1	0.01	0.05
Total bases (cmol+/kg)	2.5	0.40	0.44
Base saturation (%)	5.1	0.19	0.13
ESP A horizon (Log %)	0.1	0.17	0.25
ESP B horizon (Log %)	0.1	0.17	0.22
Clay ratio B/A	0.1	0.69	0.06

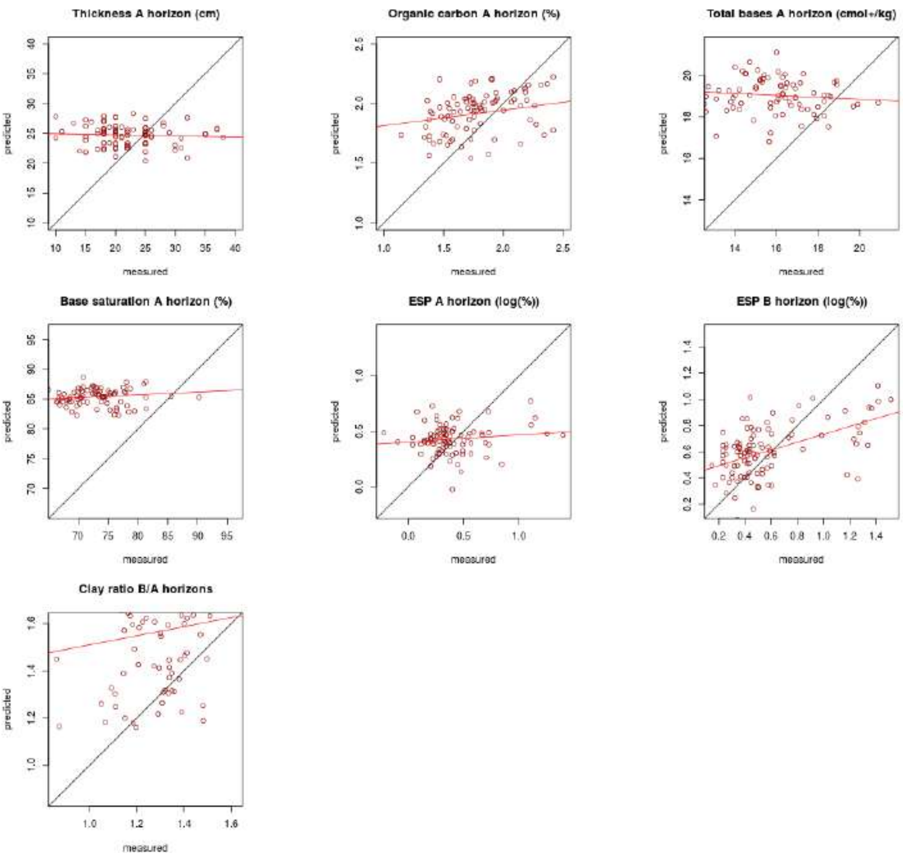


Figure 2.10: *Scatter plots of measured values of validation dataset (x-axis) vs. predicted values (y-axis), showing one-to-one (black) line and linear regression (red) line.*

cation of the accuracy of the predictions. Note that for calculating the θ statistics of the SEM results we added the measurement error variance $\boldsymbol{\theta}_{\boldsymbol{\varepsilon}}$ to Eq. 2.5 because in validation and cross-validation we compare model predictions with measurements.

The cross-validation results of $\bar{\theta}$ and $\tilde{\theta}$ show that the means vary between 1.06 and 1.11 and the medians between 0.243 and 0.481, indicating that in general the SEM model gives a fair assessment of the prediction error variance. A possible reason for the means being slightly larger than 1 is that we ignored the uncertainty in the SEM coefficients when we computed the prediction error variance. The validation means and medians show poor results for all soil properties. For base saturation, total bases and Clay ratio the means deviate substantially from 1 and the medians from 0.455, that is likely the result of the systematic difference between calibration and validation data distributions (Table 2.1). For organic carbon and ESP the variance of the calibration data is (much) larger than the variance of the validation data, which we suspect contributes to $\bar{\theta}$ and $\tilde{\theta}$ values that are much smaller than their expected values.

2.4. Discussion

2.4.1. SEM predictions

Fig. 2.8 represents the soil system interrelationships mostly as was expected based on the conceptual model. The sign of some of the coefficients, however, were opposite with what was expected. For example, soil process knowledge suggests that both ESP A horizon and ESP B horizon have a negative effect on organic carbon A horizon, but the SE model shows a positive effect of ESP B horizon. The same is true for organic carbon A horizon, total bases A horizon and base saturation A horizon, where an increase in organic carbon or total bases was expected to increase the base saturation in the same horizon, while the SE model shows a negative effect. We did not rigorously test the validity of this model because it was beyond the scope of this work. These discrepancies may be the result of misspecification, which is not uncommon in SEM (Grace et al., 2012). There are options, not used here, to constrain the model and force it to respect the expected sign.

For most soil properties, the model R^2 values (Table 2.4) are high compared to the amount of variance explained by cross-validation. In addition to the relationship between soil properties and covariates, SEM makes use of soil property interrelationships for modelling. When interrelationships are strong, then this can boost the predictive power of the model. For instance, the base saturation of the A horizon depends on total bases and organic carbon (Fig. 2.7) that help explain a proportion of the total variation of base saturation. When we apply SEM for prediction, however, soil properties can only be predicted from external drivers (Eq. 2.4) because other soil property values are not available at the prediction locations. This can result in a dramatic drop of the prediction power if the interrelationships between the target property and other properties are strong.

Soil property maps (Fig. 2.9) generally conform the conceptual model and are not biased with respect to calibration data, but their accuracy (AVE and RMSE) is poor according to validation and poor to reasonable based on cross-validation. We can think of several reasons that might explain the poor accuracy:

1. The validation and calibration datasets have more than 30 years time difference. The impact of land use and land use change, sampling and laboratory

Table 2.6: *SEM assessment. ME with the lower (l.l.) and upper (u.l.) limits of the 95% confidence interval, RMSE and mean ($\bar{\theta}$) and median ($\tilde{\theta}$) standardized squared prediction error, for validation and cross-validation procedure.*

Soil property	**l.l. ME**	**ME**	**u.l. ME**	**RMSE**	$\bar{\theta}$	$\tilde{\theta}$	**AVE**
	Cross-validation						
Thickness of A horizon		0.00		7.47	1.09	0.362	0.00
Organic carbon A horizon		0.00		0.42	1.11	0.403	0.18
Total bases A horizon		0.00		3.11	1.08	0.406	0.05
Base saturation A horizon		0.00		6.67	1.08	0.457	0.02
Log. ESP A horizon		0.00		0.38	1.07	0.446	0.17
Log. ESP B horizon		0.00		0.49	1.06	0.481	0.22
Clay ratio B/A		0.00		0.33	1.08	0.243	0.24
	Validation						
Thickness of A horizon	-10.19	-2.98	4.24	7.58	1.28	0.547	-0.21
Organic carbon A horizon	-0.34	0.01	0.35	0.44	0.79	0.378	-0.05
Total bases A horizon	-5.06	-2.86	-0.66	3.76	1.76	1.088	-2.50
Base saturation A horizon	-17.66	-12.60	-7.55	14.07	4.36	4.421	-3.72
Log. ESP A horizon	-0.27	-0.04	0.18	0.26	0.56	0.232	-0.14
Log. ESP B horizon	-0.25	-0.02	0.21	0.28	0.38	0.170	0.23
Clay ratio B/A	-0.57	-0.31	-0.05	0.41	1.49	0.881	-4.64

ME with the lower (l.l.) and upper (u.l.) limits of the 95% confidence interval, RMSE and mean ($\bar{\theta}$) and median ($\tilde{\theta}$) standardized squared prediction error, for validation and cross-validation procedure.

methods may have caused systematic and random changes in some soil properties so that the calibration data are not representative of current soil conditions. Table 2.1 shows that there are systematic differences between the distributions of the calibration and validation datasets for some soil properties. Presumably, this is the main cause of bias and large RMSE in the validation (Table 2.5).

2. The field and laboratory measurement error plays an important role. The $\mathrm{Var_{err}}/\mathrm{Var_{tot}}$ ratio (Table 2.5) shows that the measurement error is very large compared to the spatial variation in the study area. If we assume that the calibration data have similar measurement error as the validation data, the $\mathrm{Var_{err}}/\mathrm{MSE}$ ratio shows that it will be difficult to improve predictions using these data, because almost up to half of the prediction error is caused by measurement error.

3. The relationships between driving factors, which are merely proxies of the true soil-forming factors and might also have a high noise component, and state variables are not strong enough to produce accurate maps. One reason

Table 2.7: *Multiple linear regression assessment.*

Soil property	l.l. ME	ME	u.l. ME	RMSE	$\bar{\theta}$	$\tilde{\theta}$	AVE
Cross-validation							
Thickness of A horizon		0.01		7.36	1.03	0.348	0.03
Organic carbon A horizon		0.00		0.41	1.03	0.401	0.22
Total bases A horizon		-0.01		3.09	1.05	0.385	0.06
Base saturation A horizon		0.00		6.57	1.03	0.451	0.05
Log. ESP A horizon		0.00		0.37	1.03	0.421	0.24
Log. ESP B horizon		0.00		0.46	1.03	0.481	0.32
Clay ratio B/A		0.00		0.32	1.05	0.244	0.26
Validation							
Thickness of A horizon	-10.21	-2.96	4.28	7.60	1.24	0.528	-0.21
Organic carbon A horizon	-0.37	0.01	0.38	0.47	0.84	0.451	-0.16
Total bases A horizon	-5.13	-2.85	-0.57	3.82	1.76	1.135	-2.62
Base saturation A horizon	-17.68	-12.63	-7.57	14.11	4.22	4.044	-3.72
Log. ESP A horizon	-0.29	-0.05	0.19	0.28	0.58	0.242	-0.21
Log. ESP B horizon	-0.28	-0.04	0.21	0.30	0.42	0.179	0.12
Clay ratio B/A	-0.57	-0.31	-0.05	0.41	1.44	0.854	-4.62

ME with the lower (l.l.) and upper (u.l.) limits of the 95% confidence interval, RMSE and mean ($\bar{\theta}$) and median ($\tilde{\theta}$) standardized squared prediction error, for validation and cross-validation procedure.

for the apparently weak relationships can be that the study area is spatially fairly homogeneous with respect to soil conditions. Absolute ranges of variation in magnitude are small and these small differences might not be easily captured by spatially structured covariates. This might also explain the large $\mathrm{Var}_{\mathrm{err}}/\mathrm{Var}_{\mathrm{tot}}$ ratio.

4. Fig. 2.8 shows that the strongest relationships occurred between soil properties, which should help to increase the accuracy in the prediction. However, for spatial prediction these relationships cannot be fully exploited, because at prediction locations only external drivers are available. The interrelationships between soil properties are used, but only by inserting predicted soil properties in the prediction equations (as effectively achieved by the inversion of $I - B$ in Eq. 2.4). Given the weak explanatory power of the external drivers, the resulting prediction maps shown in Fig. 2.10 are overly smooth and not accurate.

5. We did not use some of the more advanced functions for fitting SE models (Grace et al., 2010, 2012). For instance, we only used linear relationships and continuous properties. Interrelationships between external drivers were not

> taken into account either, which may cause misspecification, and thus loss of global fitting. Latent variables were constructed from single measured properties, which are not recommended in the literature (Jöreskog and Sörbom, 1982). Also, Grace (2006) suggests that estimated parameters are not reliable when global fitting is poor, as was the case in this case study. This also explains the large differences between the model-R^2 and cross-validation or validation results. We will look into more advanced SEM functions, such as non-linear relationships (Grace et al., 2012; Rosseel, 2012) in future research.

To verify if these are sensible reasons for poor model performance we modelled each soil property with a multiple linear regression (MLR) model using the same covariates as used in the SEM model. For each property, all covariates were offered to the MLR model. A stepwise algorithm was used for model selection. We assessed prediction accuracy of these models by cross-validation and independent validation as we did for the SEM model and computed the $\bar{\theta}$ and $\tilde{\theta}$ (Table 2.7). The results are consistent with the (cross-)validation results of the SE model (Table 2.6). Only the AVE of the MLR models is a fraction higher than those of the SEM models. The cross-validation $\bar{\theta}$ is slightly larger than one which might be caused by ignoring the uncertainty in the estimated regression coefficients in the calculation of the prediction error variance. The MLR (cross-)validation results confirm the suspicion that the poor results of SEM are caused by the data and not by the SE model.

2.4.2. SEM strengths and limitations

SEM is quite flexible with respect to what you measure: there need not be a one-to-one correspondence between a latent variable and a measurement variable. This opens up the possibility to measure "hidden" latent variables through multiple measurement variables, to measure combined latent variables through a single measurement variable, et cetera. This is quite common in the social sciences (Jöreskog and Sörbom, 1982; Bollen, 1989), where SEM has its origin. Intelligence, for instance, is measured based on answers to a large number of questions of an IQ test. This flexibility can be useful in DSM as well. For example, one could generate a map of soil fertility, which would be a latent variable that is measured through different soil properties, such as organic matter, pH, nitrogen and phosphorous concentration, and total bases.

SEM cannot reproduce the true physical, chemical and biological processes, such as achieved by dynamic soil-landscape models (e.g. Temme et al., 2006; Finke and Hutson, 2008; Vanwalleghem et al., 2013). These approaches, however, are typically deterministic (do not quantify uncertainties) and can be cumbersome, requiring a lot of modelling steps and many inputs. Instead, the model structure in SEM is based on hypotheses of the functioning of a soil-landscape system that is formalised in a conceptual model. This can be used as a tool to understand the interactions of the soil-landscape system, and thus its genesis and functioning.

Conventional soil maps are usually supplemented with reports that help the user to understand the maps. Likewise, SEM can be used in DSM to understand how soil property maps are linked together (as exemplified in the case study). In this sense, SEM may be useful to predict what happens when a system is intervened. For example, if we want to increase the amount of organic carbon in the soil, but the SE model shows that there is an indirect effect from concentration of sodium cation to organic carbon concentration, then a better target of our intervention may be to leach the sodium from the solum rather than to increase the input of organic carbon by manuring the soil. Literature about prediction in SEM often refers to this type of analysis (Grace et al., 2012).

Prediction is not commonly done in SEM applications. This is because in the social sciences and also in ecology the main interest is often inferential: which variables influence each other and what is the strength of the relationships? (Grace and Keeley, 2006; Delgado-Baquerizo et al., 2013). SE models are typically used to gain insight in the (causal) processes that drive a system. In DSM, we are typically more interested in prediction (mapping), although we can learn from the other science domains that we should also look at what the model tells us about processes, and whether these are conform our knowledge of the soil-landscape system.

SEM is a framework that does not consider autocorrelation per se. In this respect, Matteson et al. (2013) developed a procedure to test autocorrelation in model residuals using Moran's index (Moran, 1950) and to correct parameters based on this result. Alternatively, semivariogram analysis and kriging could be integrated into SEM. Lamb et al. (2014) developed an approach to apply spatial explicit SEM. Even though it is necessary to demonstrate the efficiency of this approach in practical applications, it looks promising.

It is possible to use categorical variables in SEM, but this would lead to a violation of the assumptions made in Eqns. 2.2 and 2.3 (Bollen, 1989). For this reason, some authors (Bollen, 1989; Rosseel, 2012) have developed techniques to use categorical variables as dummy variables by changing the model assumptions and parameter estimation methodology. As categorical variables are common in soil science, it would be useful to include these in future research.

SEM does not take the estimation errors of the model parameters into account. This might result in an underestimation of the prediction error variance. However, the influence of parameter estimation errors usually will be relatively small compared to that of the system noise, especially when the number of calibration observations is large.

Though both SEM and MLR showed similar poor performance (Tables 2.6 and 2.7), the AVE of MLR was slightly larger. This is not unexpected. MLR will always perform better than SEM as it is implemented here. The SEM implementation in this chapter is linear, and since MLR based on ordinary least squares is the best linear unbiased estimator in terms of minimum variance among all linear estimators it will outperform a linear SEM. Nevertheless, there are some important advantages of using SEM. SEM can be used for causal interpretation, unlike MLR that aims to get optimal predictions on the basis of empirical correlations. SEM can thus provide insight in the functioning of the soil-landscape system, and can be used as a tool to communicate soil process knowledge to soil scientists and surveyors. Vice versa, it allows explicit incorporation of conceptual soil-landscape knowledge in a statistical modelling framework for soil spatial prediction. With respect to the latter, however, we note that the functioning of a soil-landscape system as perceived by pedologists might not necessarily be correct (Brungard et al., 2015, e.g.).

2.5. Conclusions

We introduced and illustrated structural equations modelling for DSM. SEM takes a hybrid approach between mechanistic and empirical modelling. It converts a conceptual soil-landscape model into a statistically explicit model and uses this for prediction and testing whether the hypotheses involved are supported by observed data.

The main conclusions of this work are:

- SEM can be applied for soil spatial prediction.
- SEM can improve the consistency between multiple predicted soil properties because it uses knowledge about interrelationships between soil properties and predicts these properties simultaneously.
- SEM enhances the understanding of soil-landscape processes which is beneficial to land management.
- SEM handles measurement errors explicitly, which facilitates the separation of model error from measurement error and can have a marked effect on the results of soil spatial prediction.

- Accurate predictions cannot be expected in homogeneous areas with low signal-to-noise ratios.

- Validation studies must take measurement errors in validation data into account.

Appendix: lavaan syntax

This Section describes the lavaan syntax of the structural equation model used in this article and provides a summary of the fitted model. More advanced and detailed information about lavaan can be found in Rosseel (2012, 2013), cited in the article, or at http://lavaan.ugent.be/tutorial/index.html.

In order to translate the graphical conceptual model to lavaan, three major model parts have to be specified: the measurement model, the structural model and the measurement error.

Measurement model

```
my_model <- '
        # measurement model
        thick.Ar =~ 1*thick.A
        oc.Ar =~ 1*oc.A
        tb.Ar =~ 1*tb.A
        sat.Ar =~ 1*sat.A
        esp.Ar =~ 1*esp.A
        esp.Br =~ 1*esp.B
        btr =~ 1*bt
        '
```

The measurement model describes how each latent variable is measured. A latent variable can be a conceptual variable, such as soil fertility or soil degradation, that can only be measured indirectly, through one or more indicators. Although this is the most common use of latent variables, it is also possible to interpret a latent variable as the sum of the true (“real”) value of a variable and a measurement error. This is how we used the latent variable concept in this work. The variables ending with “r” are the “real” soil properties that “are measured through” (the “=∼” operator) the measured property. For example, while oc.Ar is the true (unknown) organic matter content of the A horizon, oc.A is its measured value. The measured properties are preceded by “1∗”, which indicates the scale of the latent variable relative to that of the measurement. In this case the scale is 1 because the observed property is a direct measurement of the latent variable.

Structural model

```
'
# structural model
thick.Ar ~  dem + wdist + mrvbf + vdchn + twi + river +
                slope + maxc + evim + evisd
oc.Ar ~     lstm +  lstsd + evim + evisd + dem + wdist +
                mrvbf + vdchn + twi + esp.Br + esp.Ar + btr +
                thick.Ar
tb.Ar ~     evim + evisd + lstm + lstsd + dem + wdist +
                mrvbf + vdchn + twi + river + oc.Ar + btr
sat.Ar ~    evim + evisd + lstm + lstsd + dem + wdist +
                mrvbf + vdchn +  twi + river + tb.Ar + oc.Ar
esp.Ar ~    lstm +  lstsd + dem + wdist + mrvbf + vdchn +
                twi + river + esp.Br
esp.Br ~    lstm +  lstsd + dem + wdist + mrvbf + vdchn +
                twi + river
btr ~       lstm +  lstsd + wdist + vdchn + twi + dem +
                river + mrvbf + esp.Br + esp.Ar
'
```

The structural model defines the interrelationships between state variables. In this case, seven equations define the dependencies of the soil properties on other soil properties and external covariates. The coefficients of the (linear) equations are estimated in the calibration process.

Measurement error

```
'
# measurement error
thick.A ~~  0.25*thick.A
oc.A ~~     0.20*oc.A
tb.A ~~     0.20*tb.A
sat.A ~~    0.20*sat.A
esp.A ~~    0.20*esp.A
esp.B ~~    0.10*esp.B
bt  ~~      0.25*bt
'
```

In this part the measurement error variance is defined. The "∼∼" operator represents variance–covariance between variables. In this case, the covariances are not specified which implies that they are assumed zero. The variances are given by the numbers in front of the symbol "∗". For instance, the soil organic carbon measurement error variance is specified as 0.25.

After defining the model, it is next calibrated using a data set, as follows:

```
my_fit <- sem(model = my_model, data = my_data, estimator = "ML")
```

lavaan summary

The first part of the summary report shows the lavaan version, the number of iterations used to reach convergence, the number of observations used to calibrate the model and the estimation method, which in this case is Maximum Likelihood (ML). The p-value (Chi-square) indicates whether the difference between the model and the data is statistically significant (a low *p* value indicates that differences are significant).

The second part of the report shows the estimated coefficients, their standard errors, their Z-value (because the coefficients were estimated using ML) and the p-value of the estimates. Coefficients without standard error are the ones that were fixed in the model specification part. Summary information is obtained with `summary(my_fit)`.

```
## lavaan (0.5-20) converged normally after  72 iterations
##
##   Number of observations                           320
##
##   Estimator                                         ML
##   Minimum Function Test Statistic              146.971
##   Degrees of freedom                                31
##   P-value (Chi-square)                           0.000
##
## Parameter Estimates:
##
##   Information                                 Expected
##   Standard Errors                             Standard
##
## Latent Variables:
##                    Estimate  Std.Err  Z-value  P(>|z|)
##   thick.Ar =~
##     thick.A           1.000
##   oc.Ar =~
##     oc.A              1.000
##   tb.Ar =~
##     tb.A              1.000
##   sat.Ar =~
##     sat.A             1.000
##   esp.Ar =~
##     esp.A             1.000
##   esp.Br =~
##     esp.B             1.000
##   btr =~
##     bt                1.000
##
## Regressions:
##                    Estimate  Std.Err  Z-value  P(>|z|)
##   thick.Ar ~
##     dem               0.258    0.107    2.409    0.016
##     wdist            -0.062    0.061   -1.009    0.313
##     mrvbf             0.034    0.078    0.441    0.659
##     vdchn             0.027    0.072    0.374    0.708
##     twi              -0.014    0.070   -0.204    0.838
##     river            -0.121    0.102   -1.189    0.234
##     slope             0.058    0.080    0.725    0.469
##     maxc             -0.104    0.059   -1.766    0.077
##     evim              0.029    0.057    0.513    0.608
##     evisd             0.152    0.068    2.245    0.025
##   oc.Ar ~
##     lstm             -0.330    0.079   -4.194    0.000
##     lstsd             0.212    0.080    2.637    0.008
##     evim             -0.016    0.049   -0.326    0.745
##     evisd             0.198    0.060    3.307    0.001
##     dem               0.076    0.095    0.802    0.422
##     wdist            -0.069    0.068   -1.004    0.316
##     mrvbf             0.112    0.082    1.367    0.171
##     vdchn             0.033    0.086    0.387    0.699
##     twi              -0.051    0.077   -0.664    0.507
##     esp.Br            0.943    0.862    1.094    0.274
##     esp.Ar           -1.374    0.972   -1.413    0.158
##     btr               0.160    0.246    0.652    0.514
##     thick.Ar         -0.006    0.063   -0.093    0.926
##   tb.Ar ~
##     evim             -0.097    0.052   -1.869    0.062
##     evisd             0.123    0.065    1.889    0.059
##     lstm              0.045    0.082    0.552    0.581
##     lstsd             0.020    0.065    0.303    0.762
##     dem              -0.139    0.115   -1.203    0.229
##     wdist            -0.108    0.057   -1.906    0.057
##     mrvbf            -0.025    0.063   -0.398    0.690
##     vdchn            -0.017    0.063   -0.273    0.785
##     twi              -0.012    0.062   -0.193    0.847
##     river            -0.153    0.132   -1.153    0.249
##     oc.Ar             0.542    0.081    6.720    0.000
##     btr              -0.041    0.086   -0.480    0.632
##   sat.Ar ~
##     evim             -0.003    0.051   -0.067    0.947
##     evisd            -0.202    0.063   -3.198    0.001
##     lstm             -0.171    0.078   -2.184    0.029
##     lstsd             0.083    0.062    1.346    0.178
##     dem              -0.020    0.112   -0.181    0.857
##     wdist             0.111    0.055    2.005    0.045
##     mrvbf            -0.008    0.061   -0.127    0.899
##     vdchn             0.013    0.061    0.222    0.825
##     twi               0.086    0.060    1.431    0.152
##     river             0.255    0.119    2.152    0.031
##     tb.Ar             0.934    0.086   10.880    0.000
##     oc.Ar            -0.737    0.097   -7.634    0.000
##   esp.Ar ~
##     lstm              0.052    0.047    1.097    0.273
##     lstsd             0.043    0.043    1.008    0.313
##     dem               0.007    0.070    0.095    0.925
##     wdist            -0.004    0.037   -0.117    0.907
##     mrvbf             0.044    0.042    1.049    0.294
##     vdchn             0.037    0.043    0.869    0.385
##     twi              -0.025    0.041   -0.623    0.533
##     river            -0.064    0.073   -0.886    0.376
##     esp.Br            0.905    0.045   20.228    0.000
##   esp.Br ~
##     lstm             -0.087    0.072   -1.212    0.226
##     lstsd             0.295    0.060    4.924    0.000
##     dem              -0.156    0.110   -1.419    0.156
##     wdist            -0.023    0.055   -0.426    0.670
##     mrvbf            -0.258    0.060   -4.331    0.000
##     vdchn            -0.261    0.061   -4.297    0.000
##     twi               0.211    0.058    3.623    0.000
##     river             0.139    0.119    1.166    0.244
##   btr ~
##     lstm             -0.134    0.097   -1.375    0.169
##     lstsd             0.002    0.089    0.027    0.978
##     wdist             0.030    0.075    0.406    0.685
##     vdchn            -0.109    0.090   -1.218    0.223
##     twi               0.007    0.084    0.083    0.933
##     dem               0.093    0.142    0.653    0.514
##     river            -0.478    0.164   -2.921    0.003
##     mrvbf            -0.033    0.090   -0.362    0.717
##     esp.Br           -1.460    0.665   -2.195    0.028
##     esp.Ar            1.799    0.707    2.544    0.011
##
## Variances:
##                    Estimate  Std.Err  Z-value  P(>|z|)
##     thick.A           0.250
##     oc.A              0.200
##     tb.A              0.200
##     sat.A             0.200
##     esp.A             0.200
##     esp.B             0.100
##     bt                0.250
##     thick.Ar          0.679    0.073    9.245    0.000
##     oc.Ar             0.425    0.079    5.370    0.000
##     tb.Ar             0.515    0.061    8.443    0.000
##     sat.Ar            0.241    0.057    4.239    0.000
##     esp.Ar            0.061    0.026    2.326    0.020
##     esp.Br            0.639    0.058   10.937    0.000
##     btr               0.249    0.097    2.568    0.010
```

Chapter 3

Multivariate mapping of soil with structural equation modelling

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 for 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. The main conclusion of this study was that SEM can be used to predict several soil properties in multiple layers by considering the interrelations between soil properties and layers.

Based on:
Angelini, M. E.; Heuvelink, G. B. M. and Kempen, B., 2017. Multivariate mapping of soil with structural equation modelling. *European Journal of Soil Science*, **68**, pp. 575–591.

3.1. 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 and 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 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 and 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 and Vanwalleghem, 2016). 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 (Chapter 2).

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 a previous study (Chapter 2), we demonstrated that it is possible to include interrelations between soil properties in the modelling process. In a case study we predicted in 2-D 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. In Chapter 2 we 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.

3.2. Materials and methods

3.2.1. Study area

The study area covers about 23 000 km^2 in the Argentine Pampas between 35° 00' S – 33° 17' S and 58° 55' W – 61° 21' W (Fig. 3.1). 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 (Soil Survey Staff, 2014) (Phaeozems in WRB classification (IUSS Working Group World Reference Base, 2006)) in association with Natracuolls and Natracualf-es (Solonetz in WRB classification) (Morrás and Moretti, 2016). In spite of its apparent homogeneity, the loess is derived from several sources that affect the soil chemical and physical properties (Morrás and Moretti, 2016).

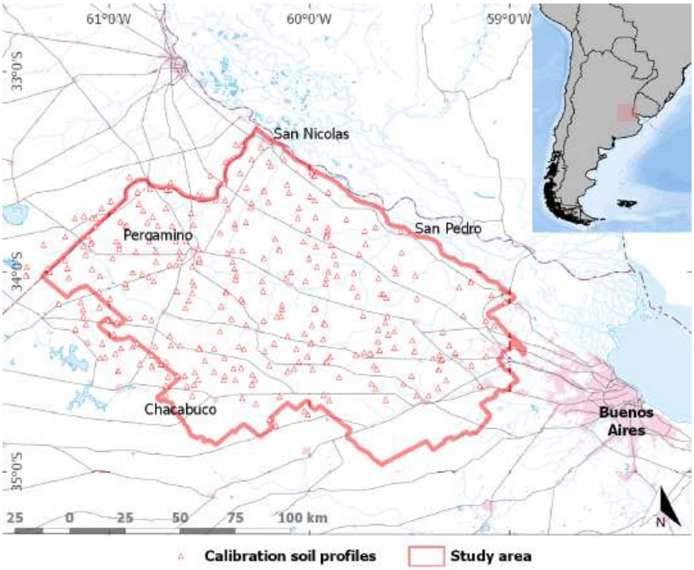


Figure 3.1: *Extent of the study area and locations of soil profiles used for calibration and cross-validation.*

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

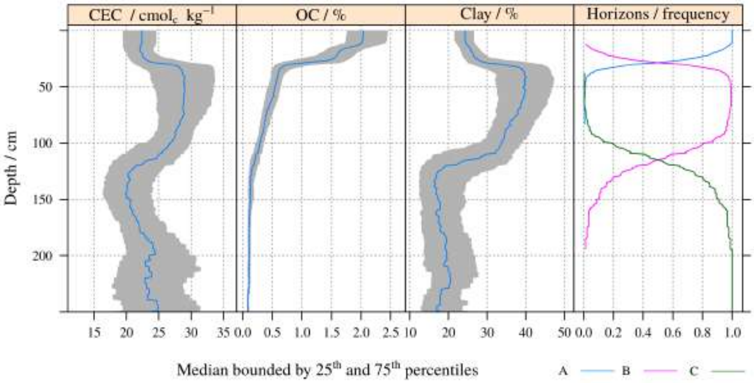


Figure 3.2: *Graphs of the median of cation exchange capacity (CEC), organic carbon (OC) and clay (Clay), as a function of depth; the grey area represents the 50% envelope between the 25th and 75th quantiles. Frequency of occurrence of each horizon type as a function of depth (Horizons).*

3.2.2. 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, 2015). Fig. 3.1 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 $\mathrm{cmol_c\,kg^{-1}}$ soil) that 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 and
- C horizon: usually represented as C, C2, R or X.

We did not include transitional horizons, such as AB, BA or BC. Fig. 3.2 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. Fig. 3.3 shows the correlations among soil properties and horizons. More detailed information about the soil data are provided in Chapter 2.

3.2.3. External factors

Table 3.1 summarises the external factors used in the modelling process. The main sources of information included the following. The 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 3.1.

Enhanced vegetation index (EVI [MOD13Q1]) and land-surface temperature and emissivity (LST [MOD11A2]) were taken from MODIS[1]. The standard deviation of a fifteen-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

[1] 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

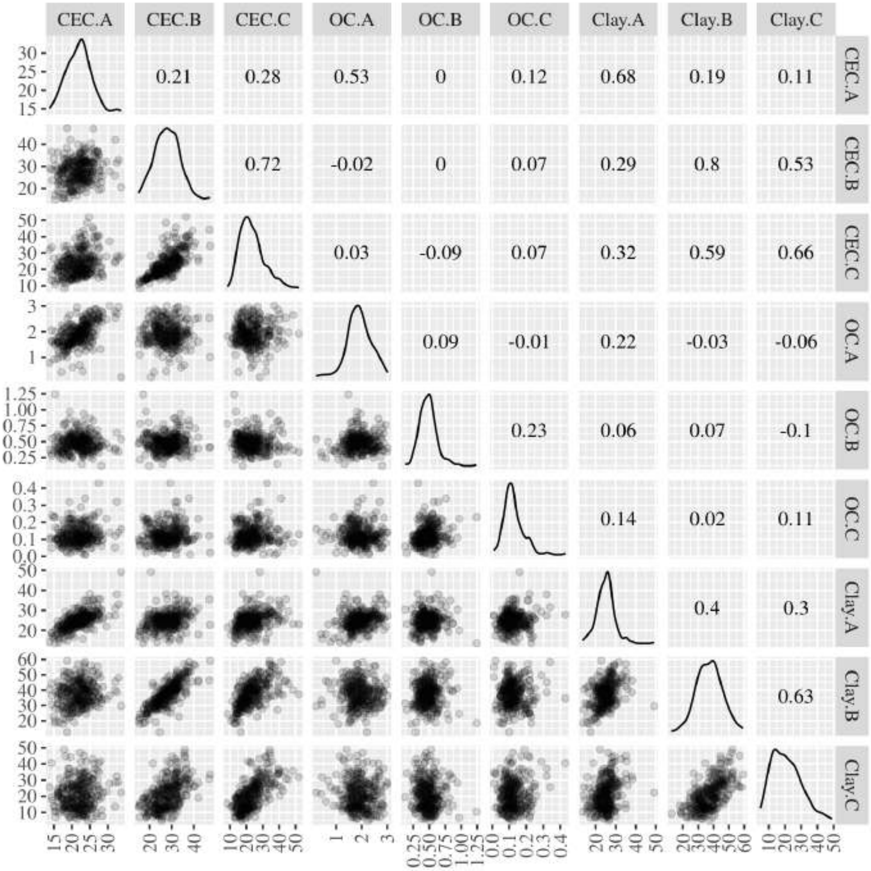


__Figure 3.3:__ Correlation graph of soil properties by horizons. The upper right triangle shows the correlation between properties, the diagonal presents the histograms of the properties and the lower left triangle the scatter plots. Soil properties are abbreviated such that the name of the soil property 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.

depends on soil texture, among other factors. We also computed the normalised 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 and Moretti, 2016). All variables were standardized by subtracting their mean and dividing by their standard deviation.

Table 3.1: *External factors*

Factor	Description	Source	Resolution
LSTM	Mean of 14 years of daytime 8-day land-surface temperature	Terra/MODIS, product MOD11A2.	1000 m
EVISD	Standard deviation of 14 years of enhanced vegetation index (EVI) 16 days	Terra/MODIS, product MOD13Q1	250 m
NDWI.A	Normalised difference water index (NDWI) bands NIR (~850 nm) and SWIR (~1240 nm). Summer season.	MODIS product MCD43A4	500 m
NDWI.B	Normalised 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)		
LAT	Latitude of plain coordinates (metres)	–	30 m
LON	Longitude of plain coordinates (metres)	–	30 m

3.2.4. Modelling framework

To formulate, apply and evaluate an SE model we divided the modelling process into seven steps (Fig. 3.4):

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):

$$\mathbf{x} = \boldsymbol{\Lambda}\boldsymbol{\xi} + \boldsymbol{\delta} \tag{3.1}$$

$$\mathbf{y} = \mathbf{K}\boldsymbol{\eta} + \boldsymbol{\varepsilon} \tag{3.2}$$

$$\boldsymbol{\eta} = \mathbf{B}\boldsymbol{\eta} + \boldsymbol{\Gamma}\boldsymbol{\xi} + \boldsymbol{\zeta} \tag{3.3}$$

   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, $\boldsymbol{\Lambda}$ and $\mathbf{K}$ are $q \times n$ and $p \times m$ coefficient matrices that link observed to latent variables, $\boldsymbol{\delta}$ and $\boldsymbol{\varepsilon}$ are vectors of measurement errors of length q and p, respectively (mutually independent and zero-mean normal deviates), $\mathbf{B}$ and $\boldsymbol{\Gamma}$ are $m \times m$ and $m \times n$ coefficient matrices of endogenous and exogenous relations and $\boldsymbol{\zeta}$ is vector of length m of model error for variable $\boldsymbol{\eta}$. Note that the diagonal elements of $\mathbf{B}$ are forced to zero so that soil properties cannot depend on themselves. Eqns. 3.1 and 3.2 define the measurement model, whereas Eq. 3.3 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 $\boldsymbol{\Theta}_{\boldsymbol{\varepsilon}}$ are $q \times q$ and $p \times p$ variance–covariance matrices of $\boldsymbol{\delta}$ and $\boldsymbol{\varepsilon}$.

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–covari-ance matrix $\Sigma(\boldsymbol{\theta})$, 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 $\boldsymbol{\Theta}_{\boldsymbol{\varepsilon}}$). 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

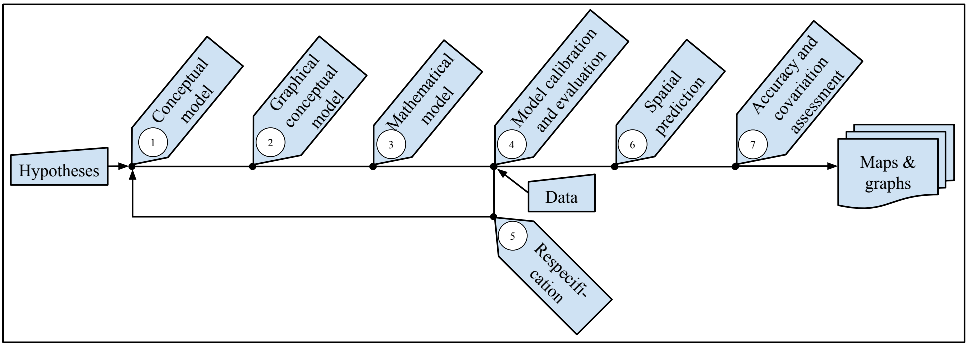


Figure 3.4: *Steps in structural equation modelling (SEM) for spatial prediction of soil properties.*

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 Eqns. 3.1 and 3.3 (Section 2.2.5):

$$\hat{\eta} = (\mathbf{I} - \mathbf{B})^{-1}\mathbf{\Gamma}\mathbf{\Lambda}^{-1}\mathbf{x} \tag{3.4}$$

   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 (Eq. 2.5).

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 Chapter 2, except for steps 4, 5 and 7. These are given in more detail below.

3.2.5. 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 Eqns. 3.1 to 3.4:

$$\boldsymbol{\Sigma}(\boldsymbol{\theta}) = (\mathbf{I} - \mathbf{B})^{-1} \left(\boldsymbol{\Gamma}\boldsymbol{\Phi}\boldsymbol{\Gamma}^{T} + \boldsymbol{\Psi}\right) \left((\mathbf{I} - \mathbf{B})^{-1}\right)^{T} + \boldsymbol{\Theta}_{\boldsymbol{\varepsilon}} \tag{3.5}$$

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 Eq. 3.3. 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 this soil mapping example, but the methodology also applies more generally (e.g. Bollen, 1989, Chapter 8).

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 and 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 cutoff value of 0.95, although it is case-dependent (Marsh et al., 2004). In addition to these measures, we computed the model R^2.

3.2.6. 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 $\mathbf{B}$, $\mathbf{\Gamma}$ and $\mathbf{\Psi}$ only, i.e. which endogenous variables depend on other endogenous and exogenous variables, and on the covariance of system noise between endogenous variables.

3.2.7. Model accuracy and assessment of covariation

In Chapter 2, 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:

$$AVE = 1 - \frac{\sum_i^n (y_i - \hat{y})^2}{\sum_i^n (y_i - \bar{y})^2} \tag{3.6}$$

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

We compute the mean ($\bar{\theta}$) and median ($\tilde{\theta}$) 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 S and the model-induced covariance matrix $\Sigma(\boldsymbol{\theta})$. 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 $\Sigma(\boldsymbol{\theta})_{MLR}$. We compared this with the correlation matrix of the observations and computed the SRMR_{MLR}.

3.3. Results

3.3.1. 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 and Weil, 2014).

The soil of the study area has small amounts of OC: one to three percent in A horizons, and typically less than one percent in B and C horizons (Fig. 3.2). 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 south-west to north-east. The loess deposits have been reworked by aeolian and fluvial processes (Morrás and 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 (Fig. 3.2). 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

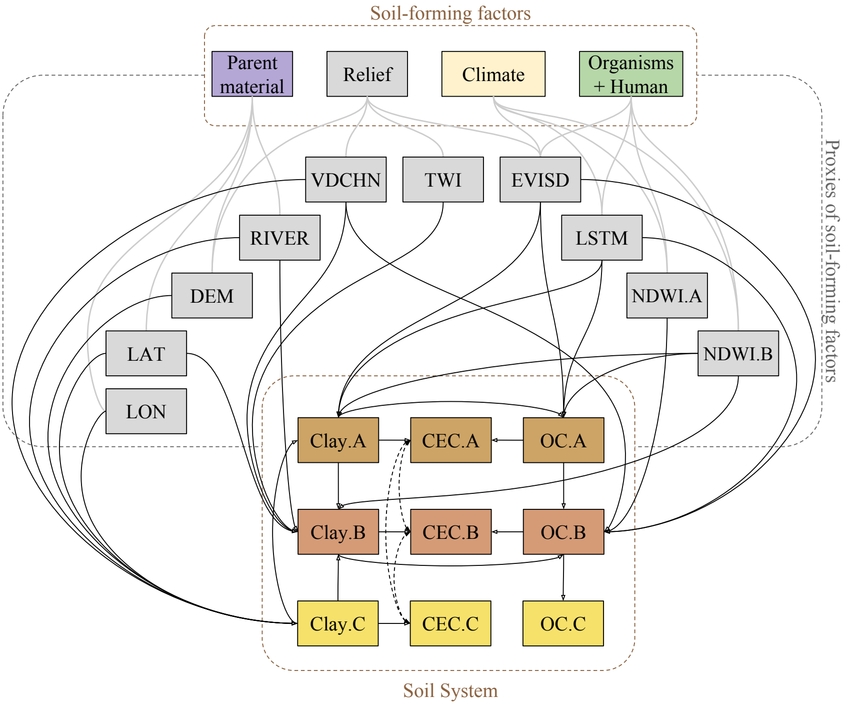


Figure 3.5: *Graphical model. Grey continuous lines represent the theoretical relation between soil-forming factors and external factors. Black continuous arrows are cause and effect links. Black dashed arrows are expected correlations between system errors. External factors are described in Table 1. 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.*

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 and Weil, 2014).

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

3.3.2. Graphical and mathematical model

The conceptual model, which characterises the main forces and processes that control the distribution of CEC, clay and OC, was transformed into a graphical model (Fig. 3.5). Fig. 3.6 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 (Eqns. 3.1 and 3.2) which comprises the matrices $\mathbf{\Lambda}$, $\mathbf{K}$, $\mathbf{\Theta_\delta}$ and $\mathbf{\Theta_\varepsilon}$. We assumed that the external factors are observed deterministic variables, therefore, $\mathbf{\Lambda}$ is an identity matrix and $\mathbf{\Theta_\delta}$ 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 characterised by $\mathbf{\Theta_\varepsilon}$. The diagonal elements of $\mathbf{\Theta_\varepsilon}$ 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 Eq. 3.3 is defined by $\mathbf{\Gamma}$, $\mathbf{B}$ and $\mathbf{\Psi}$. The elements of these matrices have a non-zero value only if there are corresponding arrows in the graphical model. Thus, we obtain:

$$\mathbf{\Gamma} = \begin{bmatrix} \gamma_{11} & \gamma_{12} & 0 & \gamma_{14} & 0 & 0 & 0 & 0 & 0 & 0 \\ \gamma_{21} & \gamma_{22} & \gamma_{23} & 0 & 0 & \gamma_{26} & 0 & 0 & 0 & 0 \\ 0 & \gamma_{32} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \gamma_{41} & 0 & 0 & \gamma_{44} & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \gamma_{54} & \gamma_{55} & \gamma_{56} & \gamma_{57} & 0 & \gamma_{59} & 0 \\ 0 & 0 & 0 & 0 & 0 & \gamma_{66} & \gamma_{67} & \gamma_{68} & \gamma_{69} & \gamma_{610} \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \tag{3.7}$$

$$\mathbf{B} = \begin{bmatrix} 0 & 0 & 0 & \beta_{14} & 0 & 0 & 0 & 0 & 0 \\ \beta_{21} & 0 & 0 & 0 & \beta_{25} & 0 & 0 & 0 & 0 \\ 0 & \beta_{32} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & \beta_{46} & 0 & 0 & 0 \\ 0 & 0 & 0 & \beta_{54} & 0 & \beta_{56} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \beta_{71} & 0 & 0 & \beta_{74} & 0 & 0 & 0 & 0 & 0 \\ 0 & \beta_{82} & 0 & 0 & \beta_{85} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & \beta_{96} & 0 & 0 & 0 \end{bmatrix} \tag{3.8}$$

$$\mathbf{\Psi} = \begin{bmatrix} \psi_{11} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & \psi_{22} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \psi_{33} & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \psi_{44} & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \psi_{55} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & \psi_{66} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \psi_{77} & \psi_{78} & \psi_{79} \\ 0 & 0 & 0 & 0 & 0 & 0 & \psi_{87} & \psi_{88} & \psi_{89} \\ 0 & 0 & 0 & 0 & 0 & 0 & \psi_{97} & \psi_{98} & \psi_{99} \end{bmatrix} \tag{3.9}$$

For example, γ_{12} refers to the arrow in Fig. 3.6 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), 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.

3.3.3. 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 3.2, 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 did an exploratory respecification analysis that provides suggestions of what can be included in the model.

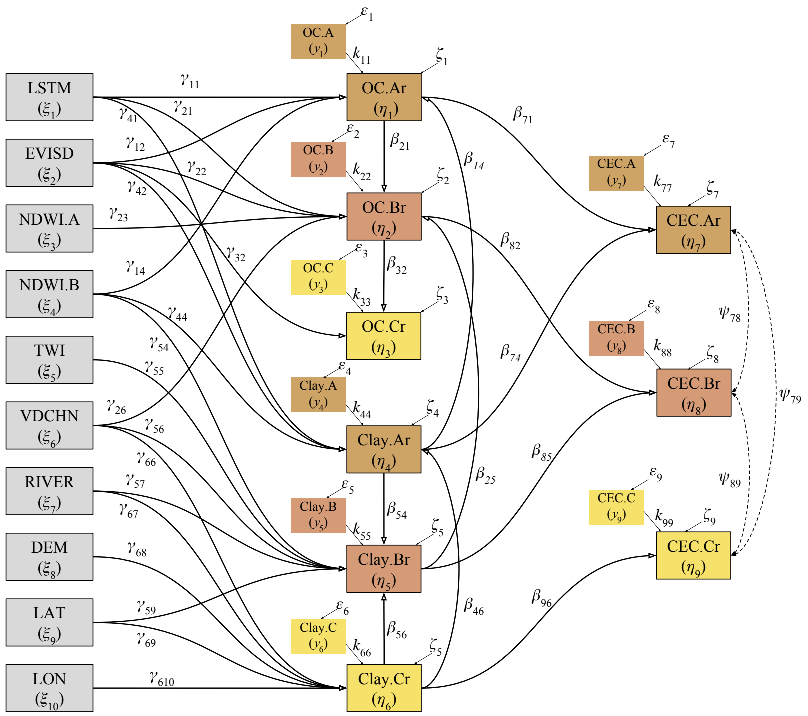

LSTM (ξ_1)
EVISD (ξ_2)
NDWI.A (ξ_3)
NDWI.B (ξ_4)
TWI (ξ_5)
VDCHN (ξ_6)
RIVER (ξ_7)
DEM (ξ_8)
LAT (ξ_9)
LON (ξ_{10})
γ_{11} γ_{21} γ_{41} γ_{12} γ_{22} γ_{42} γ_{23} γ_{14} γ_{32} γ_{44} γ_{54} γ_{55} γ_{26} γ_{56} γ_{66} γ_{57} γ_{67} γ_{68} γ_{59} γ_{69} γ_{610}
OC.A (y_1) ε_1 k_{11}
OC.Ar (η_1) ζ_1
OC.B (y_2) ε_2 k_{22}
OC.Br (η_2) ζ_2
OC.C (y_3) ε_3 k_{33}
OC.Cr (η_3) ζ_3
Clay.A (y_4) ε_4 k_{44}
Clay.Ar (η_4) ζ_4
Clay.B (y_5) ε_5 k_{55}
Clay.Br (η_5) ζ_5
Clay.C (y_6) ε_6 k_{66}
Clay.Cr (η_6) ζ_5
β_{21} β_{32} β_{54} β_{56} β_{14} β_{25} β_{46}
β_{71} β_{82} β_{74} β_{85} β_{96}
CEC.A (y_7) ε_7 k_{77}
CEC.Ar (η_7) ζ_7
CEC.B (y_8) ε_8 k_{88}
CEC.Br (η_8) ζ_8
CEC.C (y_9) ε_9 k_{99}
CEC.Cr (η_9) ζ_9
ψ_{78} ψ_{79} ψ_{89}

Figure 3.6: *Graphical model with parameters. Thick continuous arrows represent* **B** *and* **Γ** *matrices, thin continuous arrows represent* **K**, **Ψ** *and* $\mathbf{\Theta}_\varepsilon$ *matrices, and dashed double-headed arrows represent the model error correlations. External factors (grey boxes) are described in Table 3.1. 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).*

3.3.4. 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 Fig. 3.6 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 from 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 from RIVER on soil properties (γ_{67}, γ_{57}). After these modifications we obtained new measures of model fit (Table 3.2, 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 3.3 lists the first group of suggestions that were included (step 2). These modifications improved all measures (Table 3.2, 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 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 3.3). The measures of fit show a large improvement with CFI and GFI close to one (Table 3.2, step 3).

Finally, we included suggestions for the residual variance–covariance (Table 3.3, 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 with 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 3.2, step 4).

Table 3.2: *Changes in model-fitting measures after every respecification step.*

Step	χ^2	df	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

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

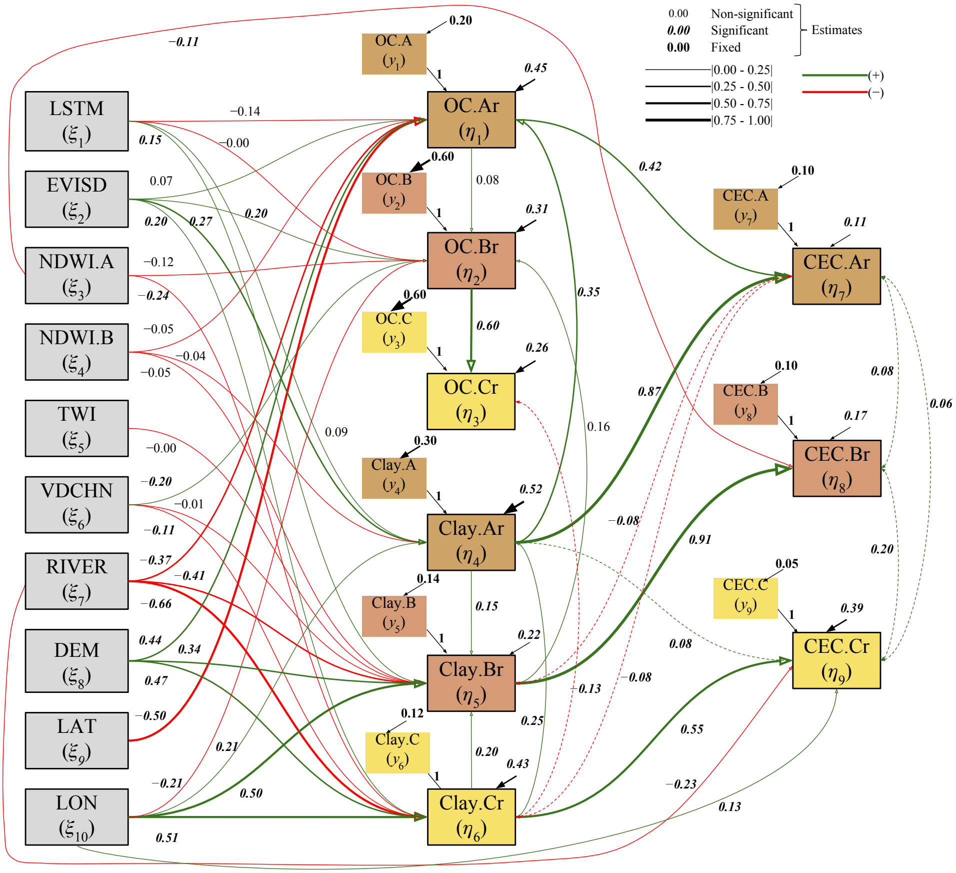


Figure 3.7: *Final graphical fitted model. Arrow thickness represents the magnitude of coefficients and their colour is the sign. Black arrows represent elements of* $\mathbf{K}$, $\mathbf{\Psi}$ *and* $\mathbf{\Theta}_\varepsilon$ *matrices. Dashed arrows represent model error correlations. Bold italic 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.*

Table 3.3: *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 re-specification process (Section 3.3.3). 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`.

The respecified model was fitted by maximum likelihood estimation. The resulting graphical model with parameter estimates is shown in Fig. 3.7. 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.

3.3.5. Spatial prediction

Fig. 3.8 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 (north-east boundary), which was used to represent parent material. The same

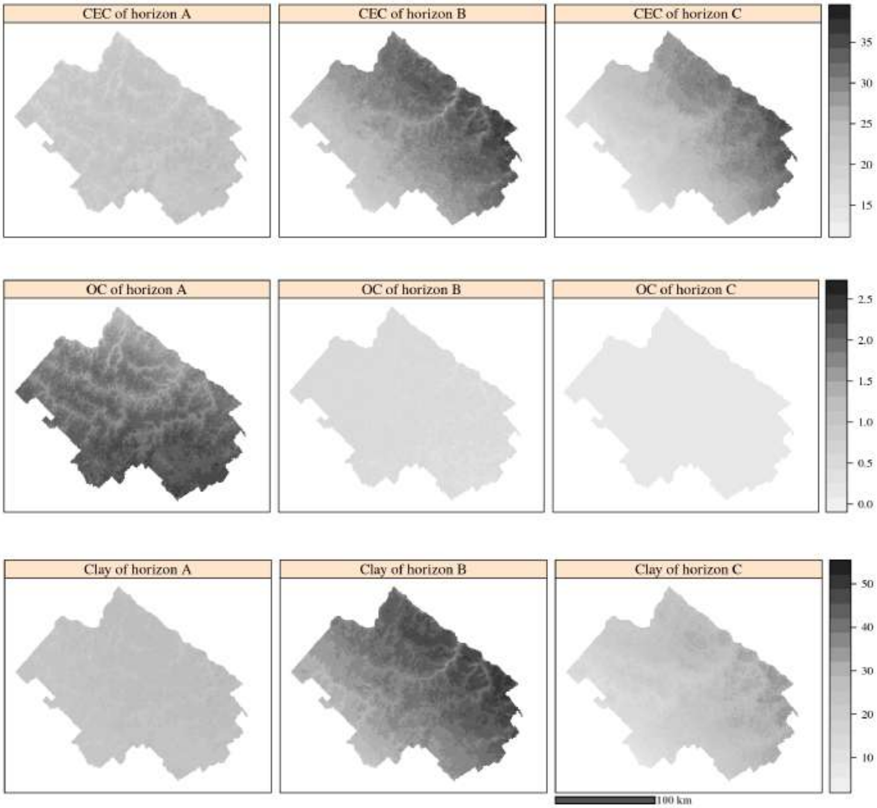


Figure 3.8: *Maps of cation exchange capacity (CEC) (cmol$_c$ kg^{-1}), organic carbon (OC) (mass %) and clay (mass %) for the A, B and C horizons.*

pattern also occurs in the maps of clay, which was expected because of the strong relation 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.

3.3.6. Model accuracy and assessment of covariation

Table 3.4 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

Table 3.4: *Cross-validation and measures of model fit.*

SP	**Hor.**	**ME**	**RMSE**	$\bar{\theta}$	$\tilde{\theta}$	**AVE**	$\mathbf{R^2}$
		SEM					
CEC		−0.004	4.30			0.53	
OC	Joint	0.000	0.25			0.91	
Clay		−0.007	5.45			0.72	
	A	0.002	3.16	1.03	0.40	0.18	0.21
CEC	B	−0.009	3.00	1.05	0.39	0.50	0.52
	C	−0.008	5.46	0.97	0.23	0.45	0.47
	A	0.000	0.40	1.07	0.38	0.24	0.27
OC	B	0.000	0.14	1.06	0.33	0.03	0.06
	C	0.000	0.06	1.02	0.37	0.02	0.03
	A	0.000	4.05	1.03	0.30	0.15	0.18
Clay	B	−0.013	5.14	0.99	0.40	0.60	0.62
	C	−0.013	6.87	1.05	0.53	0.41	0.44
		MLR					
	A	0.006	3.23	1.05	0.38	0.14	0.21
CEC	B	−0.009	4.04	1.06	0.36	0.49	0.53
	C	−0.003	5.48	1.03	0.24	0.45	0.49
	A	0.000	0.41	1.05	0.42	0.22	0.28
OC	B	0.000	0.14	1.04	0.35	0.00	0.08
	C	0.000	0.06	1.05	0.35	−0.05	0.04
	A	0.007	4.17	1.07	0.31	0.10	0.19
Clay	B	−0.010	5.21	1.05	0.41	0.59	0.63
	C	−0.011	6.90	1.04	0.52	0.40	0.45

Soil property (SP), horizon (Hor.), Mean error (ME), root mean squared error (RMSE), mean ($\bar{\theta}$) and median ($\tilde{\theta}$) of the standardized squared prediction error, amount of variance explained (AVE) and R^2 is the coefficient of determination of the model fit. OC, organic carbon; CEC, cation exchange capacity.

clay of the B and C horizons is relatively large, and so is that for CEC. Fig. 3.9 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 3.4. 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 3.4) 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 ($\bar{\theta}$) close to 1, although their medians ($\tilde{\theta}$) have slightly smaller values than the theoretical value 0.455. The RMSE shows that prediction accuracy decreases with depth

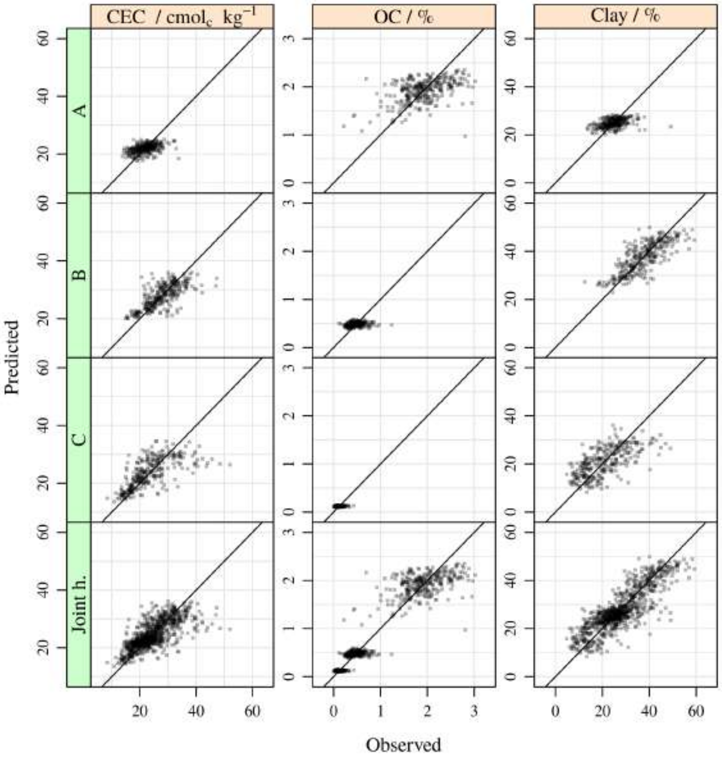


Figure 3.9: *Scatter plots of observed against predicted soil properties obtained by cross-validation. Columns of graphs are soil properties: cation exchange capacity (CEC), organic carbon (OC) and clay. Rows of graphs are horizons A, B and C, and "Joint h." represents the three horizons joined.*

for CEC and clay, which have maximum values of 5.5 cmol$_c$ kg^{-1} for CEC.C and almost 7% for Clay.C.

Fig. 3.10 shows the S, $\Sigma(\theta)$ and $\Sigma(\theta)_{MLR}$ 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 $\Sigma(\theta)$ and S than between $\Sigma(\theta)_{MLR}$ and S. Fig. 3.11 shows the absolute values of $S - \Sigma(\theta)$ and $S - \Sigma(\theta)_{MLR}$, which confirms this result. Improved performance of SEM is also confirmed by the SRMR, which is 0.024 for SEM, whereas SRMR$_{\text{MLR}}$ 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

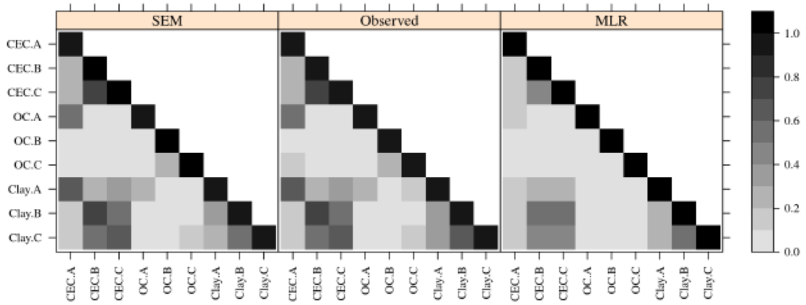


Figure 3.10: *Correlation matrix of observations (Observed), derived from the structural equation model (SEM) and with multiple linear regression (MLR).*

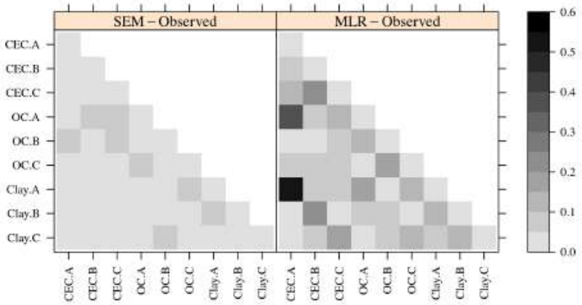


Figure 3.11: *Absolute difference between correlation matrix of original data and structural equation modelling (SEM – Observed) and multiple linear regression (MLR – Observed).*

observed covariation much better.

3.4. Discussion

3.4.1. The conceptual soil-landscape model

The fitted graphical model in Fig. 3.7 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 and 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, Fig. 3.7 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 the hypothesis of Kröhling and Iriondo (2003), 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 to 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.

Fig. 3.7 shows that NDWI of spring (NDWI.B) and TWI have a small effect on soil properties, which means that either their 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. Fig. 3.7 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.

3.4.2. 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 cutoff value of these measures because they are case dependent

(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 done 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 data set to prevent overfitting of the model (Bollen, 1989). We used cross-validation for this without using the observation that was put aside.

3.4.3. Representing soil information with SEM

The resulting SEM graph (Fig. 3.7) in combination with the maps (Fig. 3.8) 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 the fat arrow between these properties in Fig. 3.7. The 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).

3.4.4. Model accuracy

The maps of clay and consequently CEC from B and C horizons are reasonably accurate (Table 3.4 and Fig. 3.9). The maps of OC of the B and C horizons show little spatial variation (Fig. 3.8) and have poor accuracy (Table 3.4). The latter might be caused by the lack of spatial variation, the small amount of OC in these horizons and relatively large measurement error (Fig. 3.7). 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 (Kröhling and Iriondo, 2003). In general, landscape properties can explain variation in soil properties of the top layers with greater accuracy than for deeper layers (e.g. Kempen, 2011). In this 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 3.4 and Fig. 3.9). 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 of 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 3.4). 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 overfitting 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 the model results could be improved further by taking spatial autocorrelation into account, we fitted variograms (Webster and Oliver, 2007) to the SEM cross-validation residuals (Fig. 3.12). They show that spatial correlation in the residuals of 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 and Weisberg, 2011), which quantifies the cross-correlations between residuals of the linear regressions for each soil property. A MvLR model was fitted to the 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 re-

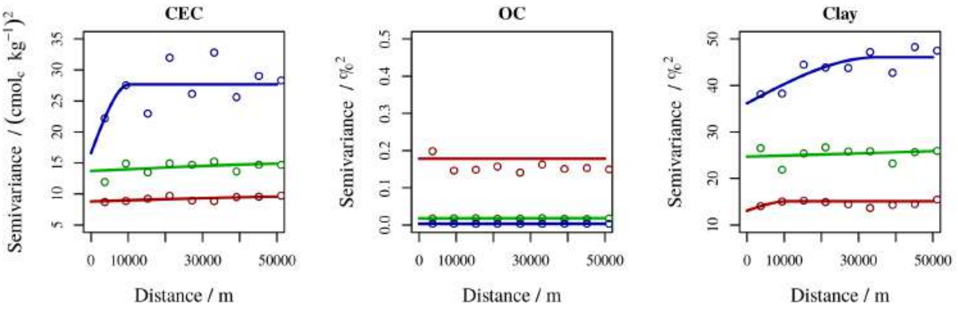


Figure 3.12: *Experimental (dots) and fitted (solid line) variograms of cross-validation residuals of cation exchange capacity (CEC), organic carbon (OC) and clay. Red lines and dots represent the A horizon, green the B horizon and blue the C horizon.*

production of the cross-correlations can be achieved. A MvLR model is rarely fitted in practice this approach adds many extra parameters that need to be estimated. In this case, with nine soil properties the MvLR model would involve $9 \cdot (9 - 1)/2 = 36$ 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 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.

3.5. 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.

The conclusions of this chapter are:

- 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 to interrelations between soil properties and to 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.

Chapter 4

Extrapolation of a structural equation model for digital soil mapping

In theory, two separate regions with the same soil-forming factors should develop similar soil conditions. This theoretical finding has been used in digital soil mapping (DSM) to extrapolate a prediction model from one area to another, which usually do not work out well. One reason for failure could be that most of these studies used empirical methods. Structural equation modelling (SEM) is a semi-mechanistic technique, which can explicitly include expert knowledge. We therefore hypothesise that structural equation (SE) models are more suitable for extrapolation than purely empirical models in DSM. The objective of this study was to investigate the extrapolation capability of SEM by comparing different model settings. We applied a SE model from a previous study in Argentina to a similar soil-landscape in the Great Plains of the United States to predict clay, organic carbon, and cation exchange capacity for three major horizons: A, B, and C. We evaluated the performance of the SEM mathematical model, as well as the extrapolation of the conceptual model. We concluded that system relationships that were well supported by pedological knowledge showed consistent and equal behaviour in both study areas. In addition, a deeper understanding of indicators of soil-forming factors could strengthen conceptual models for extrapolating DSM models. We also found that for model extrapolation, knowledge-based links between system variables are more effective than data-driven links. In particular, model modifications can improve local prediction but harm the predictive power of extrapolation.

Based on:
Angelini, M. E.; Kempen, B.; Temme, A. J. A. M.; Heuvelink, G. B. M. and Ransom M. D., 2017. Extrapolation of a structural equation model for digital soil mapping. *Soil Science Society of America Journal.* Submitted.

4.1. Introduction

In theory, two separate regions with the same soil-forming factors should develop similar soil conditions. This is explained by the fundamental theory of soil development developed by Dokuchaev (1883). Although this theory makes us of an oversimplification of the soil-forming factors, it is still very useful to digital soil mapping. Mallavan et al. (2010) used this assumption to develop the homosoil concept, with the objective of being able to extrapolate a model from one area to another. Several other authors have tried to do so as well (Lagacherie et al., 1995; Bui and Moran, 2003; Grinand et al., 2008; Malone et al., 2016; Silva et al., 2016). Most of them aimed to predict soil classes, while little has been done in terms of model extrapolation for soil properties.

The extrapolation of a model from one area to another faces several challenges. First, the soil-forming factors of the calibration area are hardly ever equal to those of the extrapolation area. For example, Malone et al. (2016) found differences between covariates of a calibration area and the same covariates of a extrapolation area located within a similar region. Second, even though the soil-forming factors may be equal in the present, the type and intensity of soil-forming factors may not have been equal in the past (Sommer et al., 2008; Temme and Veldkamp, 2009). Third, soil evolution is not linear and might exhibit chaotic behaviour, which is dynamic that hardly ever can be equal in two different regions (e.g. Huggett, 1998; D'Amico et al., 2017). Fourth, the lack of reproducing the actual mechanistic processes in empirical models makes them more difficult to be extrapolated than mechanistic models (Clark and Gelfand, 2006). As a result, map accuracy may drop substantially between the calibration and extrapolation area.

Structural equation modelling (SEM) is a semi-mechanistic multivariate technique commonly used to model cause-effect relationships of complex systems (Grace et al., 2012). SEM allows explicit use of process knowledge in a statistical modelling framework. It has been mainly applied in the social sciences and ecology (e.g. Anderson and Gerbing, 1988; Grace et al., 2010). Previously, we applied SEM to model and predict seven soil properties in a 23 000 km^2 study area in the Argentinian Pampas (Chapter 2). Following this, we tested SEM for simultaneous prediction of three soil properties (Clay, CEC and SOC) at three major horizons (A, B and C) for the same study area (Chapter 3).

The main objective of this study is to analyse whether SEM is a suitable method for extrapolation of digital soil mapping models, since its hybrid features between empirical and mechanistic approaches allows to explicitly include pedological knowledge. We therefore hypothesise that an SE model has better extrapolation capabili-

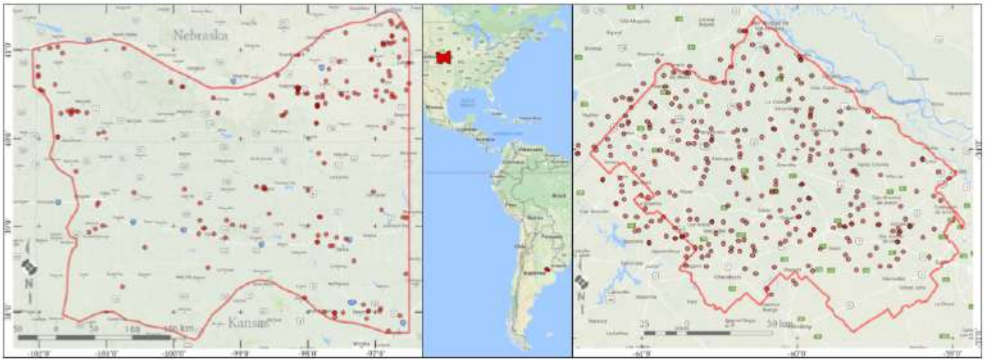


Figure 4.1: *Maps of the US (left) and Argentinian (right) study areas. Red dots represent soil profiles, red lines the study area boundaries. Central map shows locations of both areas (in red).*

ties than a purely empirical approach, such as multiple linear regression (MLR). The specific objectives of this study are: (1) to investigate the extrapolation capability of SEM by applying a structural equation (SE) model from the Argentinian pampas to a similar soil-landscape in the Great Plains of the United States, (2) to compare the performance of SEM extrapolation with that of a MLR model, (3) to quantify the effect of model respecification on the extrapolation, and (4) to discuss the differences between the models from the calibration and extrapolation areas.

4.2. Materials and methods

4.2.1. Study areas

The study areas chosen for this study share similarities in term of parent material (loess and loess-like sediments) and soil types (Phaeozems in WRB classification (IUSS Working Group World Reference Base, 2006), Ustolls and Udolls in Soil Taxonomy classification (Soil Survey Staff, 2014)). The Soil data Section summarises the differences in soil property distributions between both datasets.

Argentina

The 23 000 km^2 Argentinian study area is in the Rolling Pampas region (Fig 4.1). Originally, these were grassland plains developed under loess materials. Land use changed a century ago to cropland (Viglizzo et al., 2004) that now dominates the area. Typic and Vertic Argiudolls (Soil Survey Staff, 2014) are prevalent in uplands

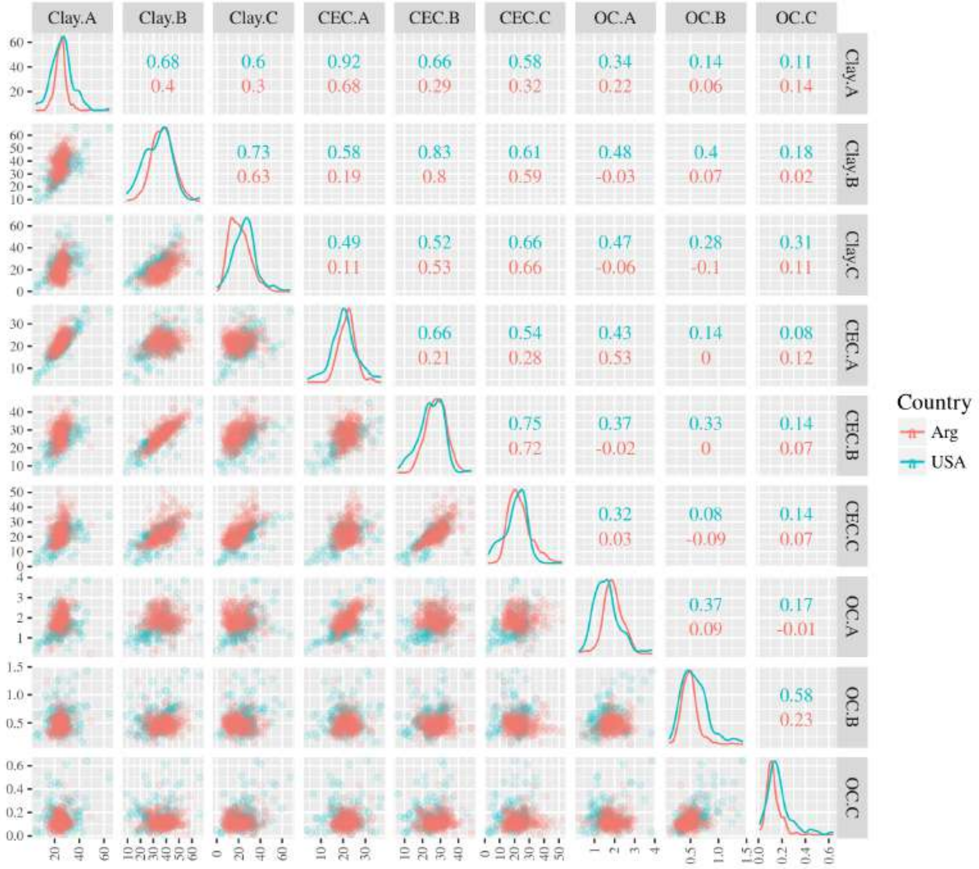


Figure 4.2: Correlation graph of soil properties by horizons and by country (in colours). The upper right triangle shows the Pearson correlation between properties, the diagonal presents the histograms (as density plot) of the properties and the lower left triangle the scatter plots. Soil properties are abbreviated such that the name of the soil property is followed by the horizon name and separated by a dot.

associated with Natraquolls, Natrudolls and Natraqualfs in most wetlands. The main clay minerals in the area are illite and smectite (Morrás and Moretti, 2016). The annual precipitation is between 900 and 1000 mm. The average summer temperature is 23℃, and the average temperature in winter is 10℃. The summer is characterised with a water deficit, while during winter there is rainfall excess (Cabrini and Calcaterra, 2008).

United States

The 150 000 km^2 US study area is in the Great Plains, covering parts of Nebraska and Kansas. The Platte and Arkansas rivers form the area's northern and southern boundaries, the western boundary generally follows the Kansas–Colorado boundary, and the eastern boundary runs north–south through Manhattan, Kansas. The original land cover in the area was grassland, developed on a landscape formed in horizontally stratified limestone-shale sequences (the Shale Hills) that are widely covered in loess and other aeolian sediment. Smectite is the predominant mineral of the clay fraction of these materials, along with randomly interstratified mica-smectite (Gunal and Ransom, 2006). In the east of the study area, land use has mainly changed into crop land, whereas grassland remains in use for grazing cattle in the west. Upland soils, developed in the loess-covered limestone-shale lithology include Argiudolls, Argiustolls, and younger soils in river valleys range from Hapludolls and Haplustolls (Soil Survey Staff, 2016). Precipitation varies from about 800 mm in the east to about 500 mm in the west of the study area (Goodin, 1995). The average summer temperature is about 26℃, and the average temperature in winter is about 1℃.

4.2.2. Soil data

The soil data from Argentina are described in detail in Chapter 3. We give a brief summary here. We obtained 344 soil profiles from the INTA database (INTA, 2015). Three soil properties were selected for modelling and mapping: soil organic carbon (OC, in weight percentage), clay content (Clay, in weight percentage) and cation exchange capacity (CEC, in cmol$_c$ kg^{-1} soil) at three major horizons: A, B and C. We did not include transitional horizons, such as AB, BA or BC.

For the US study area, we took 492 soil profiles from the SSURGO2 database (Soil Survey Staff, 2016). These were selected according to the same criteria as used for the Argentinian data: a profile must have an A, B and C horizon, must not have missing values for CEC, OC and Clay, and the horizons should not belong to a buried soil profile that might be indicative of a parent material discontinuation. We grouped all subdivisions of A, B and C horizons, such as Ap, A1, A2, Bt, Bw, etc. to the master horizon level. We excluded transitional horizons, such as AB, BA, AC, BC, etc. The locations of the final set of profiles for both study areas are shown in Fig. 4.1.

One important aspect of the two study areas to consider is the potential difference in parent material. Though both areas have parent material of aeolian origin, the amount of clay and mineralogy can be different. This could complicate the extrapolation of a model from one area to the other. Graphs of the distributions of each soil property for the two datasets are given in Fig. 4.2. The lower left panel graphs show scatter plots of the soil properties whereas the upper right part presents Pearson correlation coefficients. The diagonal graphs show the distribution of each soil property as a density plot. The graphs and statistics show that total clay percentages for both areas are very similar for A and B horizons, and that the C horizon in Argentina generally has a bit less clay than in Kansas–Nebraska. CEC values also are quite similar, with mean CEC about 20–25 cmol$_c$ kg^{-1} for A horizons, 25–30 cmol$_c$ kg^{-1} for B horizons and 20–25 cmol$_c$ kg^{-1} for C horizons, regardless of the study area. If the clay mineralogy would have been substantially different, the scatter plots between CEC and clay should show different pattern for the two areas. This does not seem to be the case. For example, the scatter plot between CEC.A and Clay.A has a similar slope in both study areas. The same holds for CEC and Clay of the B horizon. CEC and Clay of the C horizon, instead, shows that for the same amount of clay, the values of CEC in the Argentina study area are slightly larger than in the US study area, but there is also a large overlap. Although it is only a visual assessment of the data, and it does not mean that both areas have exactly the same mineralogy, the analysis shows that the CEC of the clay fraction is comparable between study areas. In general, for most soil properties the distributions are fairly similar between the two study areas, with the largest difference between Clay of the C horizon and OC of the A horizon. However, we also found that the correlation coefficients show that correlations between soil properties differ between the two study areas. For instance, in the US study area the correlation between Clay and CEC, and between OC of the A, B and C horizon is generally stronger than in Argentina.

4.2.3. Environmental covariates

Environmental covariates, similar as those used in Argentina Chapter 3, were derived for the US study area from freely accessible data, such as the SRTM DEM (Farr and Kobrick, 2000) and MODIS products. The DEM was used to derive altitude (DEM), terrain wetness index (TWI), and vertical distance to channel network (VDCHN). We also derived the mean and standard deviation from a 15 years' time series of the enhanced vegetation index (EVI) (MOD13Q1) and land-surface temperature and emissivity (LST) (MOD11A2). Finally, we computed the normalised difference of water index (NDWI) (MCD43A4) using the methodology described in Poggio et al. (2013). We computed the mean values for summer (NDWI.A) and spring (NDWI.B) periods. More detailed information about these processing steps are given in Chapter 3. Fig. 4.3 shows the histograms of covariates for both study areas. Covariates VDCHN, TWI and NDWI.A were transformed to obtain a near-normal distribution.

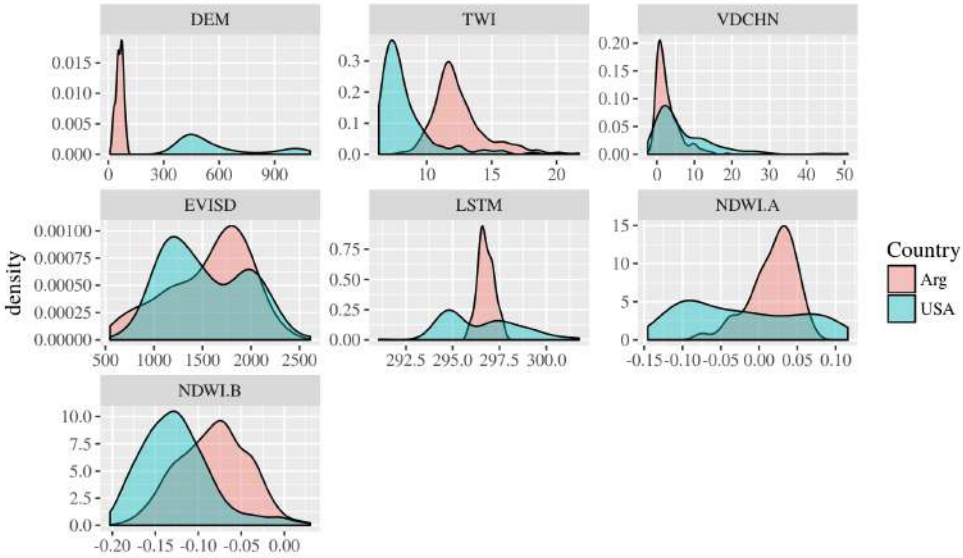


Figure 4.3: *Comparative histograms of covariates for both study areas. DEM: altitude (SRTM); VDCHN: vertical distance to channel network; TWI: terrain wetness index; EVISD: standard deviation of enhanced vegetation index; LSTM: mean of land surface temperature; NDWI.A and NDWI.B: mean of normalised difference wetness index of summer season and spring season.*

Except for VDCHN, the covariates show large differences between both areas. Since all covariates were standardized prior to modelling, by subtracting the mean and dividing by the standard deviation, we do not expect that these differences have a large impact on the models and affect extrapolation. This was confirmed by visual inspection of the histograms of the standardized covariates (not shown).

4.2.4. SEM steps

The use of SEM for DSM has been extensively described in Chapter 2 and 3. We will therefore only give a brief outline here. SEM can be subdivided into seven main steps (Fig. 4.4):

1. *Conceptual model*: this integrates the mechanistic processes that explain the functioning of a system in the form of a piece of text. This step links concepts to variables and explains the main relationships among system variables on the basis of (hypothesised) physical, chemical and biological laws that define the system.

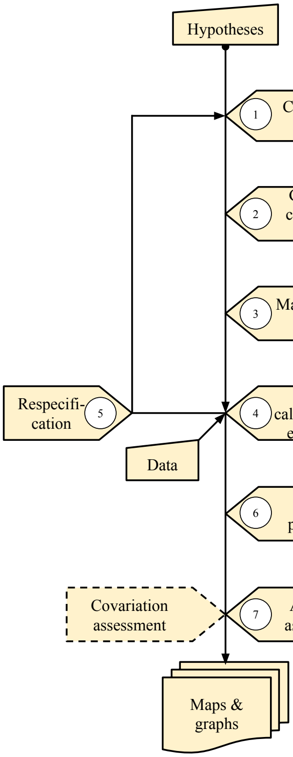


Figure 4.4: *Steps of SEM. Polygons with numbers represent the steps. Arrows represent the work flow.*

2. *Graphical model*: this step involves the representation of interrelationships between system variables in a schematic way, where variables are connected by arrows that indicate either the cause–effect links between variables or correlations in the errors of the variables. The structure of the graphical model is referred to as the model specification, since it specifies how variables influence each other.

3. *Mathematical model*: in this step the graphical model is made mathematically explicit:

$$\mathbf{x} = \boldsymbol{\Lambda}\boldsymbol{\xi} + \boldsymbol{\delta} \tag{4.1}$$

$$\mathbf{y} = \mathbf{K}\boldsymbol{\eta} + \boldsymbol{\varepsilon} \tag{4.2}$$

$$\boldsymbol{\eta} = \mathbf{B}\boldsymbol{\eta} + \boldsymbol{\Gamma}\boldsymbol{\xi} + \boldsymbol{\zeta} \tag{4.3}$$

Here, $\mathbf{x}$ and $\mathbf{y}$ are vectors of observed independent and dependent variables (i.e., external drivers and soil properties), $\boldsymbol{\Lambda}$ and $\mathbf{K}$ are matrices of coefficients that connect measured with latent variables, $\boldsymbol{\xi}$ and $\boldsymbol{\eta}$ are vectors of latent independent and dependent variables, $\boldsymbol{\delta}$ and $\boldsymbol{\varepsilon}$ are vectors of measurement errors (mutually independent and zero-mean normal deviates), $\mathbf{B}$ and $\boldsymbol{\Gamma}$ are matrices of coefficients, and $\boldsymbol{\zeta}$ is a vector of system errors for variable $\boldsymbol{\eta}$. The first two equations (Eqns. 4.1 and 4.2) define what is called in SEM literature the measurement model and Eq. 4.3 is known as the structural model. The mathematical model is defined by three more matrices: $\boldsymbol{\Psi}$ is the variance–covariance matrix of $\boldsymbol{\zeta}$, whose off-diagonal elements represent correlations of system noise of dependent latent variables, $\boldsymbol{\Theta}_{\boldsymbol{\delta}}$ and $\boldsymbol{\Theta}_{\boldsymbol{\varepsilon}}$ are the variance–covariance matrices of $\boldsymbol{\delta}$ and $\boldsymbol{\varepsilon}$, respectively.

4. *Model calibration and evaluation*: a SE model is calibrated by fitting the variance–covariance matrix of the data, $\mathbf{S}$, to the model-implied variance–covariance matrix, $\Sigma(\boldsymbol{\theta})$. The vector $\boldsymbol{\theta}$ contains all model parameters ($\mathbf{B}$, $\boldsymbol{\Gamma}$, $\mathbf{K}$, $\boldsymbol{\Lambda}$, $\boldsymbol{\Psi}$, $\boldsymbol{\Theta}_{\boldsymbol{\delta}}$, and $\boldsymbol{\Theta}_{\boldsymbol{\varepsilon}}$). Parameter estimation is usually done using a maximum likelihood estimator. After the model is calibrated, one has to assess whether the coefficients are coherent in terms of sign and magnitude. It may happen that an expected positive relationship is represented by a negative coefficient or that a coefficient is unusually large because the calibrated model compensates for mistakes in the model specifications (called misspecification). Also, the overall fit of the model is evaluated with different fitting measures to analyse to what degree the relationships established in the model depart from the correlations found in the data.

5. *Model respecification*: If the model that was defined in steps 2 and 3 and calibrated in step 4 has a low overall fitting of coefficients, or has other important weaknesses, it may be adapted in this step (thus, it is respecified). The conceptual model may be incorrectly specified because either the researcher does not have a complete understanding of the system, or the model needs to be simplified. This can be done by checking that the model coefficients make sense from a pedological point of view, or by a so-called "exploratory analysis". The last method uses the differences found between the model-implied relationships and the relationships evidenced by the data to provide suggestions that may improve the performance of the model (Chapter 3).

6. *Spatial Prediction*: the calibrated mathematical model is used to predict the

Table 4.1: *Settings of the different models used for extrapolation.*

	Model type	**Extrapolation**	**Data for respecification**	**Data for calibration**	**Prediction locations**
Model 1	SEM	MM	Arg.	Arg.	USA
Model 2	SEM	CM	Arg.	USA	USA
Model 3	SEM	CM	–	USA	USA
Model 4	SEM	CM	USA	USA	USA
Model 5	MLR	MM	–	Arg.	USA
Model 6	MLR	MM	–	USA	USA
Reference	SEM	MM	Arg.	Arg.	Arg.

Column "Model type" refers to the SE model (SEM) or a multiple linear regression model (MLR). "Extrapolation" refers to which part of the model was extrapolated: either the graphical model (GM) or the mathematical model (MM). The column "Data for respecification" refers to which dataset (Argentinian [Arg.] or US [USA] dataset) was used for the respecification step. A dash (–) means "No respecification". The last two columns detail which datasets were used for calibration and (cross-)validation.

dependent variables ($\mathbf{y}$) from the measured independent variables ($\mathbf{x}$). It is derived from Eqns. 4.1 and 4.3:

$$\hat{\eta} = (\mathbf{I} - \mathbf{B})^{-1}\mathbf{\Gamma}\mathbf{\Lambda}^{-1}\mathbf{x} \tag{4.4}$$

The prediction error variance can also be computed (Section 2.2.5).

7. *Model accuracy and covariation assessment*: the model is assessed in terms of accuracy and system error covariation among predicted soil properties. The mean error (ME; a measure of prediction bias), root mean square error (RMSE; a measure of prediction accuracy) and amount of variance explained by the model (AVE) are generally estimated for each soil property using leave-one-out cross-validation (LOOCV) (Chapter 3). In a conventional application, we would assess covariation using the Standardized Root Mean Square Residual (Hu and Bentler, 1999), but in this work will not evaluate it and focus on model comparison.

4.2.5. Model extrapolation

The aim is to test the SEM capability for model extrapolation and compare it with a purely empirical MLR model for extrapolation mapping CEC, OC content, and clay content of the A, B and C horizons. We do not only consider the mathematical (fitted) model for this purpose, but we will also look at the conceptual model and assess the effect of model respecification on extrapolation. We therefore define seven model settings: four models that are gradually adapted to the US case study (Table 4.1, Model 1 to Model 4), two models that serve as a comparison between SEM and MLR (Table 4.1, Model 5 and Model 6), and one Reference model that was designed and calibrated for the Argentinian case study (Table 4.1, Reference model). The Reference model (from Chapter 3) was included as a benchmark to evaluate the prediction performance of Models 1 to 6. Whenever the models used the same dataset for calibration as for validation, we validated the prediction with cross-validation (LOOCV method).

- *Model 1* is the extrapolation of the SEM mathematical model, including calibrated model parameters, from the Argentinian study area to the US study area. Note that this model was specified, calibrated, evaluated, respecified, and calibrated again, so it was adapted to the Argentinian study area. This model is equal to the Reference model in terms of specifications and coefficients, but here it is used to predict the soil properties of the US study area.

- *Model 2* is the extrapolation of the graphical model (GM) adapted for Argentina. This means that we took the model structure of Model 1, but not the coefficients estimated from the Argentinian data. Instead, we calibrated this model with US data.

- *Model 3* is the extrapolation of the original graphical model developed for Argentinian study area. Thus, we took the model structure that corresponded to the initial Argentinian conceptual model without respecifications. Like Model 2, the model was calibrated with US data.

- *Model 4* is the extrapolation of the Argentinian graphical model that was respecified using US data. Thus, we took the same model structure of Model 3 and adapted (respecified) it to the US case. Like Models 2 and 3, this model was calibrated with US data.

- *Model 5* is the extrapolation of an MLR mathematical model from the Argentinian study area to the US study area. This means that we used a MLR model with coefficients estimated with the Argentinian dataset to predict soil properties across the US study area.

- *Model 6* is a MLR model calibrated with the US dataset.

Several covariates used for the Argentinian case study (Chapter 3) could not be used for the US case study. One of these was "distance to river" that was used as parent material proxy in Argentina. In the US study area, there is no such relationship. This covariate was therefore not used in Models 1 to 6. Furthermore, covariates "latitude" and "longitude" were not used in Models 1 and 5 since the relationship between geographic coordinates and soil properties cannot be extrapolated from Argentina to the US. These two covariates were retained however, in the graphical model of Models 2, 3 and 4, as well as in Model 6.

4.3. Results

4.3.1. Conceptual model

The Argentinian conceptual model was detailed in a previous chapter (Chapter 3) and is only summarised here. In this conceptual model, we focused on the relations between Clay, OC and CEC, and how these are affected by soil-forming factors. It is known that CEC depends on the type and quantity of colloids of the soils, the main sources of which are humus and clay particles. Since CEC increases with decreasing particle size, not only the amount of clay and humus matter, but also the type of humus and the mineralogy of layer silicates in the clay fraction (Brady and Weil, 2014).

The soils of the Argentinian study area have generally 1–3% OC in the A horizon, with a decrease with depth. In terms of Clay, almost all profiles have at least one Bt horizon and a C horizon rich in clay as well (between 10% and 50% clay) (Fig. 4.2). This can be explained by the fact that clay illuviation and organic matter accumulation are the dominant soil forming processes in the area (Imbellone et al., 2010; Morrás and Moretti, 2016). Also, parent material strongly controls clay spatial distribution, since its granulometry decreases in size from southwest to northeast and its mineralogy varies in a different spatial pattern. Clay illuviation and organic matter accumulation are also affected by relief and climate, in addition to the activities of organisms, including humans, through the ages. The soils of the Argentinian study area are relatively homogenous in term of their geologic age, as Zárate (2003) reported that the top sediments (3 to 5 meters) of this region are from the Late Pleistocene or Holocene period.

Soil forming factors cannot be measured directly but can be represented through proxies. Thus, we can use remote sensing products to characterise land cover, such as greenness vegetation indices and land surface temperature. Fig. 4.5 shows how

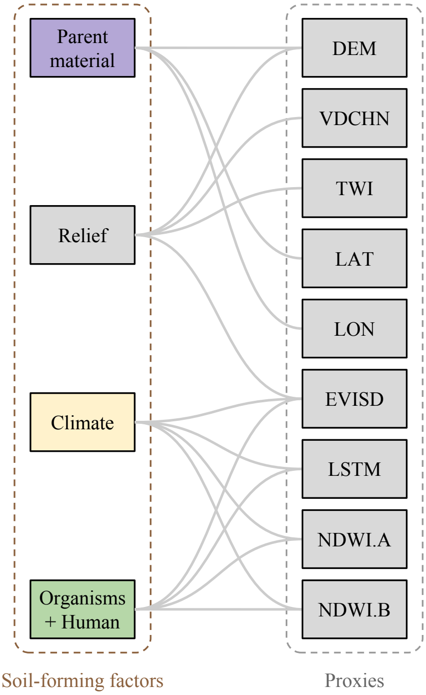


Figure 4.5: *Representation of soil forming factors through proxies. DEM: altitude (SRTM); VDCHN: vertical distance to channel network; TWI: terrain wetness index; LAT (latitude) and LON (longitude) are the X and Y axes of coordinate system; EVISD: standard deviation of enhanced vegetation index; LSTM: mean of land surface temperature; NDWI.A and NDWI.B: mean of normalised difference wetness index of summer season and spring season.*

soil-forming factors in the study area are theoretically represented through proxies. Note that proxies are not exclusive indicators of a single soil forming factor but are often the result of a combination of factors.

We do not expect the same soil evolution in the US study area. However, since the parent materials are similar, the soils also belong to a similar age (Late Pleistocene or Holocene) with similar sequences of soil horizons (Gunal and Ransom, 2006), and the quantities of the three soil properties are comparable (Fig. 4.2), we expect the relations between system variables to be homologous.

4.3.2. Graphical model

The Argentinian conceptual model was converted to a graphical model in Chapter 3. Fig. 4.6 shows the hypothesised relationships between system variables and

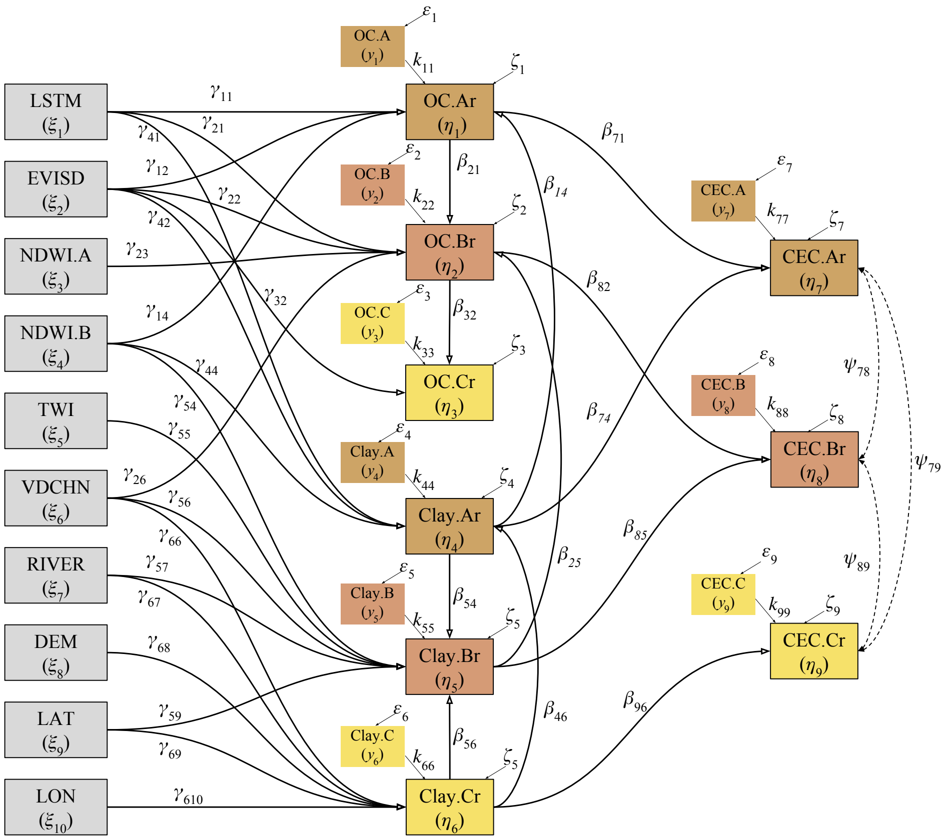


Figure 4.6: *Graphical conceptual model with parameters (from Chapter 3). Thick continuous arrows represent* $\mathbf{B}$ *and* $\mathbf{\Gamma}$ *matrices, thin continuous arrows represent* $\mathbf{K}$, $\mathbf{\Psi}$ *and* $\mathbf{\Theta}_{\varepsilon}$ *matrices, and dashed double-headed arrows represent the model error correlations. Letter "r" at the end of variable names denotes difference between true and measured soil properties (e.g., OC.A is the measured OC of the A horizon, OC.Ar is the true ("real") OC of the A horizon that is unknown).*

the parameters that represent these relationships in the mathematical model. This model was calibrated, evaluated, and respecified based on Argentinian data (Chapter 3), and presented as a Reference model here. The results of the cross-validation (LOOCV) are summarised in Table 4.2 under "Reference". The model explained 91% of OC, 72% of Clay, and 53% of CEC lateral and vertical variation, although the model performance dramatically decreases when considered by horizon. The amount of variance explained (AVE) for B and C horizon Clay were relatively large, thus CEC also had a large AVE in those horizons. The mean error (ME) shows that SEM predictions are unbiased. The RMSE shows that accuracy tends to decrease with depth for CEC and Clay.

4.3.3. Model extrapolation

Model 1: Extrapolation of the mathematical model The accuracy measures for Model 1 show a very poor performance (Table 4.2 and Fig. 4.7). The results show overestimation of CEC values for all horizons, OC.A and Clay.B (negative ME) and underestimation of OC.B, Clay.A and Clay.C (positive ME). This may be the result of differences in extreme values in the original data (Fig. 4.2). The accuracy is low (large RMSE), and the amount of variance explained (AVE) is negative for CEC and OC at the horizon level, meaning that using the horizon means gives more accurate predictions than Model 1 for these properties (Table 4.2). Table 4.2 also shows that the AVE is quite large (0.64) for OC when we evaluate the prediction performance for the three horizons jointly ("Joint" label of Fig. 4.7), indicating that the variation between horizons is much larger than within horizons. However, Fig. 4.7 should be interpreted with care. If we would use the mean OC value for each horizon as derived from US data as a predictor, this would result in an AVE of 0.66. For CEC and Clay, a joint evaluation of the horizons results in low AVE values: 0.08 for CEC and 0.12 for Clay.

Model 2: Extrapolation of the respecified graphical model In case of Model 2 the ME does not show bias in the predictions and the RMSE values are smaller than those of Model 1, except for clay and CEC of the A horizon (Table 4.2 and Fig. 4.7). The AVE values are higher than those of Model 1, with the same exception as for the RMSE. Although improved when compared with Model 1, the model performance remains poor.

Model 3: Extrapolation of the original graphical model Model 3 further improves performance. Again, predictions are unbiased and the prediction accuracy (RMSE) slightly increases with respect to those of the previous models. Also, Model 3 performs similar to Model 6 (see below) in terms of AVE for CEC and Clay, and marginally better for OC.

Model 4: Extrapolation of the graphical model respecified on the basis of US data We took Model 3 as a starting point and respecified it on the basis of expert knowledge and exploratory analysis using the US data. In total we included eight more links. After respecification, the model was again calibrated. The final model

***Table 4.2:** Accuracy measures of the different models.*

		ME	RMSE	AVE	ME	RMSE	AVE	ME	RMSE	AVE	ME	RMSE	AVE
SP	**Hor.**	**Model 1**			**Model 2**			**Model 3**			**Model 4**		
CEC	**A**	−1.42	5.97	−0.03	0.01	6.23	−0.12	0.01	5.73	0.05	0.01	5.56	0.11
	B	−3.18	7.04	−0.13	−0.01	6.33	0.09	0.00	6.12	0.15	0.01	5.87	0.22
	C	−3.12	7.41	−0.04	−0.01	7.01	0.06	0.00	6.63	0.16	0.00	6.39	0.22
OC	**A**	−0.31	0.69	−0.22	0.00	0.59	0.12	0.00	0.58	0.13	0.00	0.56	0.19
	B	0.09	0.26	−0.18	0.00	0.25	−0.09	0.00	0.24	0.00	0.00	0.23	0.01
	C	0.05	0.12	−0.22	0.00	0.11	0.02	0.00	0.11	0.00	0.00	0.11	0.04
Clay	**A**	1.02	8.42	0.08	0.00	8.94	−0.03	0.01	8.26	0.12	0.00	7.92	0.19
	B	−3.84	9.65	0.14	−0.02	9.35	0.19	−0.01	9.00	0.25	0.00	8.68	0.30
	C	4.85	10.71	−0.01	0.00	10.01	0.12	0.00	9.62	0.19	0.01	9.38	0.23
CEC	**Joint**	−2.57	6.84	0.02	0.00	6.53	0.11	0.00	6.17	0.20	0.00	5.95	0.26
OC		−0.05	0.43	0.64	0.00	0.37	0.73	0.00	0.37	0.74	0.00	0.36	0.75
Clay		0.68	9.64	0.18	−0.01	9.45	0.21	0.00	8.97	0.29	0.00	8.68	0.33
		Model 5			**Model 6**			**Reference**					
CEC	**A**	−1.42	6.02	−0.05	0.01	5.57	0.10	0.00	3.16	0.18			
	B	−3.18	7.23	−0.18	−0.04	6.15	0.14	−0.01	3.00	0.50			
	C	−3.12	7.58	−0.09	−0.03	6.63	0.16	−0.01	5.46	0.45			
OC	**A**	−0.31	0.69	−0.23	0.00	0.58	0.14	0.00	0.40	0.24			
	B	0.09	0.25	−0.16	0.00	0.24	−0.04	0.00	0.14	0.03			
	C	0.05	0.12	−0.21	0.00	0.11	−0.05	0.00	0.06	0.02			
Clay	**A**	1.02	8.41	0.09	0.00	7.86	0.20	0.00	4.05	0.15			
	B	−3.84	9.74	0.12	−0.04	8.95	0.26	−0.01	5.14	0.60			
	C	4.85	10.73	−0.01	0.01	9.54	0.20	−0.01	6.87	0.41			
CEC	**Joint**	−2.57	6.98	−0.02	−0.02	6.13	0.21	0.00	4.30	0.53			
OC		−0.05	0.43	0.64	0.00	0.37	0.74	0.00	0.25	0.91			
Clay		0.68	9.67	0.17	−0.01	8.81	0.31	−0.01	5.45	0.72			

Mean error (ME), root mean square error (RMSE) and amount of variance explained (AVE) of models estimated on the basis of leave-one-out cross-validation. Soil properties are abbreviated such that the name of the soil property is followed by the horizon name and separated by a dot.

is shown in Fig. 4.8. Cross-validation results show that Model 4 is the best model in terms of accuracy measures when compared with the others (Table 4.2).

Model 5: Extrapolation of the MLR model Extrapolation of the mathematical MLR model from Argentina to the United States resulted in accuracy measures very similar or slightly worse than those of Model 1. The AVE of Model 5 for CEC and OC of all horizons are negative, and there is bias in the predictions. Judging on the AVE for the joint horizons, this model shows slightly poorer performance in CEC and Clay than Model 1, and the same performance for OC.

Model 6: MLR model based on US data This model was mainly included for comparison of its performance with that of previous models, particularly Models 3 and 4. The ME values show no bias in the predictions and the RMSE values are

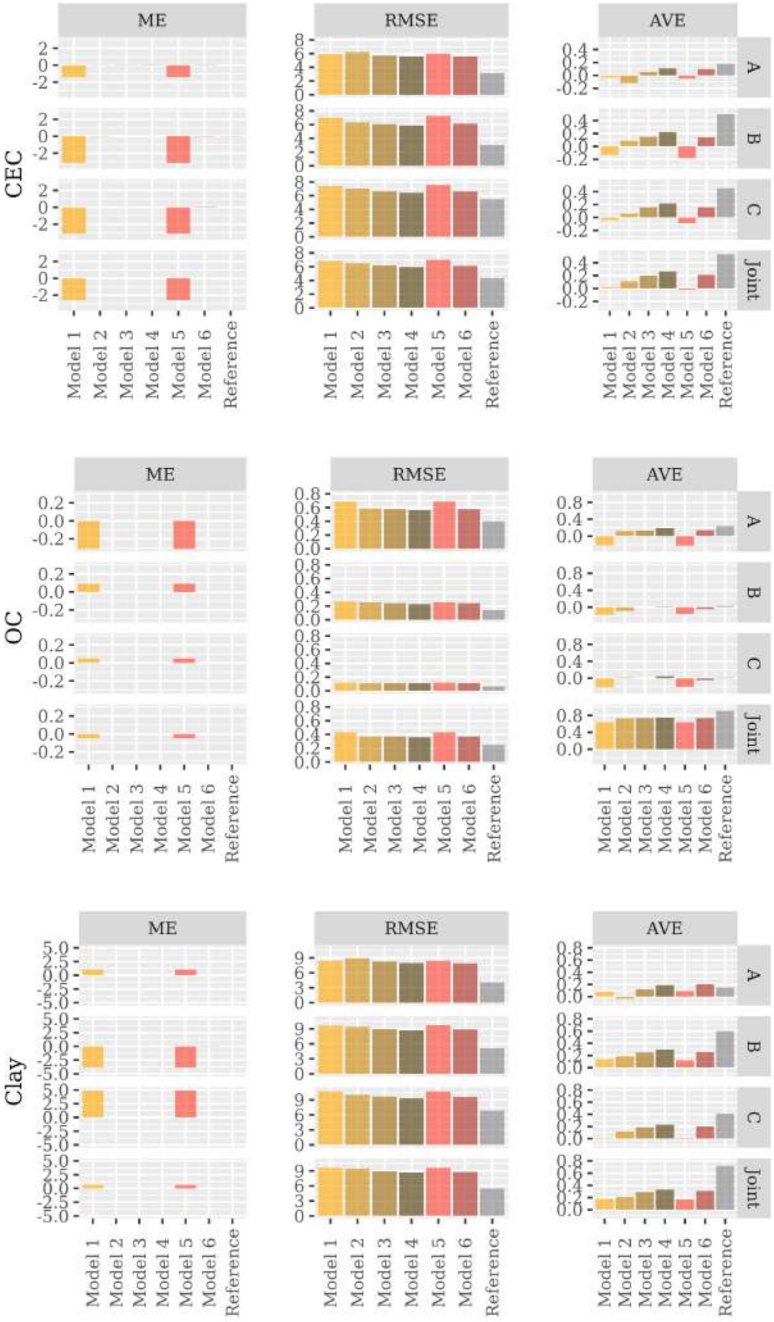


Figure 4.7: *Model performances: comparison of cross-validation mean error (ME), root mean square error (RMSE) and amount of variance explained (AVE) for the seven models by soil properties and horizons. "Joint" represents the measures considering the results at the three horizons together. The measures are computed using the US data for Models 1 to 6 and using the Argentina data for the Reference model.*

comparable to those of Models 3 and 4. The AVE values, instead, show in all cases poorer performance than Model 4 (Table 4.2).

4.4. Discussion

4.4.1. Model performance

To determine if SEM is more accurate in extrapolating the mathematical model than MLR, we compared Models 1 and 5, which were calibrated in Argentina and applied in the United States. The SEM errors were slightly smaller than the MLR errors for CEC for the individual horizons and for the joint horizons, and were almost equal for OC and Clay, where the differences were very small between horizons and the same for the joint horizons. It may be possible that the empirical part of SEM, which is the data-driven coefficient estimation procedure, harms the predictive power of SEM similarly as for MLR.

Both models show poor accuracy, similarly as reported in other extrapolation studies (Grinand et al., 2008; Malone et al., 2016). One reason for that may be the differences in the system variable correlations (Fig. 4.2), despite that the areas have fairly similar soils and parent material. For instance, the correlation between Clay of the B horizon and OC of the A horizon is −0.03 in the Argentinian case and 0.44 in the US case. The same can be observed for some correlations between covariates and soil properties (not shown in Fig. 4.2).

Models 2, 3 and 4 were compared to assess how well the Argentinian graphical model (that represents the conceptual model at different respecification steps) can be extrapolated and how model respecification affects extrapolation capability. Again accuracy measures were generally poor. However, we can still learn a few useful things from the models by comparing the results relative to each other. Models 2 to 4 show that the best model for the Argentinian study area (Model 2, Table 4.1) gives the worst prediction for the US study area, which means that the respecifications done for the Argentinian case result in misspecifications for the US area. If we look at the performance of Model 3 (Table 4.2), where we removed the respecifications (thus making the model more general), there is a clear improvement with respect to Model 2. Model 3 however, performed in between Model 2 and Model 4. The respecification of Model 4 for the US dataset improved the prediction, presumably because it took local conditions into account that were not included in Model 3. Thus, respecifications help to improve local predictions but harm extrapolation capability.

Finally, let us compare the performance of Model 6 with those of Models 2 to 4.

Bearing in mind that Model 3 is the representation of the initial (Argentinian) conceptual model, as it was not subject to respecifications, it is interesting to see that it performed similarly to Model 6. Also, Model 4 was superior to Model 6 in terms of AVE, which confirms the results of the previous chapter (Chapter 3), where SEM also had slightly better performance than MLR.

4.4.2. Comparison of Argentinian and US models

The original graphical model (Chapter 3) was respecified for two different study areas, here represented by the Reference model and Model 4. Both models are presented schematically in Fig. 4.8. This figure allows to analyse how different data change the graphical model through the respecification process. In order to compare these models, we focus first on the differences between links among soil properties, and then on the differences between links among soil properties and covariates.

Some of the differences among soil property interrelationships are the connections from the B horizon Clay to A and B horizon OC in the US case that are absent in Argentina. Instead, there is a link between A horizon Clay and A horizon OC in Argentina that is absent in United States. In other words, the clay content of the B horizon controls the OC content of the A and B horizons in the US case, while it is the clay content of the A horizon that controls the OC of the A horizon in Argentina. Both models show a stronger relationship between Clay and CEC than between OC and CEC, which fits with literature for soils that are clayey like these (Brady and Weil, 2014). Also, B and C horizon OC do not affect CEC because of the negligible amount of organic matter in these horizons (Fig. 4.2). Coincidently, both models present a negative error covariance between C horizon Clay and C horizon OC and between B horizon Clay and A horizon CEC, which means that an overestimation of one soil property occurs simultaneously with an underestimation of the other soil property, and vice versa, which could be related to a change in the type of clay mineralogy or in another soil property, such as pH.

Models differed strongly in the relationships between covariates and soil properties. There are strong relationships present in one case while they are absent in the other. For example, in the Argentinian case (Fig. 4.8) land surface temperature (LSTM) slightly influences Clay of the B horizon, while LSTM is an important covariate to determine the OC of the A horizon and Clay of the C horizon in the US case (Fig. 4.8). Something similar occurs with NDWI (.A and .B) and OC of the A horizon, which in Argentina do not have any effect, while in the United States these are important covariates. Poggio et al. (2013) also found that NDWI was significantly important to predict OC. It seems that the Argentinian case is an exception to this rule. Finally,

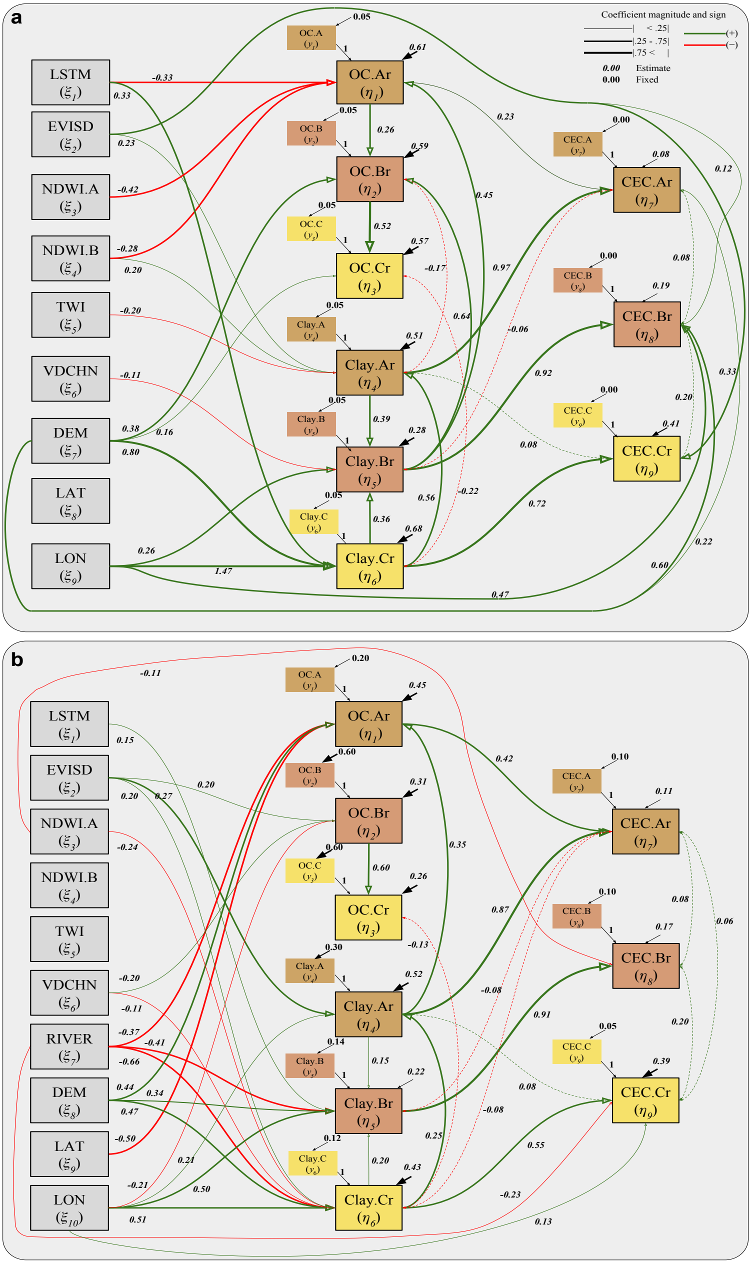


Figure 4.8: *Graphical mathematical models with parameters from the United States case study* **(a)** *and Argentina case study* **(b)**. *Only statistically significant coefficients are shown in both graphs. Continuous red and green arrows represent* $\mathbf{B}$ *and* $\mathbf{\Gamma}$ *matrices, continuous black arrows represent* $\mathbf{K}$, $\mathbf{\Psi}$ *and* $\mathbf{\Theta_\varepsilon}$ *matrices. Dashed double-headed arrows represent model error correlations. Line thickness represents coefficient magnitude, red colours are for negative coefficients and green colours for positive coefficients. Observed soil properties are* y_i *boxes that are named with their property name and horizon, i.e. Clay.A. Latent variables are* η_i *boxes with the name of the observed variable and an "r" for "real", such as Clay.Ar.*

some relations were quite similar in both cases, such as those between DEM and C horizon Clay and between EVI standard deviation and A horizon Clay.

4.4.3. Validity of the conceptual model

Model validity assessment (Bagozzi and Yi, 1988), may become relevant for soil mapping. It consists of calibrating a model with data from different datasets. The main objective in most SEM studies is to analysed whether a model is adequate enough to support the hypotheses that have led to the conceptual model. The smaller the differences between the model for different datasets, the larger the robustness of the model. In our case, this procedure might make the model more generic and less precise (poorer fitting measures) for a particular study area, but more accurate for extrapolation purposes.

We can assess the validity of the SEM conceptual model by comparison of Model 4 and the Reference model, since these were generated from the same hypotheses. Pairs of variables with similar coefficients support the pedological theories behind the coefficients, while large differences between coefficients of the same pair of variables indicate a lack of support. We found that the conceptual model has a theoretically well-supported component, which is to characterise the relationships between soil properties, and a weaker component related to relationships between covariate and soil properties.

The well-supported component is consistent in both models (Fig. 4.8), although there are some differences that can have pedological implications. For example, the US case study shows, as expected, strong relationships between the clay content of the three horizons, but in Argentina these relationships are weak, especially those between the A horizon and the other horizons. This might support the idea that the A horizon is younger than the B and C horizons (Kröhling and Iriondo, 2003). Comparison with the US case strengthens this idea since in the US case study, where all horizons that belong to a second parent material were removed, the soil properties show a stronger correlation between horizons, particularly in the case of Clay. From a pedological point of view, we may think that a higher clay percentage in the B horizon implies a larger water holding capacity for plants and other organisms, which would improve biomass production, hence increase OC in the A horizon. However, this is only true when water is a limiting resource for organisms, such as in the US study area. If weather conditions are more humid, like in the Argentinian case, an increase in clay percentage in the B horizon may not affect organic matter production in the A horizon.

When looking at the relationships between covariates and soil properties, the weakly-sup-ported component of both models (Fig. 4.8), latitude, longitude and altitude have a large influence on the soil properties in both models, particularly on the clay percentage, which may be explained by the spatial distribution of the parent material. Also, mean LST, standard deviation of EVI and NDWI moderately contribute to explain soil property variance, but the mechanisms behind these links are harder to explain than in the other relations, since they are partial proxies of different soil-forming factors.

Model 4 shows poorer performance than the Reference model (Table 4.2). This might be due to the large difference in sampling density and area size between both study areas. The relationships between covariates and soil properties may be different in large and heterogeneous areas, since it may include different environmental conditions. In this regard, some US covariates tend to have a bimodal distribution (Fig. 4.3), while the distributions of the soil properties are unimodal. These bimodal distributions, therefore, can mean that there are two distinct environments within the area.

4.4.4. Challenges of extrapolating soil mapping models

Extrapolation of soil mapping models faces a range of difficulties. For example, Grinand et al. (2008) found that the predictive power of their model remained low when applied to an extrapolation area, but was much higher when applied to a validation dataset from the same calibration area. They associated this finding to the fact that validation data were spatially autocorrelated with calibration data, which increased its accuracy measures, and so the lack of spatial correlation between calibration and extrapolation locations decreased the same accuracy measures. On the other hand, Malone et al. (2016) extrapolated a model from one area to another area located in the same region. They tested the similarity between the areas on the basis of the available covariates and found that only about half of the area was similar. Then, they found that the predictive power of a model depended on the similarity between calibration and extrapolation study areas.

In this study, we found that the predictive power of the model depends largely on the similarity of system variable relations. Dissimilarity of covariates might cause differences in the system variable relationships, but not necessarily. The main factor of model extrapolation failure is ultimately the mismatch of the model structure and/or model coefficients, which are defined on the basis of those system variable interrelationships. In this study, we found that some relationships between the same pair of variables were opposite in sign between calibration and extrapolation areas. This, of course, damages the validity of the coefficients and consequently, the performance of the whole model.

4.5. Conclusions

We tested the extrapolation potential of a SE model by applying a SE model developed for an Argentinian study area to a study area in the Great Plains of the United States for prediction of three soil properties of three major horizons. We compared the performances of several SE models, which differed in the degree in which we allowed data from the extrapolation area to change the structure and coefficients of the SE model. We also added MLR models to the comparison. The main conclusions of this study are:

- An extrapolated conceptual model from Argentina calibrated with US data performs slightly better than a MLR model calibrated with US data.
- System relationships that were well supported by pedological knowledge, such as soil property relationships, showed consistent and equal behaviour in both study areas. Other relationships, such as between covariates and soil properties, differed much more between models. An explanation for this can be that the covariates used here were poor proxies of the true soil-forming factors and that covariates of the two regions were not exactly the same.
- A deeper understanding of the real soil-forming factors and the soil evolution might lead to create better covariates to strengthen conceptual models for DSM.
- Knowledge-based model specifications are more effective than data-driven respecifications for extrapolation of the SEM conceptual model.
- Respecifications (model modifications) can improve local prediction but will usually harm the extrapolation capability of a model.
- While extrapolation of a calibrated model may often be discouraged because of poor prediction accuracy, extrapolation of the conceptual model through the SEM graphical model is a viable alternative. In this way, pedological knowledge about the calibration area can be combined with soil and covariate data from the extrapolation area.

Chapter 5

Including spatial correlation in structural equation modelling of soil properties

Digital soil mapping techniques usually take an entirely data-driven approach and model soil properties individually and layer by layer, without consideration of interactions. In previous studies we implemented a structural equation modelling (SEM) approach to include pedological knowledge and between-properties and between-layer interactions in the mapping process. However, as SEM is commonly applied in the social sciences and econometrics, it typically does not consider spatial correlation. Therefore, the goal of this chapter was to extend SEM by accounting for residual spatial correlation using a geostatistical approach. We assumed second-order stationarity and estimated the semivariogram parameters, together with the usual SEM parameters, using maximum likelihood estimation. Next, spatial prediction was done using regression kriging. We summarise the mathematics of both SEM and the geostatistical model, as well as the process to combine them. The methodology is applied to mapping cation exchange capacity, clay content and soil organic carbon for three soil horizons in a 150 100 km^2 study area in the Great Plains of the United States. The calibration process included all parameters used in `lavaan`*, a software implementation of SEM, plus two extra parameters to model residual spatial correlation. The residuals showed substantial spatial correlation, which indicates that including spatial correlation yields more accurate predictions. This was confirmed by cross-validation. We also compared the standard SEM and the spatial SEM approaches in terms of SEM model coefficients. Differences were significant but none of the coefficients changed sign. Presence of residual spatial correlation suggests that some of the causal factors that explain soil variation were not captured by the set of covariates. In such case it is worthwhile to search for additional covariates leaving only unstructured residual noise, but as long as this is not achieved, it pays off to include residual spatial correlation in mapping using SEM.*

Based on:
Angelini, M. E. and Heuvelink, G. B. M., 2017. Including spatial correlation in structural equation modelling of soil properties. *Spatial Statistics*. Submitted.

5.1. Introduction

Many national and international programs require accurate and detailed soil information, such as the Sustainable Development Goals of the United Nations (defined in 2014) and the *4 per 1000* initiative (Conference of the Parties to the United Nations Framework Convention on Climate Change in Paris in 2015). It is now common to produce spatially explicit soil information at global and national scales with digital soil mapping techniques (Hengl et al., 2017; Minasny et al., 2017).

Digital soil mapping (DSM) makes use of field and laboratory soil data, environmental covariates and a statistical model to predict soil properties or soil type at unmeasured locations. Most DSM studies take an entirely data-driven approach and model the soil spatial variation layer by layer and for individual soil properties separately. Mechanistic soil process knowledge and interactions between layers and soil properties is often ignored.

In previous studies we have investigated the potential use of structural equation modelling (SEM) (Chapter 2, Chapter 3 and Chapter 4) for DSM. The SEM methodology consists of a system of equations derived from a pedological conceptual model, which defines and supports interrelations in the soil-landscape system. The system of equations that forms the core of SEM was calibrated using empirical data from a study area in the Argentine Pampas. We analysed to what degree the empirical data agreed with the conceptual model and used the calibrated SEM for simultaneous prediction of multiple soil properties at multiple layers.

The strength of SEM is that it blends process-driven and data-driven approaches, by using mechanistic principles to define the model structure and empirical principles to calibrate and refine the model. However, SEM typically does not account for spatial correlation, other than through the spatial structure of the covariates. Spatial SEM applications could benefit from taking spatial correlation into account explicitly. Lamb et al. (2014) removed spatial autocorrelation from the system variables by analysing model correlations at different lag distances. Wall (2012) extended SEM by assuming that the system variables are spatially dependent, through (cross-) correlation of the stochastic residuals. They took a spatial lattice approach and combined SEM with a conditional autoregressive model to represent behavioural risk factor surveillance survey data.

In this study we extend SEM by accounting for residual spatial correlation using a geostatistical approach. We take a geostatistical approach because this better meets the characteristics of soil data, that typically vary continuously in space. We assume second-order stationarity and estimate the semivariogram parameters, simultaneously with the usual SEM parameters, using a maximum likelihood approach. Next, spatial prediction is done using regression kriging, where the trend part is defined by the SEM structural equation. The methodology is applied to mapping cation exchange capacity, clay content and soil organic carbon for three soil horizons in a 150 100 km^2 study area in the US Great Plains.

5.2. Materials and methods

5.2.1. Structural equation model

SEM methodology has been described in various text books, such as Jöreskog and Sörbom (1981); Bollen (1989). We have used SEM to predict spatially several soil properties (Chapters 2, 3, and 4). Here, we first summarise the main equations of SEM, including parameter estimation and prediction, and next extend the methodology to include spatial correlation. The components of the various equations presented in this section are described in Table 5.1.

Model definition

The system of equations of SEM is characterised by the structural model and the measurement model. The first is defined as follows:

$$\begin{gathered}\boldsymbol{\eta} = \mathbf{B}\boldsymbol{\eta} + \boldsymbol{\Gamma}\boldsymbol{\xi} + \boldsymbol{\zeta} \\ E[\boldsymbol{\xi}] = E[\boldsymbol{\zeta}] = 0 \\ Var(\boldsymbol{\xi}) = \boldsymbol{\Phi}, Var(\boldsymbol{\zeta}) = \boldsymbol{\Psi}\end{gathered} \tag{5.1}$$

where $\boldsymbol{\eta}$ is a vector of latent endogenous variables, $\boldsymbol{\xi}$ a vector of latent exogenous variables and $\boldsymbol{\zeta}$ a vector of structural errors. The diagonal elements of $\mathbf{B}$ are zero, $\boldsymbol{\xi}$ and $\boldsymbol{\zeta}$ are mutually independent and normally distributed. Their variance–covariance matrices are given by $\boldsymbol{\Phi}$ and $\boldsymbol{\Psi}$, respectively.

The measurement model is given by:

$$\begin{gathered}\mathbf{Y} = \mathbf{K}\boldsymbol{\eta} + \boldsymbol{\varepsilon} \\ \mathbf{X} = \boldsymbol{\Lambda}\boldsymbol{\xi} + \boldsymbol{\delta} \\ E[\boldsymbol{\varepsilon}] = E[\boldsymbol{\delta}] = 0 \\ Var(\boldsymbol{\varepsilon}) = \boldsymbol{\Theta}_{\varepsilon}, \quad Var(\boldsymbol{\delta}) = \boldsymbol{\Theta}_{\delta}\end{gathered} \tag{5.2}$$

Table 5.1: *SEM components, adapted from Bollen (1989)*

Symbol	Dimension	Definition
		Structural model
$\boldsymbol{\eta}$	$n \times 1$	Latent endogenous variable
$\boldsymbol{\xi}$	$m \times 1$	Latent exogenous variable
$\boldsymbol{\zeta}$	$n \times 1$	System error
$\mathbf{B}$	$n \times n$	Coefficients for latent endogenous variable
$\boldsymbol{\Gamma}$	$n \times m$	Coefficients for latent exogenous variables
$\boldsymbol{\Phi}$	$m \times m$	Variance–covariance matrix of $\boldsymbol{\xi}$
$\boldsymbol{\Psi}$	$n \times n$	Variance–covariance matrix of $\boldsymbol{\zeta}$
		Measurement model
$\mathbf{Y}$	$p \times 1$	Measured variable of $\boldsymbol{\eta}$
$\mathbf{X}$	$q \times 1$	Measured variable of $\boldsymbol{\xi}$
$\mathbf{z}$	$(p+q) \times 1$	Values of $\mathbf{y}$ and $\mathbf{x}$ variables
$\boldsymbol{\varepsilon}$	$p \times 1$	Measurement error of $\mathbf{y}$
$\boldsymbol{\delta}$	$q \times 1$	Measurement error of $\mathbf{x}$
$\mathbf{K}$	$q \times m$	Coefficients between $\mathbf{y}$ and $\boldsymbol{\eta}$
$\boldsymbol{\Lambda}$	$p \times n$	Coefficients between $\mathbf{x}$ and $\boldsymbol{\xi}$
$\boldsymbol{\Theta}_{\varepsilon}$	$p \times p$	Variance–covariance matrix of $\boldsymbol{\varepsilon}$
$\boldsymbol{\Theta}_{\delta}$	$q \times q$	Variance–covariance matrix of $\boldsymbol{\delta}$
		Spatial model
$\mathbf{s}, \mathbf{s_k}$	2×1	Measurement locations ($k = 1 \ldots N$)
a	scalar	Range
h	scalar	Euclidean distance
α	scalar	Nugget-to-sill ratio
$\mathbf{c}(h)$	scalar	Correlation function value at distance h

where $\boldsymbol{\varepsilon}$ and $\boldsymbol{\delta}$ are mutually independent normally distributed variables that are independent of all previously defined variables. In what follows, we will assume that $\mathbf{K}$ and $\boldsymbol{\Lambda}$ are identity matrices (hence $p = n$ and $q = m$). Note also that all variables are assumed to have zero mean. In practice, this is accommodated by standardising the observations of individual measurement variables prior to modelling.

Parameter estimation

The variance–covariance matrix of the vector of measurement variables $\mathbf{Z} = [\mathbf{Y}^T\mathbf{X}^T]^T$ follows from the SEM system of equations. It is given by:

$$Var(\mathbf{Z}) = \Sigma(\boldsymbol{\theta}) = \begin{bmatrix} (\mathbf{I}-\mathbf{B})^{-1}(\boldsymbol{\Gamma\Phi\Gamma}^T + \boldsymbol{\Psi})((\mathbf{I}-\mathbf{B})^{-1})^T + \boldsymbol{\Theta}_{\boldsymbol{\varepsilon}} & (\mathbf{I}-\mathbf{B})^{-1}\boldsymbol{\Gamma\Phi} \\ \boldsymbol{\Phi\Gamma}^T((\mathbf{I}-\mathbf{B})^{-1})^T & \boldsymbol{\Phi} + \boldsymbol{\Theta}_{\boldsymbol{\delta}} \end{bmatrix} \tag{5.3}$$

Here, $\boldsymbol{\theta}$ represents all parameters contained in $\mathbf{B}$, $\boldsymbol{\Gamma}$, $\boldsymbol{\Phi}$, $\boldsymbol{\Psi}$, $\boldsymbol{\Theta}_{\boldsymbol{\varepsilon}}$ and $\boldsymbol{\Theta}_{\boldsymbol{\delta}}$. The parameters $\boldsymbol{\theta}$ are most commonly derived using maximum likelihood estimation. In practice, some of the elements of $\boldsymbol{\theta}$ will not be estimated but assumed known. For instance, many elements of $\mathbf{B}$ and $\boldsymbol{\Phi}$ will be assumed zero because not all exogenous and endogenous variables have a direct effect on (other) endogenous variables. In what follows, we will further assume that $\boldsymbol{\Theta}_{\boldsymbol{\delta}} = 0$ and that $\boldsymbol{\Theta}_{\boldsymbol{\varepsilon}}$ is a known diagonal matrix. We will represent $\boldsymbol{\Phi}$ with the empirical variance–covariance matrix derived from the N observation vectors $\mathbf{x_k}, k = 1 \ldots N$.

We summarise the maximum likelihood estimation procedure explained in Bollen (1989, Appendix 4A). Since we assume multivariate normality and all measured variables have zero mean, the probability density $f\left(\mathbf{z}; \Sigma(\boldsymbol{\theta})\right)$ of $\mathbf{Z}$ is given by:

$$f\left(\mathbf{z}; \Sigma(\boldsymbol{\theta})\right) = (2\pi)^{-(p+q)/2}|\Sigma(\boldsymbol{\theta})|^{-1/2} \exp\left[-\tfrac{1}{2}\mathbf{z}^T\Sigma(\boldsymbol{\theta})^{-1}\mathbf{z}\right] \tag{5.4}$$

In conventional SEM we assume that the N observation vectors $\mathbf{z}_k (k = 1 \ldots N)$ are realisations of independent random vectors $\mathbf{Z}$. Because of independence, their joint density is the product of the marginal densities:

$$\begin{gathered} f(\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_N; \Sigma(\boldsymbol{\theta})) = \\ f(\mathbf{z}_1; \Sigma(\boldsymbol{\theta})) \cdot f(\mathbf{z}_2; \Sigma(\boldsymbol{\theta})) \ldots f(\mathbf{z}_N; \Sigma(\boldsymbol{\theta})) = \\ (2\pi)^{-N(p+q)/2}|\Sigma(\boldsymbol{\theta})|^{-N/2} \exp\left[-\tfrac{1}{2}\sum_{k=1}^{N}\mathbf{z}_k^T\Sigma(\boldsymbol{\theta})^{-1}\mathbf{z}_k\right] \end{gathered} \tag{5.5}$$

For parameter estimation we treat this as a function of the parameters $\boldsymbol{\theta}$ and minimise the negative log-likelihood:

$$-log\ L(\boldsymbol{\theta}) = \frac{N(p+q)}{2} log(2\pi) + \frac{N}{2} log(|\Sigma(\boldsymbol{\theta})|) + \frac{1}{2}\sum_{k=1}^{N}\mathbf{z}_k^T\Sigma(\boldsymbol{\theta})^{-1}\mathbf{z}_k \tag{5.6}$$

Prediction

Prediction of the latent endogenous variable $\boldsymbol{\eta}$ is not very common in the SEM literature because SEM is mostly applied in the social sciences, where parameter estimation and interpretation are the main objectives. However, in a soil mapping context prediction is important because the ultimate aim is to make a soil map. Prediction of the endogenous variables from observations of the exogenous variables is easily achieved as (Section 2.2.5):

$$\hat{\boldsymbol{\eta}} = (\mathbf{I} - \mathbf{B})^{-1}\boldsymbol{\Gamma}\mathbf{x} \tag{5.7}$$

with prediction error variance given by:

$$Var(\boldsymbol{\eta} - \hat{\boldsymbol{\eta}}) = (\mathbf{I} - \mathbf{B})^{-1}\boldsymbol{\Psi}((\mathbf{I} - \mathbf{B})^{-1})^T \tag{5.8}$$

Note that these results make use of the assumption that $\boldsymbol{\Lambda} = I$ and $\boldsymbol{\Theta}_{\boldsymbol{\delta}} = 0$, but that it would not be difficult to generalise these to a case where these assumptions are not made. Note also that Eq. 5.8 assumes that the model parameters are known and hence it does not include uncertainty caused by parameter estimation error.

5.2.2. Generalisation to the spatially correlated case

We now drop the assumption that vectors $\mathbf{Z}_k = \mathbf{Z}(\mathbf{s}_k), k = 1 \ldots N$ are statistically independent (note the change of notation by making explicit that observations are taken at geographic locations $\mathbf{s}_k$). This is not a realistic assumption in a spatial setting, where variables $\mathbf{Z}(\mathbf{s}_k)$ and $\mathbf{Z}(\mathbf{s}_l)$ are likely to be correlated when the distance between locations $\mathbf{s}_k$ and $\mathbf{s}_l$ is relatively small. We extend the basic model Eq. 5.1 by allowing $\boldsymbol{\xi}$ and $\boldsymbol{\zeta}$ to be spatially correlated. We model the covariance between $\mathbf{Z}(\mathbf{s}_k)$ and $\mathbf{Z}(\mathbf{s}_l)$ as:

$$Cov(\mathbf{Z}(\mathbf{s}_k), \mathbf{Z}(\mathbf{s}_l)) = \boldsymbol{\Sigma}(\boldsymbol{\theta})\; c(|\mathbf{s}_k - \mathbf{s}_l|) \tag{5.9}$$

where the correlation function c is given by:

$$c(h) = \begin{cases} 1 & \text{if } h = 0 \\ (1 - \alpha)\exp\left(-\frac{h}{a}\right) & \text{if } h > 0 \end{cases} \tag{5.10}$$

where α is the nugget-to-sill ratio and a the “range” parameter (i.e., a measure of the spatial correlation length). Note that we assumed an isotropic, stationarity exponential correlation model that has the same parameter values for all endogenous and exogenous variables. More flexibility would be offered if the parameters were allowed to be variable-specific (see Discussion Section).

Parameter estimation

We use maximum likelihood estimation as before. However, since the N observation vectors are no longer independent, we can no longer use Eq. 5.5. Instead, the joint density of the $\mathbf{Z}_k, k = 1 \dots N$ is given by:

$$\begin{gathered} f(\mathbf{z}(\mathbf{s}_1), \mathbf{z}(\mathbf{s}_2), \dots, \mathbf{z}(\mathbf{s}_N); \Sigma(\boldsymbol{\theta}), \alpha, a) = \\ (2\pi)^{-N(p+q)/2} |\Sigma_{all}(\boldsymbol{\theta}, \alpha, a)|^{-1/2} \exp\left[-\tfrac{1}{2}\mathbf{z}_{all}^T \Sigma_{all}(\boldsymbol{\theta}, \alpha, a)^{-1} \mathbf{z}_{all}\right] \end{gathered} \tag{5.11}$$

where the $N \cdot (p+q)$ vector $\mathbf{z}_{all}$ is a concatenation of all $\mathbf{z}(\mathbf{s}_k), k = 1 \dots N$ and where the $N \cdot (p+q) \times N \cdot (p+q)$ matrix $\Sigma_{all}(\boldsymbol{\theta}, \alpha, a)$ is given by a Kronecker product:

$$\Sigma_{all}(\boldsymbol{\theta}, \alpha, a) = \Sigma(\boldsymbol{\theta}) \otimes \mathbf{C}(\alpha, a) \tag{5.12}$$

The $N \times N$ matrix $\mathbf{C}(\alpha, a)$ contains the correlations at distances between observation locations:

$$\mathbf{C}(\alpha, a)[k, l] = c(|\mathbf{s}_k - \mathbf{s}_l|) \qquad k, l = 1 \dots N \tag{5.13}$$

The corresponding negative log-likelihood now becomes:

$$-log\, L(\boldsymbol{\theta}) = \frac{N(p+q)}{2} log(2\pi) + \frac{1}{2} log(|\Sigma_{all}(\boldsymbol{\theta}, \alpha, a)|) + \frac{1}{2}\mathbf{z}_{all}^T \Sigma_{all}(\boldsymbol{\theta}, \alpha, a)^{-1} \mathbf{z}_{all} \tag{5.14}$$

This can be minimised using numerical search algorithms (see Section 5.2.3). Computations can take much time because each iteration of the numerical search algorithm involves evaluation of the inverse and determinant of Σ_{all}. This can be a large matrix (in the case study discussed in Section 5.2.3 it is $2,754 \times 2,754$), but computation of its determinant and inverse can be speeded up dramatically by making use of properties of the Kronecker product (Steeb, 2012).

Prediction

Prediction of the endogenous variable $\boldsymbol{\eta}$ starts with Eq. 5.7, but since the prediction error $\boldsymbol{\eta} - \hat{\boldsymbol{\eta}}$ is spatially correlated, mapping accuracy can be improved by predicting $\boldsymbol{\eta} - \hat{\boldsymbol{\eta}}$ at all non-observation locations using the "observed" prediction errors $\mathbf{y}_k - \hat{\boldsymbol{\eta}}_k, k = 1 \ldots N$. The interpolated residuals can then be added to the "trend" obtained with Eq. 5.7. This approach boils down to regression kriging (Hengl et al., 2004). In this case, where we have multiple variables that are jointly normal and whose means are known (i.e., zero), the residuals are predicted using simple cokriging, which boils down to computing the conditional normal distribution.

For two jointly normal vectors U and V, the distribution of U given $V = v$ is given by (Hogg and Craig, 1995, Chapter 2):

$$\{U|V = v\} \sim N\left(\mathrm{C}_{UV}\mathrm{C}_{VV}^{-1}v, \mathrm{C}_{UU} - \mathrm{C}_{UV}\mathrm{C}_{VV}^{-1}\mathrm{C}_{VU}\right) \tag{5.15}$$

We make use of this by substituting the n-vector $\boldsymbol{\eta}(\mathbf{s}_0) - \hat{\boldsymbol{\eta}}(\mathbf{s}_0)$ for U and the $n \cdot N$ concatenation vector $\{\mathrm{Y}(\mathbf{s}_k) - \hat{\boldsymbol{\eta}}(\mathbf{s}_k), k = 1 \ldots N\}$ for V. The elements of the variance–covariance matrices $\mathrm{C}_{UU}, \mathrm{C}_{VV}, \mathrm{C}_{UV}$ and C_{VU} are derived using Eqns. 5.8 and 5.9:

$$var(\boldsymbol{\eta}(\mathbf{s}_0) - \hat{\boldsymbol{\eta}}(\mathbf{s}_0) = (\mathbf{I} - \mathbf{B})^{-1}\boldsymbol{\Psi}((\mathbf{I} - \mathbf{B})^{-1})^T \tag{5.16}$$

$$\begin{aligned} var(\mathrm{Y}(\mathbf{s}_k) - \hat{\boldsymbol{\eta}}(\mathbf{s}_k)) &= var(\boldsymbol{\eta}(\mathbf{s}_k) + \boldsymbol{\varepsilon}(\mathbf{s}_k) - \hat{\boldsymbol{\eta}}(\mathbf{s}_k)) \\ &= (\mathbf{I} - \mathbf{B})^{-1}\boldsymbol{\Psi}((\mathbf{I} - \mathbf{B})^{-1})^T + \boldsymbol{\Theta}_{\boldsymbol{\varepsilon}}, \quad k = 1 \ldots N \end{aligned} \tag{5.17}$$

$$cov(\mathrm{Y}(\mathbf{s}_k) - \hat{\boldsymbol{\eta}}(\mathbf{s}_k), \mathrm{Y}(\mathbf{s}_l) - \hat{\boldsymbol{\eta}}(\mathbf{s}_l)) = ((\mathbf{I} - \mathbf{B})^{-1}\boldsymbol{\Psi}((\mathbf{I} - \mathbf{B})^{-1})^T + \boldsymbol{\Theta}_{\boldsymbol{\varepsilon}}) \cdot c(|\mathbf{s}_k - \mathbf{s}_l|), \quad k, l = 1 \ldots N \tag{5.18}$$

Thus, prediction maps of all soil properties are obtained by adding the conditional mean as given in Eq. 5.15 to the trend obtained with Eq. 5.7, while prediction error variance maps are derived from the conditional variance given in Eq. 5.15.

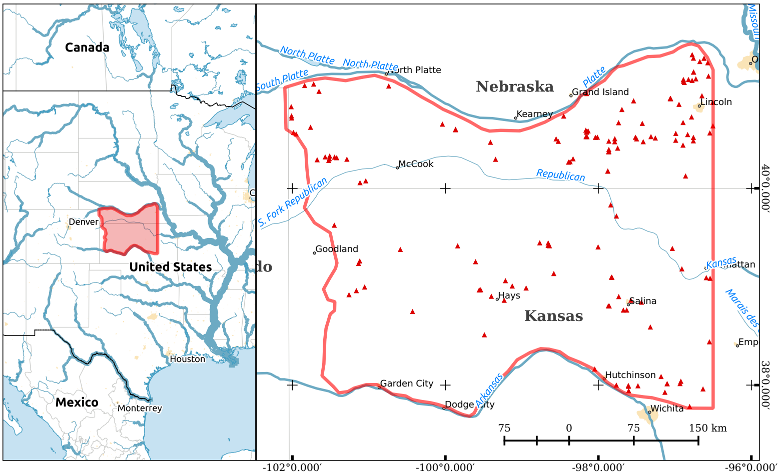


Figure 5.1: *Study area and boundary (red line) with locations of soil profiles (red triangles).*

5.2.3. Case study

We used the spatial SEM approach for modelling the interrelations among soil properties and environmental variables ("covariates") and predict the soil properties from soil observations at sampling locations and covariate maps. Soil observations are generally collected from soil profiles. Each soil profile is divided into layers, generally named horizons, defined by their depth and thickness, which can vary between profiles. At each sampling location, soil properties are measured at each horizon. Covariate maps are available as raster maps, exhaustively distributed over the whole study area.

Study area

The study area is located in the USA Great Plains, between latitudes 37°42' N and 41° 30' N and longitudes 96° 30' W and 102° 06' W. The area is about 150 100 km^2 in size and covers parts of the states Nebraska and Kansas (Fig. 5.1). The main parent materials of this region are limestone and shale, covered by aeolian sediments (loess). Although the area was originally grassland, much of it has been changed into cropland. The remaining grassland is used for grazing cattle. The annual precipitation ranges from 500 mm in the west to 800 mm in the east.

Soil profile data were taken from the SSURGO2 database (Soil Survey Staff, 2016). Three soil properties were selected for modelling and mapping: soil organic carbon

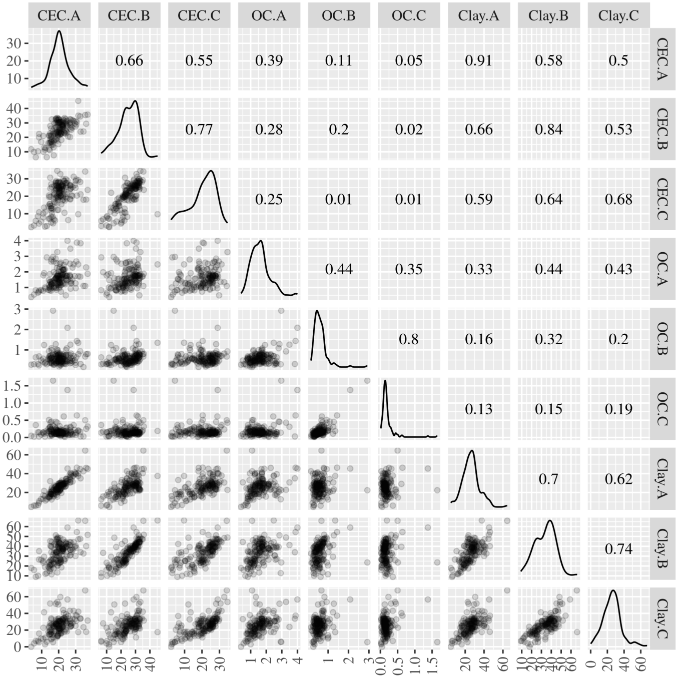


__Figure 5.2:__ Correlation graphs of soil properties by horizon. The upper right triangle shows Pearson correlation coefficients between properties, the diagonal presents histograms and the lower left triangle scatter plots. Soil properties are abbreviated such that the name of the soil property is followed by the horizon name and separated by a dot. 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.

(OC, in mass percentage), clay content (in mass percentage), and cation exchange capacity (CEC, in cmol$_c$ kg^{-1} soil) for three major horizons: A, B and C. Initially, 492 soil profiles were fetched, but only those profiles that fulfilled the following criteria were selected: (1) it must have an A, B and C horizon; (2) it must not have missing values for CEC, OC and Clay; (3) the horizons should not belong to a buried soil profile that might be indicative of a parent material discontinuation; and (4) multiple profiles must not share the same spatial coordinates. We grouped all subdivisions of A, B and C horizons, such as Ap, A1, A2, Bt, Bw, etc. to the master horizon level, and excluded transitional horizons, such as AB, BA, AC, BC, etc. The total number of retained soil profiles was 147 (Fig. 5.1). Fig. 5.2 shows a correlation graphs of the soil properties by horizon.

Environmental covariates were derived from freely accessible and globally available remote sensing products and processed using the same methodology as in Chapter 3. The source data are the SRTM DEM (Farr and Kobrick, 2000) and products from the moderate-resolution imaging spectroradiometer (MODIS). From the SRTM DEM we derived altitude (DEM), terrain wetness index (TWI) and vertical distance to channel network (VDCHN). From a 15 years time series of MODIS images, we computed the standard deviation of the enhanced vegetation index (EVISD) and the mean land-surface temperature and emissivity (LSTM). Finally, we computed time series of the normalised difference of water index, using the procedure described in Poggio et al. (2013). From this series, we computed the mean values from 4 July to 20 August (NDWI.A) and the mean from 7 April to 1 May (NDWI.B). Covariates VDCHN, TWI were log-transformed and NDWI.A was transformed to the cubic root to obtain a sufficiently symmetric distribution. As in a previous chapter (e.g. Chapter 4), we also included the geographical latitude (LAT) and longitude (LON) as covariates, since these might indicate regional differences in parent material.

Model specifications and accuracy assessment

In order to assess the added value of the new approach, we compared the performance of a non-spatial, standard SE model fitted in a previous study for this region (Chapter 4) with the performance of the spatial SE model. The standard SE model takes the interrelations between soil properties and between soil properties and covariates into account. Details of the conceptual model and model specification are given in Chapter 4. The standard SEM and the spatial SEM did not differ in their specification, which means that both have the same SEM system parameters (although the optimised parameter values may differ, as we will see below), but the spatial SEM has two additional parameters, namely α and a. Standard SEM was specified and calibrated with the `lavaan` package (Rosseel, 2012). Resulting parameter estimates and two coefficient equal to 0.5 for α and a were set as starting values for calibration of the spatial SEM, which was calibrated under the R (R Core Team, 2017) environment. We used the *PORT* routines (*nlminb* function of the `stats` package) to minimise the negative log-likelihood. Standard errors were estimated using the `lavaan` approach.

To assess the performance of both models, we computed the mean error (ME), the root mean square error (RMSE) and the amount of variance explained (AVE) of the predicted soil properties. We computed these statistics using leave-one-out cross-validation, in the same way as in Chapter 3. Thus, soil data from all N sampling locations were put aside one by one, each time using the remaining data to calibrate the models and predict at the location that was left out.

5.3. Results and discussion

5.3.1. Graphical model

The nugget-to-sill ratio α of the calibrated spatial SE model was 0.41. The range parameter a was estimated as 47.2 km, which means that the effective range of spatial correlation is about 142 km, which is considerable, given the extent of the study area (about 350 $km \times 400\ km$). The large range and modest nugget-to-sill ratio together indicate that there is substantial residual spatial correlation.

Fig. 5.3 shows the graphical model. The estimated values of several coefficients differed from those of the standard SEM model calibrated with the same data (Chapter 4), although differences were never greater than 0.3, and none of the coefficients changed sign. Note that all variables were standardized to have zero mean and standard deviation equal one. Some relations were no longer significant, such as those between EVISD and CEC of B horizon, between LSTM and OC of A horizon and between NDWI.B and OC of A horizon. Note that this does not mean that these links do not exist, but that there was not enough statistical evidence that the coefficients are different from zero. Legendre (1993) reported that including spatial correlation increases the standard error of the estimates, what can make that some coefficients become non significant.

We notice a substantial increase of the LON coefficient but this effect is compensated by an increase of the DEM coefficient. The two effects compensate each other because LON and DEM are negatively correlated (correlation coefficient -0.97). So it is difficult to interpret these changes from a pedological point of view.

Presence of a fairly strong residual spatial autocorrelation suggests that one or more important external factors are missing, and hence not captured by the used set of covariates. We hypothesise that the climate factor is well represented by the annual mean land surface temperature (LSTM) and the two wetness indices (NDWI. A and NDWI.B). The Organism factor might be partially represented, because we did not included land use type. Some areas within the region are being irrigated, which highly impacts the response of the MODIS products. By including the land use we could model these region in a different way. Probably, the parent material is the worst represented of the soil-forming factors. Gunal and Ransom (2006) reported variation in clay mineralogy within the study area, which might have a large impact in the CEC. This could be solved by including data from gamma-ray sensor which are highly correlated with the parent material (Cook et al., 1996).

Fig. 5.4 shows the semivariograms and cross-semivariograms of the residuals

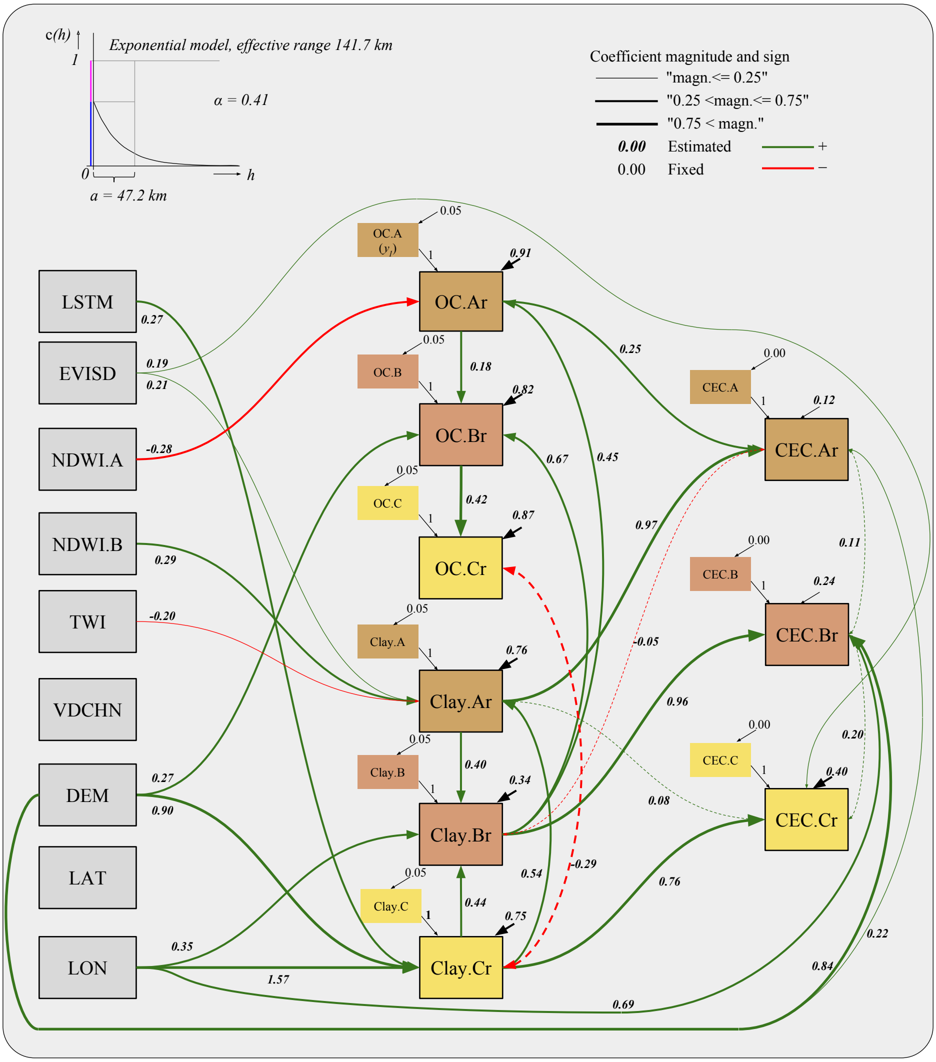


Figure 5.3: *Graphical model, with model coefficients for the spatial model. Italic bold numbers represent calibrated coefficients, normal font numbers are fixed parameters. Red and green continuous arrows represent* **B** *and* **Γ** *matrices, where the thickness reflects the magnitude of the coefficient and the colour its sign; black continuous arrows represent non-zero elements in the* **K**, **Ψ** *and* $\mathbf{\Theta}_\varepsilon$ *matrices. Dashed double-headed arrows represent model error correlations. Acronyms of external factors (grey boxes) and soil properties are described in the main text. 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).*

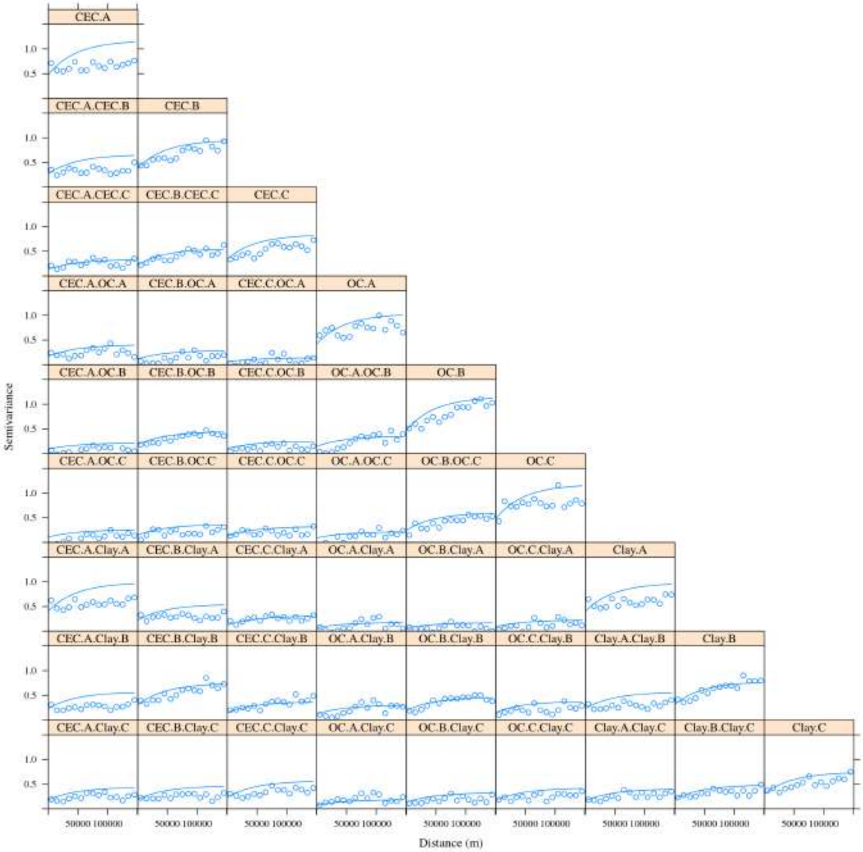


Figure 5.4: *Semivariograms and cross-semivariograms of SEM residuals. Blue circles represent the experimental semivariograms, blue lines the semivariogram models.*

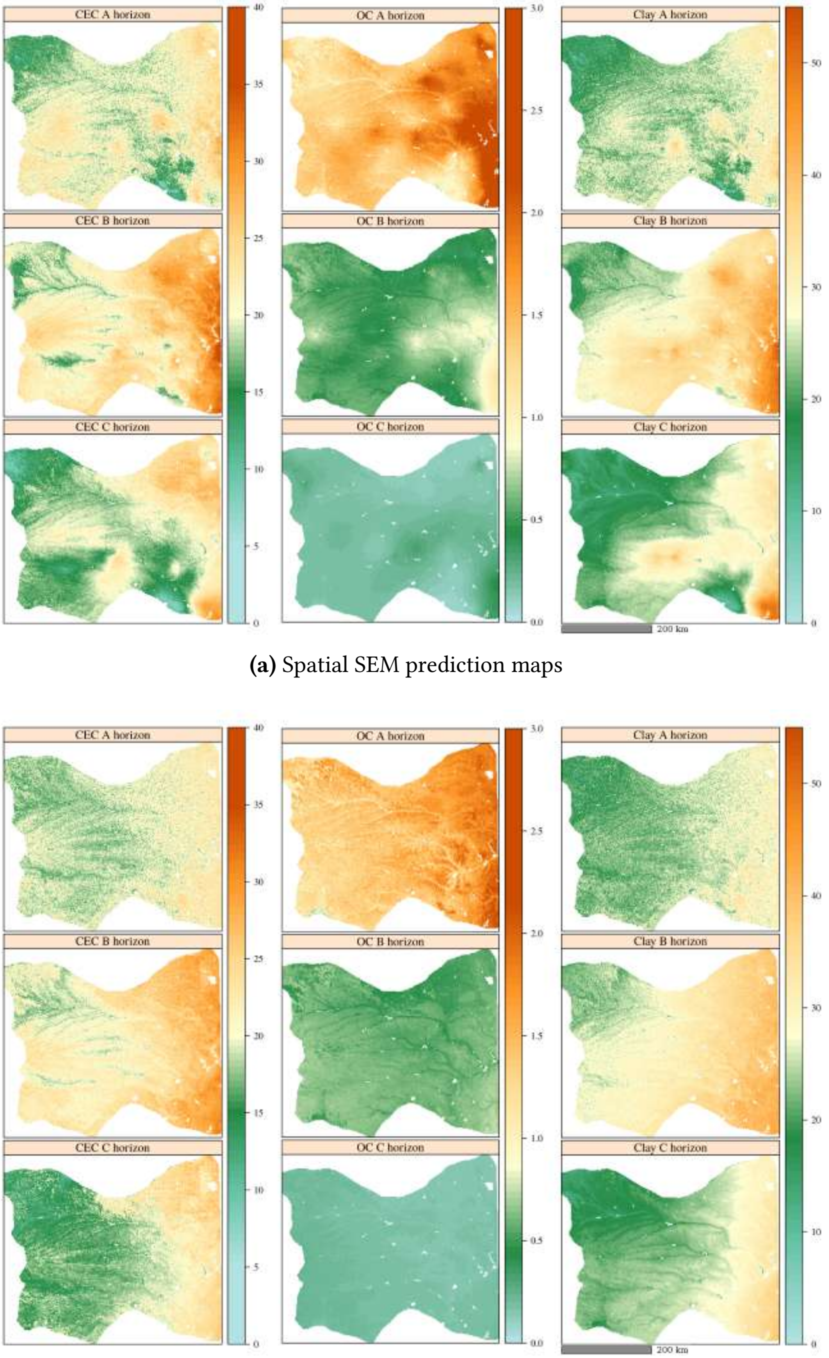
($(\mathbf{I} - \mathbf{B})^{-1}\boldsymbol{\zeta}$) derived from the correlation function c and covariance matrix $\Sigma(\boldsymbol{\theta})$. The semivariogram models fit the experimental semivariograms quite well, even though parameters α and a were imposed to be the same for all system variables. There is a slight tendency to overestimate the partial sill of some variables (e.g., CEC and Clay of A horizon). A more flexible approach would allow that each variable and pair of variables had their own spatial correlation parameters, but this would lead to a dramatic increase of the total number of parameters, which will complicate the parameter estimation process. Moreover, it might become difficult to verify that the calibrated system satisfies the positive-definiteness requirement. An intermediate solution that would add flexibility while still ensuring positive-definiteness, would be to adopt the Linear Model of Coregionalisation (Wackernagel, 1995).

5.3.2. Prediction maps and accuracy assessment

Fig. 5.5 shows the prediction maps of CEC, OC and Clay for all three horizons obtained with spatial SEM (Fig. 5.5a) and standard SEM (Fig. 5.5b). Note that we masked pixels that correspond with urban areas and water bodies, these are shown as white pixels inside the rasters. In general terms, the maps of clay and CEC for A and B horizons produced with spatial SEM show a similar pattern as those obtained with standard SEM. However, where residuals deviate from zero spatial SEM increases or decreases the predictions, which are shown as darker and lighter patches in the maps. This effect is most pronounced in the maps of OC of A and B horizons, where the standard SEM map shows a pattern mainly controlled by drainage and the spatial SEM shows an important kriging effect. The kriging effect is also present in the C horizon Clay and C horizon CEC maps. For instance, it increases spatial variation of predictions in the south-eastern part of the study area, as well as in the western parts.

Interestingly, the spatial SEM map of CEC produces a similar pattern as the spatial SEM clay map in some regions, such as in the south-east, and different patterns in other regions, such as in the south-west. The strong link between CEC and Clay of the C horizon (Fig. 5.3) indicates that an increase (or decrease) of Clay should also increase (or decrease) the CEC, unless there is a dramatic change in the type of clay mineralogy, which might be the case in the south-western region of the study area. Note that in this region the sampling density is low (Fig. 5.1), thus a change in mineralogy of an isolated soil sample affects the surroundings, even though the sample may not be representative of the entire region. Although a deeper investigation is needed to explain these discordances, the method implemented reflects the added value of joint modelling of multiple soil properties using SEM.

The spatial SEM approach outperformed the standard SEM in terms of accuracy (Table 5.2). The MEs of the spatial SEM show a slight bias for some variables, but their magnitude is small compared with RMSE. Including residual spatial autocorrelation dramatically improved the AVE of CEC of A and C horizons and the AVE of clay of B and C horizons. When grouping predictions over horizons, spatial SEM leads to notable improvement for CEC and clay, while for OC the improvement is negligibly small. Fig. 5.6 shows the observations against the leave-one-out cross-validation predictions. This confirms that the best predictions are obtained for clay, followed by CEC and OC. It also confirms that a substantial part of the spatial variation of the soil properties is not explained by the model.



(a) Spatial SEM prediction maps

(b) Standard SEM prediction maps

Figure 5.5: *Prediction maps of cation exchange capacity (CEC) (cmol$_c$ kg^{-1}), organic carbon (OC) (mass %), and Clay (mass %) for the A, B, and C horizons using spatial SE model* ***(a)*** *and standard SE model* ***(b)****.*

Table 5.2: *Leave-one-out cross-validation accuracy measures of standard SEM and spatial SEM.*

		Standard SEM			Spatial SEM		
SP	Hor.	ME	RMSE	AVE	ME	RMSE	AVE
CEC	A	0.01	5.56	0.11	−0.07	5.13	0.24
	B	0.01	5.87	0.22	−0.11	5.46	0.32
	C	0.00	6.39	0.22	−0.11	5.37	0.45
OC	A	0.00	0.56	0.19	0.00	0.53	0.27
	B	0.00	0.23	0.01	0.00	0.21	0.18
	C	0.00	0.11	0.04	0.00	0.10	0.18
Clay	A	0.00	7.92	0.19	−0.13	7.44	0.29
	B	0.00	8.68	0.30	−0.17	7.95	0.42
	C	0.01	9.38	0.23	−0.19	7.70	0.48
CEC	Joint	0.00	5.95	0.26	−0.10	5.32	0.41
OC		0.00	0.36	0.75	0.00	0.34	0.78
Clay		0.00	8.68	0.33	−0.17	7.70	0.48

SP is soil properties, Hor. is horizons, ME is mean error, RMSE is mean root square error, AVE is amount of variance explained.

5.3.3. Spatial SEM scope and limitations

The methodology presented showed how residual spatial correlation can be successfully integrated in the SEM framework. Future work could extend this by incorporating an integrated model evaluation approach (Bollen, 1989, Chapter 7, pp. 256), adding more flexibility in parametrisation of the multivariate residual, including uncertainty of parameter estimates in the prediction process, as well as incorporation of the method in generic and freely available software.

The first of these is probably the most essential. We did not measure the overall fit of the model, which target to measure in what extent the sample variance–covariance matrix departs from the model-implied variance–covariance matrix. We explained the model evaluation step with more details in Chapter 3. In this case, we can only argue that it must be more precise than the standard SEM, since the accuracy assessment that we have done shows substantial improvements.

A potential advantage of the spatial SEM approach is that it may aid the improvement of the conceptual model. Residual spatial autocorrelation is caused by mechanistic processes (Legendre, 1993), and hence presence of residual spatial correlation indicates that some causal factors have not been adequately incorporated in

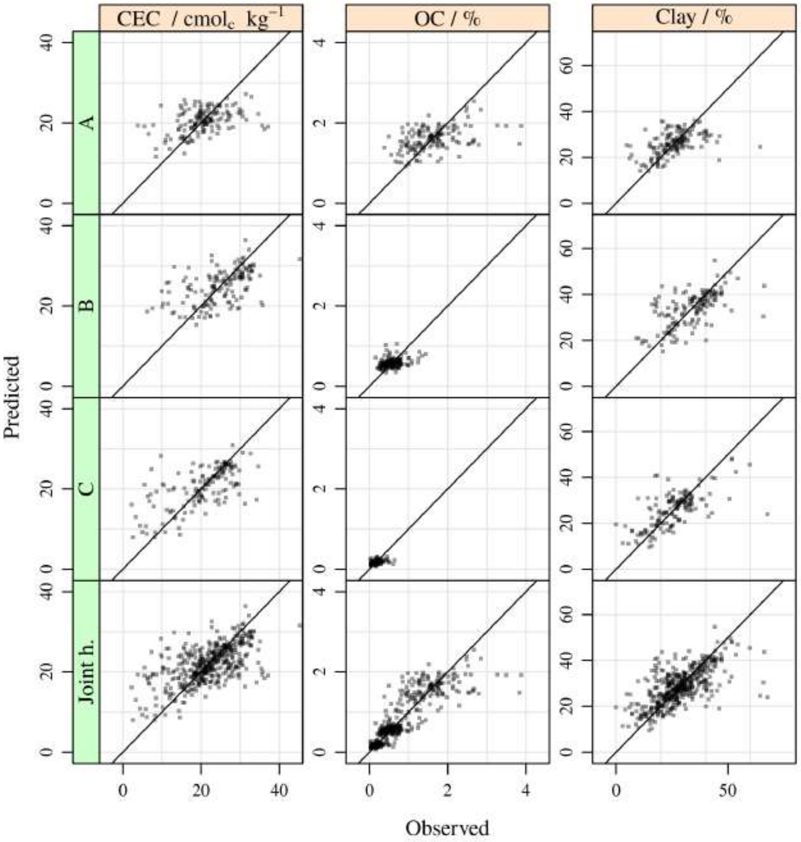


Figure 5.6: *Scatterplots of observed against predicted soil properties obtained by leave-one-out cross-validation using spatial SEM. Columns are soil properties: cation exchange capacity (CEC), organic carbon (OC) and Clay. Rows are horizons A, B, C, and "Joint h." (data for three horizons joined).*

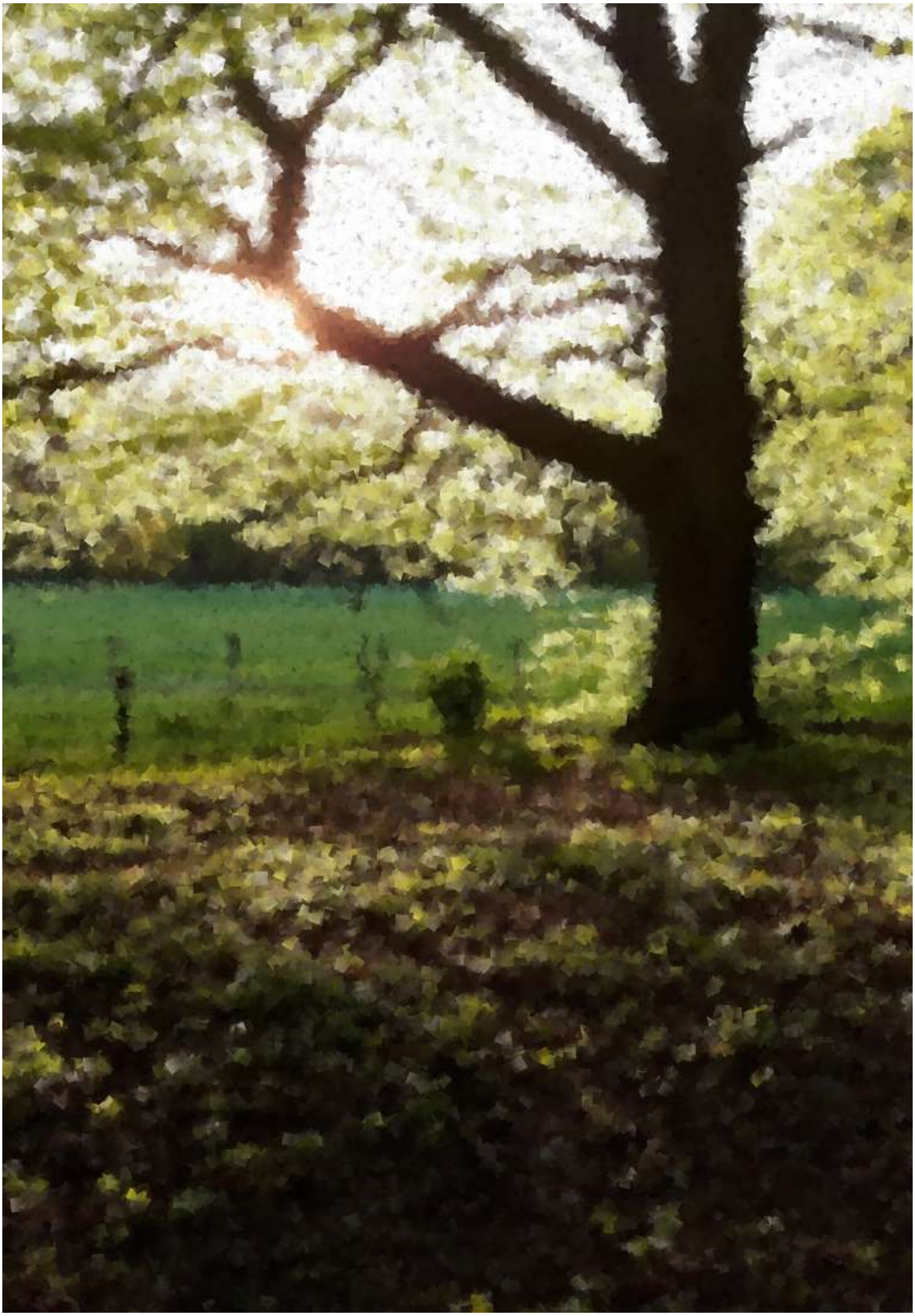
the model. Comparison of the strength of spatial correlation between different soil properties might indicate which soil forming factors are poorly represented. In the case study we imposed the same nugget-to-sill ratio and variogram range for all properties. If this condition was relaxed and these parameters allowed to vary by soil property, it might reveal which soil properties would benefit most from adding (proxies of) soil forming factors. This finding may then be combined with pedologic knowledge in search for appropriate additional covariates. If extension of the set of covariates would remove all residual spatial variation then this would indicate that all spatially structured soil forming factors have been adequately represented. Thus, analysing and modelling residual spatial autocorrelation can aid the conceptual model building process of SEM.

Taking it a step further, we might also try to include a causal model of spatial correlation. As yet this is very rare in DSM, but it has been applied in ecology to support theories of animal behaviour (e.g. Legendre and Fortin, 1989). In pedology, spatial processes such as lateral water flow, erosion, sedimentation and vegetation growth might be included in DSM. This brings us close to mechanistic soil modelling (Opolot et al., 2015; Temme and Vanwalleghem, 2016), which potentially is of great importance to DSM, especially when DSM is extended from mapping static soil properties to modelling and prediction of the distribution of soil properties in space and time.

5.4. Conclusions

In this study we have shown how to combine SEM with geostatistical interpolation. We illustrated the approach with a case where the spatial distribution of three soil properties over three soil horizons was modelled simultaneously. The main conclusions of this study are:

- Including residual spatial correlation in SEM can be achieved using a regression kriging approach.
- The spatial SEM method outperformed the standard SEM method in terms of prediction accuracy for a case study in the Great Plains of the United States.
- Modelling of residual spatial correlation would benefit from more flexibility than used in this study. More flexibility can be achieved by adopting the linear model of coregionalisation.
- Including residual spatial correlation influences the magnitude of calibrated SEM coefficients, as well as their standard errors, but it is not easy to explain the differences from a causal pedological perspective.
- The computational demand of the spatial SEM approach is modest for medium-size datasets but may become problematic for large datasets and studies involving a large number of soil properties.
- Presence of residual spatial correlation indicates that important covariates are missing in the SEM model. Analysis and modelling of residual spatial correlation can help improve the conceptual modelling step of SEM and the selection of appropriate covariates.



Chapter 6

Synthesis

6.1. Introduction

In Chapter 1 I argued that digital soil mapping (DSM) will benefit from including soil process knowledge in spatial prediction models, because in order to properly describe or map soil spatial variation, we need to understand soil behaviour. This is needed to answer questions such as: which are the dominant soil processes in a certain region? How will the soil react under increased productivity pressure? How vulnerable is the soil to erosion or pollution? How much organic carbon can we store in the soil at a given location? To answer these questions it is not enough to describe the soil in form of a map, but we also need to represent our knowledge about the soil (Bui, 2004). Structural equation modelling (SEM) is a knowledge-driven statistical modelling technique that allows to model complex relationships in a system. In this thesis I explored if SEM is a suitable technique to include pedological knowledge in DSM.

In this final chapter I discuss whether the objectives of this thesis were achieved and whether SEM is indeed a valuable addition to the DSM toolbox (Section 6.2). I will also look ahead and identify topics for future research (Section 6.3). Finally I will summarise my perception about the future of SEM in DSM (Section 6.4) and give the main conclusions of the thesis (Section 6.5).

6.2. What have we learned? Main findings of this thesis

The overall objective of this PhD thesis was to extend DSM with soil process information through the *development*, *calibration*, *application* and *validation* of a structural equation (SE) model. This objective was addressed through four specific objectives with associated research questions (Section 1.4), of which the results were presented in Chapters 2 to 5.

In this section I summarise the findings of the previous chapters by four topics and identify the main strengths and limitations of using SEM for DSM. Section 6.2.1 summarises and discusses the general merits and challenges of using SEM for DSM. In Section 6.2.2 I put SEM in the context of multivariate and multi-layer soil modelling and prediction. Section 6.2.3 explores the potential of model extrapolation. Finally, Section 6.2.4 considers the extension of SEM to account for spatial correlation in observational data.

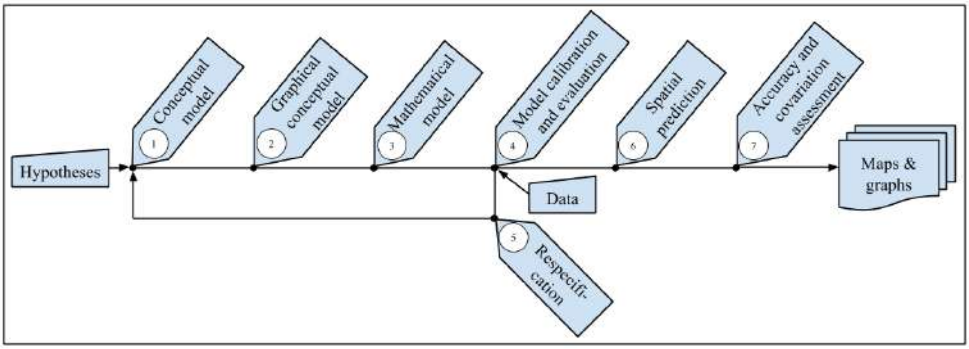


Figure 6.1: *Steps in structural equation modelling (SEM) for spatial prediction of soil properties (copied from Chapter 3).*

6.2.1. Using SEM to include soil-forming processes in DSM

Translating pedological knowledge into a SE model

Process knowledge is incorporated in SEM in the first step of the modelling process (Fig. 6.1, step 1), when one defines a conceptual model and the associated graphical model. The conceptual model combines theories or hypotheses that explain how the system under study functions. One begins with defining a general theory on system functioning. Then one links these theories to measured variables. This is not an easy task, since hardly ever there is a one-to-one correlation between a theory in the conceptual model and the measured variables (Grace et al., 2012). Once the system variables to represent the conceptual model are identified, a graph is constructed that represents causal relationships between system variables by arrows (Fig. 6.1, step 2). After the structure of the model is defined, the model is ready to be calibrated (Fig. 6.1, step 3). The remaining steps were explained in Chapter 3 (Section 3.2.6), but here I concentrate on what I experienced for the first steps.

The first challenge in translating a conceptual soil-landscape model to a graphical model is to define proxies for the soil-forming factors, because usually the covariates lack direct pedological meaning. For example, the altitude above sea level, usually represented by a digital elevation model, is an important feature of the landscape and generally well-correlated with soil properties such as clay content, but the spatial distribution of clay depends on the type of landscape and its relative position, rather than on the altitude above sea level. Also, every soil-forming factor might be (partially) represented by several proxies (see Fig. 6.2 for an illustration), each of which could function as a proxy for multiple factors.

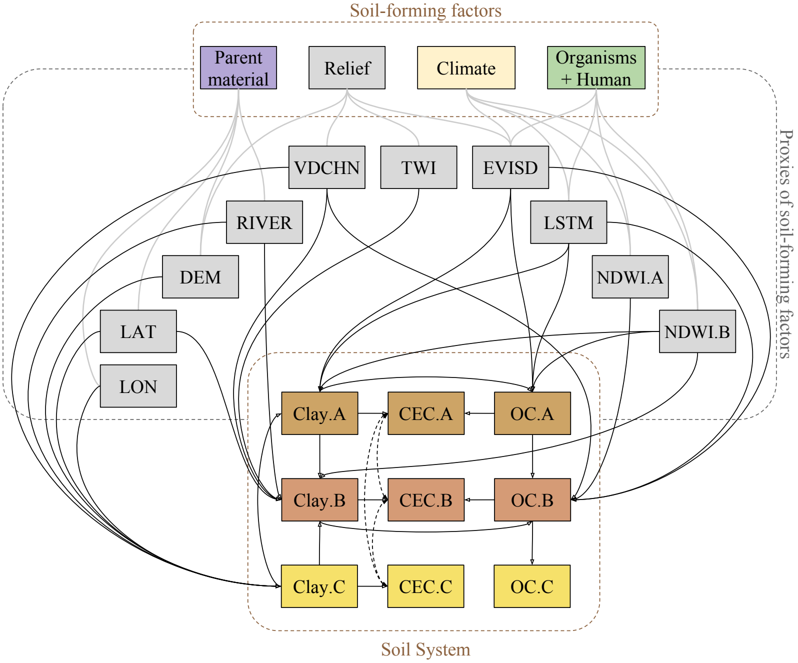


Figure 6.2: *Graphical model of three soil properties. Grey continuous lines represent the theoretical relation between soil-forming factors and external factors. Black continuous arrows are cause and effect links. Black dashed arrows are expected correlations between system errors. External factors are described in Table 3.1. 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 (copied from Chapter 3.)*

The second challenge is to represent the soil-forming processes. This cannot be done exactly because SEM is not a dynamic model, unlike a mechanistic model (e.g. Opolot et al., 2015; Temme and Vanwalleghem, 2016). The graphical model can connect soil properties and proxies of soil-forming factors that affect these properties, but not the processes themselves. Instead, the soil-forming processes are implicitly used to support the connections, which later on are characterised by coefficients that are defined on the basis of the observed correlations between variables (Fig. 6.1, step 3). Clearly, if one variable has a causal effect on another, the two are correlated. But, a soil property can also be affected by different processes simultaneously. In the face of this situation, I had to compromise decisions to avoid excessively increasing the complexity of the model.

SEM is severely handicapped by the fact that it cannot represent dynamic processes. Even though the model construction takes mechanistic principles into account it will always be a simplified approximation of the real world processes. However, SEM does include process knowledge and interrelations, much better than prevailing empirical DSM methods such as linear regression, machine learning and kriging.

Data-driven relations versus knowledge-based relations

Data-driven relations refer to the relations between covariates and soil properties generated by a model. Usually, the process of covariate selection is done with an empirical approach, for instance through stepwise selection, or by just providing all available covariates, as is typically done when using machine-learning models. The importance of each covariate is measured and reported in terms of the number of times that it is used in the model (e.g. Poggio et al., 2013; Fitzpatrick et al., 2016; Hengl et al., 2017). In this context, however, the causes of these relations remain opaque and thus do not contribute substantially to an improved knowledge of the soil system.

Knowledge-based relations help to build the graphical model. Coefficients obtained from model calibration constitute new insights for the conceptual model, so they improve the system knowledge. However, defining knowledge-based relations is not a straightforward process. The selection of variables has been extensively discussed in the SEM arena. Pearl (1998) explained that the selection of variables is one of the most frustrating issues in causal analysis (which is the probabilistic analysis of causation –Section 1.2.3), because when we want to analyse the effect of one variable X on another Y, we might want to take into account the variation caused by another variable Z. For instance, let us consider a case where we want to model the effect of land use (agriculture or pristine conditions), X, on soil organic carbon (SOC), Y, in a region with two different climates (dry or humid), Z. Let us assume that most agricultural plots are in the humid area, while the driest areas have a predominance of pristine conditions. If we use linear regression to predict the percentage of SOC by land use type (agriculture or pristine condition) without considering climate, we might conclude that agriculture has a positive effect on SOC, represented by a positive regression coefficient. But, if we include climate in the equation, we would probably see that agriculture has a negative effect on SOC, represented by a negative partial regression coefficient in a multiple linear regression. This well-known effect is known as Simpson's paradox, which states that "…any statistical relationship between two variables may be reversed or negated by including additional factors in the analysis". This effect also turns up in SEM.

We encountered this issue in Chapter 2 where we noticed, for example, that exchange sodium percentage (ESP) of the A horizon (*esp.Ar*) negatively affected the OC content of the A horizon Fig. 2.8, while at the same time the ESP of the B horizon (*esp.Br*) positively affected the OC content of the A horizon. This does not make sense from a pedological point of view as both relationships should be negative. Such problem is usually solved in the respecification process, where we add or remove arrows to the graphical model to correct for these type of problems. Therefore, Simpson's paradox might be present in any empirical approach, including SEM, but the advantage when using SEM is that it becomes apparent in the calibrated graphical model, which allows us to correct the model.

Learning from data: the respecification process

The respecification process, which consists of modifying the original model, is a standard procedure in most SEM applications (Grace et al., 2012). For this reason we implemented this step in Chapter 3. Model respecification can be assisted exclusively by expert knowledge, for example by checking that the sign and magnitude of model coefficients make sense, and removing or adding alternative paths in the system, or it can be applied using **exploratory analysis** (Section 3.2.6 Model respecification; Bollen (1989), Chapter 7).

This last approach takes into account the differences between the model-implied variance–covariance matrix and the sample variance–covariance matrix, and suggests links between variables that could decrease the differences between the two matrices. The suggestions are provided in terms of a modification index, that estimates how much improvement the model would gain if a given link is included in the model. Since this index is a univariate indicator, the number of alternative links at every modification step can be very large. The task of the soil scientist is to decide which of the proposed modifications make sense from a pedological point of view. In our case study, we knew that, in theory, cation exchange capacity (*CEC*) is not directly affected by any environmental covariate, because *CEC* is just a feature of the soil colloidal fraction. However, the model suggested to include a path between *CEC* (of the C horizon) and distance to the river (*river*). I tried to compensate this lack of fitting by including indirect links, such as *river*→ *Clay*→ *CEC*, but the model suggestions showed that it was not enough. I interpreted this difference due to the absence of other variables (not considered in the model, such as the type of parent material) correlated with both *river* and *CEC*. Therefore, I decided to include a direct link *river*→*CEC* for prediction purposes. This not only improved the performance of the prediction, but also pointed out topics that needs to be addressed in future studies. In this way, SEM is a useful tool to learn from the data.

Implications of the measurement error

Another interesting feature of SEM is its possibility to include explicitly measurement error in the model, and differentiate it from the model system error. Although most pedometricians might be aware of measurement error, it hardly ever is explicitly quantified and accounted for in DSM studies. Frequently, it is implicitly included within the error maps that are produced along with soil property maps. The error of those maps are generally assigned to the limitations of the statistical model or to the quality of the covariates, but hardly ever to the limited quality of the soil data. In SEM, instead, it is possible to differentiate the system error (the error from the model) from the measurement error by implementing latent variables. If a latent variable is measured with a single indicator, the difference between these two is given by the measurement error, the variance of which can be fixed by the researcher or estimated from the data. When a latent variable is measured by more than one indicator the process is more complex, as the measurement errors of the indicators might be correlated (for example when two variables are measured with the same instrument).

In the case studies I found that the measurement error of calibration data was high with respect to the total variance of some soil properties, such as CEC. Note that I emphasize here that the error was high in relative terms, because actually the same measurement error could be negligible in a study area that has much larger spatial variation in the soil properties. For this reason I concluded that it is difficult to get high prediction accuracy in homogeneous areas, particularly when the signal-to-noise ratio is low (Chapter 2). Another conclusion was that, since I needed to take care of this specification in SEM, I became aware of such a issue. By knowing the signal-to-noise ratio of each variable involved in the model, we can plan ahead whether to include it or not, and we can estimate the potential variance left to be explained by the model. Thus, it would be possible to predict the potential accuracy of the map.

6.2.2. SEM for mapping multi-layers and multivariate soil properties

Soil information in three dimensions

In Chapter 3 I targeted to predict multiple soil properties in three dimensions. The third dimension of the soil was represented by the three mayor genetic horizons (A, B, C), rather than by soil depth. The advantage of this approach is that there is a more natural connection between the conceptual model and the horizons, as the soil-forming processes can more easily be grouped by horizons. In this sense, SEM can nicely take care of multi-layers because it can be decided on pedological grounds which arrows between horizons should be included and which not. Thus, I built a model where clay and OC of the A, B, and C horizons were affected by covariates that represent soil-forming factors, while at the same time clay and OC affect the CEC of the three main horizons (Chapter 3). Since it is known that some soil-forming processes are more prevalent in some horizons than in others, it was possible to define how horizons are connected to one another. For instance, organic matter accumulates in the A horizon. Water flow and bioturbation move organic material to deeper layers, so the amount of organic carbon in the A horizon affects directly the amount of organic carbon in the B horizon, and this in turn affects the amount of organic carbon in the C horizon. This process not only affirms that organic carbon of different horizons should be correlated, but also determines the direction of the arrows in the graphical model. In this way we can represent soil property interrelations that show how soil properties between horizons interact with one another. In the case study, calibration confirmed that CEC (from a given horizon) was mainly affected by clay and OC content of the same horizon. Accordingly, the CEC maps of the three horizons (Fig. 3.8) preserved a combination of patterns of the clay and OC maps. These maps were, therefore, in concordance with the graphical model of Fig. 3.7.

Interpretation of digital soil maps: the graphical model

In the past, soil maps were produced along with reports that helped to interpret the soil information that they contained to support agricultural planning (Brevik and Hartemink, 2010). The new DSM methods have matured and become operational during the last years (Kempen, 2011). Reports for soil maps interpretations have not been adopted in DSM, probably because digital soil maps can be considered self-explanatory.

However, being self-explanatory may not be sufficient for users. Knowledge of the causes of spatial variation of soil properties is also important, particularly for decision makers, politicians, farmers, and other land managers. In this respect, SEM can contribute to DSM with a new tool for interpreting digital soil maps, which is the graphical model. According to Pearl (1998), graphical models are symbolic systems for concepts and relations that are not easily expressed in a system of equation. In DSM, they may explicitly represent the model, which summarises the system relationships in a simple way, and help to understand how soil properties relate to one another. Thus, it might be highly relevant for land management, especially when the model takes into account soil properties that determine crop yield and that can be changed in short periods. Since we might improve some soil conditions with land management, such as soil pH or soil compaction, our intervention in the system would be better supported if we knew which other soil properties affect the variables that we wish to improve. Graphical models can be used, then, as a complementary source of information next to digital soil maps.

Assessing covariation among soil properties

Like other multivariate models, such as multivariate linear regression modelling (MvLR), SEM is capable of reproducing the covariation between soil properties. However, SEM does this more efficiently than MvLR because it only uses meaningful parameters, while 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 with MvLR. However, using SEM allows to reproduce the covariation on the basis of a conceptual model, while MvLR would not contribute as much to the understanding of the system under study.

Covariation assessment is not a standard procedure in most DSM approaches and it has only been roughly addressed by some authors (e.g. Lacoste et al., 2014). I presented an approach to assess model covariation in Chapter 3 and proved that models that do not take covariation between target variables into account fail to get an accurate prediction error variance–covariance matrix. Note that the assessment of covariation was based on the same data that were used to calibrate the model. As this might cause some bias, it may be preferred to split the dataset into calibration and validation datasets. Also, note that to reproduce the covariation in the predicted values, it is not enough to predict the mean values, but that several realizations of the predicted variance–covariance matrix had to be simulated to reconstruct the relationships. This kind of prediction would be useful to create probabilistic maps, which answer questions such as "where are the areas that have more than 50% of clay in the B horizon with a 95% confidence level?".

6.2.3. Soil map extrapolation with SEM

Possibilities of model extrapolation with SEM

We tested different approaches to predict the spatial distribution of several soil properties across a study area in the United States. These approaches are summarised in Table 6.1 (replicated from Chapter 4). Model 1 was the extrapolation of the mathematical model, which means calibration of a SE model with Argentinian data and

Table 6.1: *Settings of the different models used for extrapolation (copied from Chapter 4, Table 4.2).*

	Model type	Extrapolation	Data for respecification	Data for calibration	Prediction locations
Model 1	SEM	MM	Arg.	Arg.	USA
Model 2	SEM	CM	Arg.	USA	USA
Model 3	SEM	CM	–	USA	USA
Model 4	SEM	CM	USA	USA	USA
Model 5	MLR	MM	–	Arg.	USA
Model 6	MLR	MM	–	USA	USA
Reference	SEM	MM	Arg.	Arg.	Arg.

Column “Model type” refers to the SE model (SEM) or a multiple linear regression model (MLR). “Extrapolation” refers to which part of the model was extrapolated: either the graphical model (GM) or the mathematical model (MM). The column “Data for respecification” refers to which dataset (Argentinian [Arg.] or US [USA] dataset) was used for the respecification step. A dash (–) means “No respecification”. The last two columns detail which datasets were used for calibration and (cross-)validation.

prediction in the US case study area. The same was done with MLR (Model 5) to have a reference for comparison of the performance of SEM. The results showed that SEM performed better than MLR (Table 4.2), although both prediction methods performed poorly. This was not surprising, as other authors reported similar experiences (Lagacherie et al., 1995; Grinand et al., 2008; Malone et al., 2016). Grinand et al. (2008) argued that the reason for the low accuracy was due to the lack of spatial correlation between samples of calibration and prediction areas, while Malone et al. (2016) argued that it was due to differences in terms of the covariate feature space between both areas. I, instead, found that the main reason of the mismatch was that the study areas present different relationships between the system variables. For example, some correlations between covariates and soil properties were positive in one dataset, while these were negative in the other dataset. Extrapolation can only produce nonsensible results in such case. In order to avoid this, we should derive and use covariates that reveal the dynamics of the soil. O’Geen (2012) showed how the soil moisture regime is a key factor of soil development. I will present possible methodologies to obtain such type of covariates in the following section.

Models 2, 3, and 4 differed in the degree of adaptation to the US study area. Model 2 had the same specifications as the reference model (the model developed in Chapter 3), which means that both models shared the same structure but not the coefficients. Model 3 used the structure of Fig. 3.6, which is the original model of the Argentinian case study without respecifications. Model 4 was respecified using the

US data. The results confirmed that the respecification done in the Argentinian case (Model 2) degraded its capability to be extrapolated; the more general model (Model 3) was more accurate than Model 2 when used for prediction in the US case study. The respecification of Model 4 further increased its prediction power, since these adapted the model to the local conditions in the US. Model 6 was a MLR model calibrated and applied to US data to be compared to Model 4. Again, SEM performed better than MLR (see Table 4.2).

Causal analysis for DSM

In Chapter 4, we applied the same conceptual model in two different areas (Models 2 to 4). We described in detail the differences between the models and their pedological implications. On the basis of this experience, I would like to discuss here two topics: the **pedological support for the system interrelations**, and in relation with this, the use of **SEM for testing conceptual models**.

In the process of defining the relation between system variables, one has to invoke pedological hypotheses to link soil properties to each other, and soil-landscape hypotheses to define relations between covariates and soil properties. Pedological hypotheses can easily be applied because there is a good understanding of many processes that work at the pedon level. On the basis of these differences it was possible to find literature that explained some of these processes (Sections 3.3.1 and 4.3.1), so it improved the knowledge about the system. However, it was not easy to link covariates and soil properties because there are not many generalisable soil-landscape hypotheses, neither good covariates to represent the landscape features. We could explain the relations between some covariates and soil properties, but for some of them, such as the standard deviation of the NDVI it was difficult to judge whether the coefficients were correct or not. Then, the comparison between both study areas in this respect was also difficult, so I concluded that these interactions need to be re-created with better covariates.

Testing a conceptual model is one of the main applications of SEM, and in this context, it might be very relevant for DSM. If we can develop robust conceptual models, for example, for a given soil type, then it would stand alone as a piece of knowledge, and at the same time, it might be used for prediction. We should not expect to be able to predict the soil properties accurately with such a model for every single case, as hardly ever the conditions between study areas are the same, but at least we should be able to analyse the differences between regions, to compare the structure that the conceptual model takes for each study area, and explain the differences. These options would represent new features for DSM.

6.2.4. Spatial SEM: a geostatistical approach

The development and application of the spatial SEM method

In Chapter 5 we combined geostatistics (Webster and Oliver, 2007) with SEM by using a maximum likelihood estimator (ML) (Bollen, 1989) for model calibration. It required the use of a full variance–covariance matrix that not only takes into account covariances at distance zero, which is what the conventional SEM ML estimator does, but also covariances at distances different than zero. Similar rationale has been already applied in SEM. For example Lamb et al. (2014), developed a method to analyse how the spatial correlation affected the system interrelationships at different scales, but did not target to make predictions. Wall (2012), instead, developed an approach similar to our needs, where they modelled the spatial correlation in the residuals of a SE model. However, they used the lattice instead the geostatistical model because they applied it to polygon-support data. We implemented the methodology in R (R Core Team, 2017) making use of several functions from `lavaan` (Rosseel, 2012).

We described the mathematics to predict following the conditional distribution theory[1]. The approach that we implemented for prediction was similar to regression kriging: the SE model coefficients and the variograms and cross-variograms parameters were estimated simultaneously. We applied the spatial SEM method in the US case study area and compared the predictions with Model 4 of Chapter 4 (called "standard SEM"), with the prediction of the new approach ("spatial SEM") done in Chapter 5 (Fig. 5.5). Visually, the main differences between the maps were the circular patterns produced by the kriging of the residuals. The spatial SEM maps proved to be substantially more accurate than the standard SEM maps on basis of cross-validation.

Causal analysis of spatial correlation

Accounting for spatial correlation had a substantial impact on the magnitude of the model parameters, but as expected, not on their signs. Several coefficients that related covariates to soil properties became insignificant, while others, such as those that connected altitude with clay (of B and C horizons), increased their magnitude. Even though some covariates lack pedological meaning. This is the case, for example, for altitude which likely represented a change in parent material, rather than an effect of altitude on clay. Adding other covariates to the model, such as gamma

[1] https://en.wikipedia.org/wiki/Multivariate_normal_distribution#Conditional_distributions

radiometer data, which are available for the US study area, could help to ascertain this.

In this context, residual spatial correlation points to missing soil-forming factors or processes in the conceptual model. Legendre (1993) argued that spatial correlation is given by physical forces and ecological processes, but he did not analyse the causes of spatial autocorrelation in depth. Instead this study focused on presenting a framework to include spatial correlation in causal analysis with analogous methods as I used in Chapter 5. However, other authors (Lichstein et al., 2002; Guisan and Thuiller, 2005; Dormann, 2007) argued that the presence of residual autocorrelation is a symptom of missing environmental covariates. In geostatistics, where the semivariance measures spatial variation that tends to increase with spatial distance, there is no claim that distance is a causal factor itself. Instead, environmental forces and processes create similarity among observations, so spatial autocorrelation is a property of the environmental covariates rather than of the soil properties. Then, as covariates are spatially exhaustive, the trend of the geostatistical model includes spatial correlation implicitly through the covariates, and the residual part of the model would become uncorrelated ("white noise" or "pure nugget"), so that kriging would not have added value. What is more, the fast release of new remote sensors will likely produce more meaningful proxies of soil forming factors (such as soil moisture), thus the spatial SEM approach will be just a nice temporal solution for those cases where the covariates do not provide the needed information.

In soil science, we can identify many processes that cause spatial correlation, such as sedimentation by aeolian or water erosion, bioturbation, water movement along the soil profile vertically and laterally, etc. Sedimentation, for example, will affect the parent material distribution, so if we counted with a spatially exhaustive map of parent material, the spatial correlation would be captured by the effect of covariate (i.e., parent material) on soil properties, and not by the residual.

The impact of spatial SEM on map accuracy

The case study where I applied the spatial SEM approach (Chapter 5) yielded a remarkable improvement in prediction accuracy. The largest improvement was in the clay of the C horizon (Table 5.2), where the amount of variance explained (AVE) increased from 0.23 to 0.48, and consequently the CEC of the C horizon, that largely depends on the clay percentage (see Fig. 5.3), also increased, in this case from 0.22 to 0.45. Other important variables, such as OC of A horizon, also improved substantially (AVE from 0.19 to 0.27), while the AVE of the other soil properties improved at least by 0.1 at the horizon level. I did not analyse the causes of the residual spatial correlation. It seems that a large part of it could be explained by the parent material distribution. Since parent material is highly correlated with gamma-ray sensor data (Cook et al., 1996), and these data are publicly available for the study area (Duval et al., 2005), it would be possible to evaluate this hypothesis and see if inclusion of a gamma-ray covariate increases the accuracy of the standard SEM maps and reduces residual spatial autocorrelation.

6.3. What more can we learn? Future research

In this thesis I have introduced SEM as a knowledge-driven method for DSM. I learned a lot from this experience, but in fact we are still at the beginning of the potential of SEM for DSM. There are many aspects of SEM that require further research, and there are a large number of applications in other science disciplines from which we could learn a lot. In this section I will explore some of these.

The role of latent variables in SEM In this thesis, the environmental covariates were treated as deterministic variables, that affected (latent variables of) soil properties. I used one single measured variable per latent variable (see Fig. 5.3, e.g. *OC.A* is the observed variable of the latent variable *OC.Ar*). In effect, I used the concept of latent variables to deal with measurement error. I also only applied non-recursive paths, which means that if *A* affects *B*, *B* does not affect *A*, neither directly nor indirectly. However, there are many more options to represent the soil-landscape system using SEM. Latent variables usually depict conceptual variables that cannot be measured directly, but through one or more indicators. Grace and Bollen (2008) explained that usually latent variables are seen as the cause of the indicators. For example, one of the causes of the average yield of crops is *soil fertility*. So, *soil fertility* could be a latent variable that is measured through the average yield of several crops for a given location. However, Grace and Bollen (2008) also points out that this is not always congruent with the researcher's needs, as indicators might cause the latent variable to take a different shape than envisaged by the researcher. For instance, we might wish *soil fertility* to be the result of nitrogen, phosphorous and potassium concentration of the soil. If crop yield is taken as an indicator of soil fertility it might not match with these concentrations, because crop yield is also influenced by weather and land management. SEM can provide a way out and Grace and Bollen (2008) explained in detail the possible type of structures that may be used for this. Hence, it would allow to create maps of latent variables that represent concepts of processes.

Categorical variables in SEM For the implementation of SEM in this thesis, the target variables for modelling were always continuous. Soil survey data however, contain many categorical variables for describing soil morphology and for classifying soils. Descriptive data, such as horizons, colour, structure, mottles, etc., have been used for soil classification and are indications of soil processes (Hartemink and Minasny, 2014). To include these in SEM would make a lot of sense. However, under the methods that we used in this thesis, categorical variables would violate the model assumptions (Bollen, 1989, pp. 433). The SEM literature describes different methods to deal with categorical data. Edwards et al. (2012) explained that these can be included as indicators (measures) of continuous or categorical latent variables. At the same time, they can be dependent or independent variables. For each of these cases there is a particular approach. A common solution in SEM is to replace the categorical observed variable by an underlying latent continuous variable.

Non-linear relationships Throughout this thesis, I assume that the structural equation is linear. In other words, I assume that the effect of covariates on soil properties and the effect of soil properties on other soil properties may be represented by a linear equation (such as Equation 3.2). Grace and Keeley (2006) mentioned that multi-group analysis, which is analogous to regression-tree analysis but more flexible, can solve some problems of non-linearity. Jöreskog and Yang (1996); Schumacker (1998) demonstrated the use of quadratic functions in SEM and discussed its drawbacks. They mentioned that there should be a good reason to implement such functions, as the model becomes more complex and the problem of non-convergence occurs more frequently.

SEM for time-series analysis One of the features of SEM is that it is a static modelling method, and we implemented SEM as such. However, there is an alternative method within SEM for time-series analysis named latent variable growth curve modelling. Duncan and Duncan (2004) described the principles of this methodology for social and behavioural science to study the nature of change of a system along a period of time. In order to model the pass of time, it includes two parameters which are the initial status (or intercept) and the rate of change (or slope) for a given number of time steps. Although it is not a competitor approach of mechanistic models, such as Temme et al. (2006); Finke (2012a); Temme and Vanwalleghem (2016), it might bring valuable alternative ways of treating time-series data. Thus, SEM could be useful for monitoring and modelling soil change, for instance in carbon stocks.

Improvement of environmental covariates In this thesis, we usually experienced poor prediction accuracy (Chapters 2, 3, and 4) and a large effect of spatial correlation on the prediction (Chapter 5). This was generally caused by the fact that the covariates were poor proxies of the soil-forming factors. This problem is frequently solved in DSM by including more and more covariates that nowadays are available (e.g. Fitzpatrick et al., 2016; Hengl et al., 2017). However, this is not an immediate solution under a SEM framework, first, because SEM cannot handle a large number of variables, and second, because it would not make sense to include a lot of variables that cannot be supported by a pedological rationale. Instead, the solution would be to develop meaningful covariates that represent the key soil-forming factors of a region in a better way. In my view, the study of Kuenzer et al. (2015) gave some hints how to get there. In this study it was noted that the source of remote sensing data is still increasing and that time-series data are already long enough to study landscape dynamics. They summarised a number of methods to do so, such as "Breaks For Additive Season and Trend —*BFAST*" (Verbesselt et al., 2010) and *TIMESAT* (Jönsson and Eklundh, 2004). These algorithms allow to extract the long term trend, seasonal systematic variation, and anomalies, among other characteristics. Assuming that one of the causes of land cover variation is the soil condition, this type of product could indirectly help to create more meaningful environmental covariates.

Increase of calibration data Other means to improve model accuracy would be by increasing the number of soil samples used for calibration. This is a common conclusion in many DSM studies (e.g. Landrum et al., 2015; Brevik et al., 2016; Heuvelink et al., 2016), however, it is difficult nowadays to get enough funding for that purpose. An alternative approach could be to include non-specialist data in DSM, namely citizen science. Rossiter et al. (2015) summarised several projects related to soil science that are already collecting and using this type of data, although not for DSM. The same authors identified potential type of data contributors, type of information that could be collected, and how this information could be used in DSM. They listed a number of challenges, mainly related to the use of such data, such as quality control, bias and uncertainty in the data, heterogeneous source of data, and people's perception.

In this sense, SEM seems very suitable to work with data that come from social surveys. In fact, SEM has a rich history in using data from questionnaires to study, for example, customer satisfaction from people perception (Kuo et al., 2009). Thus, if we need to survey people's perception about landscape features, we may need to create latent variables that summarise concepts such as soil degradation, soil quality and soil health McBratney et al. (2017), that can be answered through several indicators of visual soil assessment (Sonneveld et al., 2014), such as presence or absence of earth worms, gullies, surface salinity, etc. This type of analysis might help to develop (latent) soil indicators that could be mapped as target variables, as well as to be used as covariates that can help predict soil properties.

SEM for soil genesis Finally, I want to emphasise the potential that SEM has for soil genesis. In Chapter 2 (Section 2.3.3), I reviewed the main soil-forming processes of the Argentinian study area. Fig. 2.6 summarised a mental model and intended to show the rationale behind the relationships set in the final model (Fig. 2.8). However, I did not discuss the pedological implications of such a model because it was beyond the scope of the study. Nevertheless, I illustrated with a simple example how SEM could be used for soil genesis. In Section 4.4.3 it became clear that the A horizon of the Argentinian Pampas may be developed in different parent material than the underlying horizons.

The static nature of SEM would limit the application in soil genesis, but we can still use it to test hypotheses about a given soil-forming process. Also, by making use of latent variables, we could represent processes measurable through key soil properties. Note that such studies would require the joint effort from soil pedologists and pedometricians, which usually do not share the same niche in soil science.

6.4. Structural equation modelling for digital soil mapping

In this PhD thesis I showed that SEM can be a useful method to include pedological knowledge in DSM. During the learning process, I encountered limitations, such as the linearity of the relationships and the static nature of SEM. But I also discovered that the quality of the environmental covariates was a limiting factor for studying causal relationships –and probably for many other DSM applications as well–, not only because it affects the accuracy of prediction maps, but also because their failure to represent the real soil-forming factors does not allow a proper causal analysis. Often, the accuracy of resulting maps was disappointing, and I found that the measurement error plays an important role when the study area is homogeneous. But then, I realised that one becomes aware of all these issues just by using SEM, because one has to take care of every detail in the model. From the beginning of the modelling process, one has to consider aspects that in other DSM techniques are part of the "black box" and may go unnoticed. When using SEM, one has to disentangle the system and analyse its parts carefully (such as the model parameters, as well as the logical reasoning behind the model structure) to be able to draw sensible pedological conclusions at the end of the process. Being trained as an agronomist with much attention for pedology, soil processes and soil management, I found this process very rewarding.

In my view, SEM has many more appealing features than limitations. Rephrasing what has been discussed in previous sections of this chapter, I would like to give a few examples to illustrate this opinion:

- The step of *model respecification* is a procedure to learn from the data and improve our understanding about the system, which, as far as I know, is completely new to DSM.
- The possibility to condense pedological knowledge in a graphical scheme is highly valuable to DSM, because:
    - we can give "common sense" to correlations contained in the data,
    - we can add (cause-effect) direction to the correlations though arrows, and
    - we can use the graphical model to interpret the maps.
- It is possible to predict several soil properties at once, while preserving the covariation between them.
- By extrapolating the conceptual model from one area to another, we can reinforce the conceptual model.

To summarise, I think that SEM is a suitable framework to fill a vacant niche in DSM. It does not compete with purely empirical machine learning techniques, nor with mechanistic models, and provides a framework that requires that pedometricians and pedologists work together.

6.5. Conclusions

This thesis introduced structural equation modelling as a new method for DSM. In Chapter 2 I described soil-forming processes and linked these to key soil properties in a SE model, from which maps were produced. In Chapter 3 I included more advanced SEM tools and applied these to predict three soil properties of three major soil horizons. In Chapter 4 I explored the capabilities of SEM for model extrapolation by comparing different model settings. In Chapter 5 I combined geostatistics with SEM to model residual spatial correlation, and applied the combined model to a case study showing the improvement of prediction accuracy. In this final chapter I summarised the main findings and explored SEM extensions that may improve our understanding of soil formation and the accuracy of prediction maps. On the basis of this thesis it can be concluded that:

- SEM is a suitable framework for including pedological knowledge in DSM.
- SEM is particularly recommended for mapping multivariate soil properties at multiple soil layers within regions where a limited number of processes are considered.
- Fine-tuning pedological hypotheses using SE model suggestions helps to improve the accuracy of prediction maps.
- Covariation between soil properties are well preserved using SEM, while univariate modelling fails to reproduce these interrelations.
- Covariation assessment of predicted soil properties should be a standard procedure when more than one soil property is predicted in DSM.
- Soil mapping extrapolation from one area to another remains challenging in DSM, but we may be able to overcome this challenge by creating more robust conceptual models and using more meaningful covariates.
- The adjustment of the graphical model for the Argentinian case study produced more accurate predictions in that area, but diminished the extrapolation capability of the model.
- Using SEM helps to identify the model and data errors because it forces the researcher to be aware of all system connections and parameters.
- Environmental covariates are often poor proxies of soil-forming factors, which is the case for many applications in DSM. This is a drawback for

SEM, since it does not allow to build robust graphical models that reflect soil-forming processes.

- Including spatial correlation in SEM improved the accuracy of the predictions considerably for a case study but it did not contribute to improved understanding of mechanistic processes underlying the soil-landscape system.

I am confident that SEM could become a new tool for a more *conscious* DSM. I use the word *conscious* because using SEM forced me to become more aware of the processes behind the system interrelationships. Working towards a *conscious* DSM is not easy and poses a real challenge, if only because we need to understand the nature of the data that we use to be able to create the right proxies of soil-forming factors.

Empirical methods, such as kriging, multiple linear regression and machine learning techniques might yield very accurate maps, which indeed are useful for many purposes, but such maps do not necessary help to reveal and understand the processes behind the observed spatial variation.

In the face of the large demand for answers on global issues, such as "*what is the impact of climate change on food production?*", "*which soils best protect the main soil functions, now and in the future?*", "*how can we increase the global carbon storage of soils?*" and "*how can we reach land degradation neutrality?*", soil maps should not only focus on producing spatially explicit information about the current soil status, but they should also deliver relevant information about soil processes. Therefore, I think that SEM is a relevant tool to generate soil information that is useful for land management at a local and regional scale, where people's decisions affect the soil system and where we need to know how the system will react to those decisions. Thus, with SEM we not only describe soil spatial distribution, but we can explicitly link soil spatial distribution to the mechanistic process knowledge behind it. There is a need and there is a way. I foresee a bright future for SEM in digital soil mapping.

References

Adhikari, K., Kheir, R. B., Greve, M. B., Bøcher, P. K., Malone, B. P., Minasny, B., McBratney, A. B., Greve, M. H., 2013. High-resolution 3-D mapping of soil texture in Denmark. Soil Science Society of America Journal 77, 860–876.
URL http://dx.doi.org/10.2136/sssaj2012.0275

Adhikari, K., Minasny, B., Greve, M. B., Greve, M. H., 2014. Constructing a soil class map of Denmark based on the FAO legend using digital techniques. Geoderma 214-215, 101–113.
URL http://www.sciencedirect.com/science/article/pii/S0016706113003510

Anderson, J. C., Gerbing, D. W., 1988. Structural equation modeling in practice: A review and recommended two-step approach. Psychological Bulletin 103 (3), 411–423.

Arbuckle, J., 1997. Amos users' guide, version 3.6.

Arhonditsis, G. B., Stow, C. A., Steinberg, L. J., Kenney, M. A., Lathrop, R. C., McBride, S. J., Reckhow, K. H., 2006. Exploring ecological patterns with structural equation modeling and Bayesian analysis. Ecological Modelling 192 (3), 385–409.
URL http://www.sciencedirect.com/science/article/pii/S0304380005003765

Bagozzi, R. P., Yi, Y., 1988. On the evaluation of structural equation models. Journal of the Academy of Marketing Science 16 (1), 74–94.
URL https://doi.org/10.1177/009207038801600107

Bentler, P. M., 1990. Comparative fit indexes in structural models. Psychological Bulletin 107 (2), 238–246.

Blake, G. R., Steinhardt, G. C., Pombal, X. P., Muñoz, J. C. N., Cortizas, A. M., Arnold, R. W., Schaetzl, R. J., 2008. Pedoturbation. Springer Netherlands, Dordrecht, pp. 516–522.
URL https://doi.org/10.1007/978-1-4020-3995-9_417

Boast, C. W., 1973. Modeling the movement of chemicals in soils by water. Soil Science 115 (3), 224–230.
URL http://journals.lww.com/soilsci/Fulltext/1973/03000/MODELING_THE_MOVEMENT_OF_CHEMICALS_IN_SOILS_BY.8.aspx

Bockheim, J. G., Gennadiyev, A. N., 2000. The role of soil-forming processes in the definition of taxa in Soil Taxonomy and the World Soil Reference Base. Geoderma 95 (1), 53–72.
URL http://www.sciencedirect.com/science/article/pii/S001670619900083X

Bollen, K. A., 1989. Structural equations with latent variables. Wiley, New York, USA.

Brady, N. C., Weil, R. R., 2014. The nature and properties of soils, fourteenth Edition. Pearson, Harlow.

Brahim, N., Blavet, D., Gallali, T., Bernoux, M., Mar. 2011. Application of structural equation modeling for assessing relationships between organic carbon and soil properties in semiarid Mediterranean region. International Journal of Environmental Science & Technology 8 (2), 305–320.
URL https://doi.org/10.1007/BF03326218

Brevik, E. C., Calzolari, C., Miller, B. A., Pereira, P., Kabala, C., Baumgarten, A., Jordán, A., 2016. Soil mapping, classification, and pedologic modeling: History and future directions. Geoderma 264, 256–274.
URL http://www.sciencedirect.com/science/article/pii/S0016706115001718

Brevik, E. C., Hartemink, A. E., 2010. Early soil knowledge and the birth and development of soil science. CATENA 83 (1), 23–33.
URL http://www.sciencedirect.com/science/article/pii/S0341816210001013

Brungard, C. W., Boettinger, J. L., Duniway, M. C., Wills, S. A., Edwards, T. C., 2015. Machine learning for predicting soil classes in three semi-arid landscapes. Geoderma 239-240, 68–83.
URL http://www.sciencedirect.com/science/article/pii/S0016706114003516

Brus, D. J., Kempen, B., Heuvelink, G. B. M., 2011. Sampling for validation of digital soil maps. European Journal

of Soil Science 62 (3), 394–407.
URL http://dx.doi.org/10.1111/j.1365-2389.2011.01364.x

Bui, E. N., 2004. Soil survey as a knowledge system. Geoderma 120 (1–2), 17–26.
URL http://www.sciencedirect.com/science/article/pii/S0016706103002738

Bui, E. N., Moran, C. J., 2003. A strategy to fill gaps in soil survey over large spatial extents: An example from the Murray-Darling basin of Australia. Geoderma 111 (1–2), 21–44.
URL http://www.sciencedirect.com/science/article/pii/S0016706102002380

Buol, S. W., Southard, R. J., Graham, R. C., McDaniel, P. A., 2011. Mollisols: Grassland Soils of Steppes and Prairies. Wiley-Blackwell, pp. 331–347.
URL http://dx.doi.org/10.1002/9780470960622.ch15

Cabrini, S., Calcaterra, C., 2008. Los sistemas de producción en la cuenca del Arroyo Pergamino. Resultados de una encuesta a productores.
URL https://inta.gob.ar/documentos/

Cambule, A. H., Rossiter, D. G., Stoorvogel, J. J., 2013. A methodology for digital soil mapping in poorly-accessible areas. Geoderma 192, 341–353.
URL http://www.sciencedirect.com/science/article/pii/S0016706112003151

Clark, J. S., Gelfand, A. E., 2006. A future for models and data in environmental science. Trends in Ecology & Evolution 21 (7), 375–380.
URL http://www.sciencedirect.com/science/article/pii/S0169534706001030

Conrad, O., 2001. SAGA-GIS module library documentation: Module cluster analysis for grids.
URL http://www.saga-gis.org/saga_tool_doc/2.1.3/imagery_classification_1.html

Conrad, O., Bechtel, B., Bock, M., Dietrich, H., Fischer, E., Gerlitz, L., Wehberg, J., Wichmann, V., Böhner, J., 2015. System for automated geoscientific analyses (SAGA) v. 2.1.4. Geoscientific Model Development 8 (7), 1991–2007.

Cook, S. E., Corner, R. J., Groves, P. R., Grealish, G. J., Feb. 1996. Use of airborne gamma radiometric data for soil mapping. Soil Research 34 (1), 183–194.
URL https://doi.org/10.1071/SR9960183

Cruzate, G. A., 2001. Caracterización y cartografía de los materiales parentales de los suelos del centro de la Región Pampeana mediante el procesamiento geoestadístico de parámetros químicos y físicos. In: Convenio Facultad de Agronomía UBA-Instituto Nacional de Tecnología Agropecuaria. INTA.

D'Amico, M. E., Freppaz, M., Zanini, E., Bonifacio, E., Jun. 2017. Primary vegetation succession and the serpentine syndrome: The proglacial area of the Verra Grande glacier, North-Western Italian Alps. Plant and Soil 415 (1), 283–298.
URL https://doi.org/10.1007/s11104-016-3165-x

de Gruijter, J., Brus, D., Bierkens, M., Knotters, M., 2006. Sampling for Natural Resource Monitoring. Springer-Verlag, Berlin, Heidelberg.
URL https://doi.org/10.1007/3-540-33161-1

Delgado-Baquerizo, M., Maestre, F. T., Gallardo, A., Bowker, M. A., Wallenstein, M. D., Quero, J. L., Ochoa, V., Gozalo, B., Garcia-Gomez, M., Soliveres, S., Garcia-Palacios, P., Berdugo, M., Valencia, E., Escolar, C., Arredondo, T., Barraza-Zepeda, C., Bran, D., Carreira, J. A., Chaieb, M., Conceicao, A. A., Derak, M., Eldridge, D. J., Escudero, A., Espinosa, C. I., Gaitan, J., Gatica, M. G., Gomez-Gonzalez, S., Guzman, E., Gutierrez, J. R., Florentino, A., Hepper, E., Hernandez, R. M., Huber-Sannwald, E., Jankju, M., Liu, J., Mau, R. L., Miriti, M., Monerris, J., Naseri, K., Noumi, Z., Polo, V., Prina, A., Pucheta, E., Ramirez, E., Ramirez-Collantes, D. A., Romao, R., Tighe, M., Torres, D., Torres-Diaz, C., Ungar, E. D., Val, J., Wamiti, W., Wang, D., Zaady, E., Oct. 2013. Decoupling of soil nutrient cycles as a function of aridity in global drylands. Nature 502 (7473), 672–676.
URL http://dx.doi.org/10.1038/nature12670

Dobos, E., Carré, F., Hengl, T., Reuter, H. I., Tóth, G., 2006. Digital soil mapping as a support to production of functional maps. Office for Official Publications of the European Communites, Luxemburg 22123.

Dokuchaev, V. V., 1883. The Russian Chernozem report to the free economic society. Imperial University of Saint Petersburg, St. Petersburg, Russia.

Dormann, C. F., 2007. Effects of incorporating spatial autocorrelation into the analysis of species distribution data. Global Ecology and Biogeography 16 (2), 129–138.
URL http://dx.doi.org/10.1111/j.1466-8238.2006.00279.x

Duchaufour, P., 1998. Handbook of pedology (translated from French by VAK Sharma). Soil Sciece, 164.

Duncan, O. D., 1966. Path analysis: Sociological examples. American Journal of Sociology 72 (1), 1–16.
URL https://doi.org/10.1086/224256

Duncan, T. E., Duncan, S. C., 2004. An introduction to latent growth curve modeling. Behavior Therapy 35 (2),

333–363.
URL http://www.sciencedirect.com/science/article/pii/S000578940480042X

Durán, A., Morrás, H., Studdert, G., Liu, X., Sep. 2011. Distribution, properties, land use and management of Mollisols in South America. Chinese Geographical Science 21 (5), 511.
URL https://doi.org/10.1007/s11769-011-0491-z

Duval, J. S., Carson, J. M., Holman, P. B., Darnley, A. G., 2005. Terrestrial radioactivity and gamma-ray exposure in the United States and Canada. Open-File Report 2005-1413, U.S. Geological Survey, available online only.
URL https://pubs.usgs.gov/of/2005/1413/

Edwards, M. C., Wirth, R. J., Houts, C. R., Xi, N., 2012. Categorical Data in the Structural Equation Modeling Framework. Guilford Press New York, NY, 72 Spring Street, New York, NY 10012, Ch. 12, pp. 495–511.

Epskamp, S., 2015. semPlot: Unified visualizations of structural equation models. Structural Equation Modeling: A Multidisciplinary Journal 22 (3), 474–483.
URL http://dx.doi.org/10.1080/10705511.2014.937847

FAO, 2017. GSOC map. Cookbook Manual. FAO, 1st Edition.
URL http://www.fao.org/3/a-bs901e.pdf

Farr, T. G., Kobrick, M., 2000. Shuttle radar topography mission produces a wealth of data. Eos, Transactions American Geophysical Union 81 (48), 583–585.
URL http://dx.doi.org/10.1029/EO081i048p00583

Finke, P. A., 2012a. Modeling the genesis of luvisols as a function of topographic position in loess parent material. Quaternary International 265, 3–17.
URL http://www.sciencedirect.com/science/article/pii/S1040618211005891

Finke, P. A., 2012b. On digital soil assessment with models and the Pedometrics agenda. Geoderma 171, 3–15.
URL http://www.sciencedirect.com/science/article/pii/S0016706111000024

Finke, P. A., Hutson, J. L., 2008. Modelling soil genesis in calcareous loess. Geoderma 145 (3), 462–479.
URL http://www.sciencedirect.com/science/article/pii/S0016706108000268

Fitzpatrick, B. R., Lamb, D. W., Mengersen, K., Sep. 2016. Ultrahigh dimensional variable selection for interpolation of point referenced spatial data: A digital soil mapping case study. PLOS ONE 11 (9), 1–19.
URL https://doi.org/10.1371/journal.pone.0162489

Fox, J., Weisberg, S., 2011. An R companion to applied regression, 2nd Edition. Sage Publications, Thousand Oaks, CA.

Glinka, K. D., 1927. Dokuchaiev's ideas in the development of pedology and cognate sciences. No. I in Academy of Sciences of the Union of the Soviet Socialist Republics. Russian pedological investigations. The Academy, Leningrad.

Gonzalez Bonorino, F., 1966. Soil clay mineralogy of the Pampa plains, Argentina. Journal of Sedimentary Research 36 (4), 1026–1035.
URL http://jsedres.geoscienceworld.org/content/36/4/1026

Goodin, D. G., 1995. Climate and weather atlas of Kansas: an introduction. Kansas geological survey, University of Kansas, Laurence, Kansas.

Goovaerts, P., 1992. Factorial kriging analysis: A useful tool for exploring the structure of multivariate spatial soil information. Journal of Soil Science 43 (4), 597–619.
URL http://dx.doi.org/10.1111/j.1365-2389.1992.tb00163.x

Grace, J. B., 2006. Structural equation modeling and natural systems. Cambridge University Press, Cambridge, UK.
URL http://pubs.er.usgs.gov/publication/70185580

Grace, J. B., Adler, P. B., Stanley Harpole, W., Borer, E. T., Seabloom, E. W., 2014. Causal networks clarify productivity–richness interrelations, bivariate plots do not. Functional Ecology 28 (4), 787–798.
URL http://dx.doi.org/10.1111/1365-2435.12269

Grace, J. B., Anderson, M. T., Olff, H., Scheiner, S. M., 2010. On the specification of structural equation models for ecological systems. Ecological Monographs 80 (1), 67–87.
URL http://dx.doi.org/10.1890/09-0464.1

Grace, J. B., Bollen, K. A., Jun. 2008. Representing general theoretical concepts in structural equation models: The role of composite variables. Environmental and Ecological Statistics 15 (2), 191–213.
URL https://doi.org/10.1007/s10651-007-0047-7

Grace, J. B., Keeley, J. E., 2006. A structural equation model analysis of postfire plant diversity in California shrublands. Ecological Applications 16 (2), 503–514.
URL http://dx.doi.org/10.1890/1051-0761(2006)016[0503:ASEMAO]2.0.CO;2

Grace, J. B., Schoolmaster, D. R., Guntenspergen, G. R., Little, A. M., Mitchell, B. R., Miller, K. M., Schweiger, E. W.,

2012. Guidelines for a graph-theoretic implementation of structural equation modeling. Ecosphere 3 (8), 1–44.
URL http://dx.doi.org/10.1890/ES12-00048.1

Grinand, C., Arrouays, D., Laroche, B., Martin, M. P., 2008. Extrapolating regional soil landscapes from an existing soil map: Sampling intensity, validation procedures, and integration of spatial context. Geoderma 143 (1), 180–190.
URL http://www.sciencedirect.com/science/article/pii/S0016706107003199

Grunwald, S., 2009. Multi-criteria characterization of recent digital soil mapping and modeling approaches. Geoderma 152 (3), 195–207.
URL http://www.sciencedirect.com/science/article/pii/S0016706109001827

Guisan, A., Thuiller, W., 2005. Predicting species distribution: Offering more than simple habitat models. Ecology Letters 8 (9), 993–1009.
URL http://dx.doi.org/10.1111/j.1461-0248.2005.00792.x

Gunal, H., Ransom, M. D., Mar. 2006. Genesis and micromorphology of loess-derived soils from central Kansas. Catena 65 (3), 222–236.
URL http://www.sciencedirect.com/science/article/pii/S034181620500202X

Haavelmo, T., 1943. The statistical implications of a system of simultaneous equations. Econometrica 11 (1), 1–12.
URL http://www.jstor.org/stable/1905714

Hägglund, G., 2001. Structural equation modeling: Present and future. Scientific Software International, Ch. Milestones in the history of factor analysis, pp. 11–38.

Hartemink, A. E., Hempel, J., Lagacherie, P., McBratney, A. B., McKenzie, N. J., MacMillan, R. A., Minasny, B., Montanarella, L., de Mendonça Santos, M. L., Sanchez, P., Walsh, M., Zhang, G.-L., 2010. GlobalSoilMap.net – A New Digital Soil Map of the World. Springer Netherlands, Dordrecht, pp. 423–428.
URL https://doi.org/10.1007/978-90-481-8863-5_33

Hartemink, A. E., Minasny, B., 2014. Towards digital soil morphometrics. Geoderma 230-231, 305–317.
URL http://www.sciencedirect.com/science/article/pii/S0016706114001177

Hengl, T., Heuvelink, G. B. M., Stein, A., 2004. A generic framework for spatial prediction of soil variables based on regression-kriging. Geoderma 120 (1), 75–93.
URL http://www.sciencedirect.com/science/article/pii/S0016706103002787

Hengl, T., Mendes de Jesus, J., Heuvelink, G. B. M., Ruiperez Gonzalez, M., Kilibarda, M., Blagoti'c, A., Shangguan, W., Wright, M. N., Geng, X., Bauer-Marschallinger, B., Guevara, M. A., Vargas, R., MacMillan, R. A., Batjes, N. H., Leenaars, J. G. B., Ribeiro, E., Wheeler, I., Mantel, S., Kempen, B., Feb. 2017. Soilgrids250m: Global gridded soil information based on machine learning. PLOS ONE 12 (2), 1–40.
URL https://doi.org/10.1371/journal.pone.0169748

Heuvelink, G. B. M., 2006. Incorporating process knowledge in spatial interpolation of environmental variables. In: Proceedings of Accuracy 2006; 7th international symposium on spatial accuracy assessment in natural resources and environmental sciences. pp. 32–47.

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 (2), 201–215.
URL http://www.sciencedirect.com/science/article/pii/S2352009416300219

Heuvelink, G. B. M., Webster, R., 2001. Modelling soil variation: Past, present, and future. Geoderma 100, 269–301.
URL http://www.sciencedirect.com/science/article/pii/S0016706101000258

Hewitt, A. E., 1993. Predictive modelling in soil survey. Soils and Fertilizers 56 (3), 305–314.

Hogg, R. V., Craig, A. T., 1995. Introduction to mathematical statistics, 5th Edition. Upper Saddle River, New Jersey: Prentice Hall.

Hu, L., Bentler, P. M., 1999. Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. Structural Equation Modeling: A Multidisciplinary Journal 6 (1), 1–55.
URL http://dx.doi.org/10.1080/10705519909540118

Huggett, R. J., 1998. Soil chronosequences, soil development, and soil evolution: A critical review. CATENA 32 (3), 155–172.

Iacobucci, D., 2009. Everything you always wanted to know about SEM (structural equations modeling) but were afraid to ask. Journal of Consumer Psychology 19, 673–680.
URL https://ssrn.com/abstract=2693262

Iacobucci, D., 2010. Structural equations modeling: Fit indices, sample size, and advanced topics. Journal of Consumer Psychology 20, 90–98.
URL https://ssrn.com/abstract=2693263

Imbellone, P. A., Giménez, J. E., Mormeneo, M. L., 2014. Suelos calcáreos del litoral noreste de la Provincia de

Buenos Aires, Argentina. Suelos con acumulaciones calcáreas y yesíferas de Argentina, 57–93.

Imbellone, P. A., Giménez, J. E., Panigatti, J. L., 2010. Suelos de la Región Pampeana: Procesos de formación. No. 287. INTA, Buenos Aires.

INTA, 1964. Plan Mapa de Suelos de la Región Pampeana. Resolución CD No. 218/1964. Gaceta INTA.

INTA, 2015. SISINTA: Sistema de información de suelos del INTA.
URL http://sisinta.inta.gob.ar

IUSS Working Group World Reference Base, 2006. World reference base for soil resources 2006. A framework for international classification, correlation and communication.

Jenny, H., 1941. Factors of soil formation: A system of quantitative pedology. Courier Corporation.

Jönsson, P., Eklundh, L., 2004. TIMESAT– program for analyzing time-series of satellite sensor data. Computers & Geosciences 30 (8), 833–845.
URL http://www.sciencedirect.com/science/article/pii/S0098300404000974

Jöreskog, K. G., Sörbom, D., 1981. LISREL V: Analysis of linear structural relationships by maximum likelihood and least squares methods. University of Uppsala, Department of Statistics.

Jöreskog, K. G., Sörbom, D., 1982. Recent developments in structural equation modeling. Journal of Marketing Research 19 (4), 404–416.
URL http://www.jstor.org/stable/3151714

Jöreskog, K. G., Yang, F., 1996. Nonlinear structural equation models: The Kenny-Judd model with interaction effects. Taylor & Francis Group, 270 Madison Ave, New York NY 10016, Ch. 3, pp. 57–88.

Kempen, B., 2011. Updating soil information with digital soil mapping. Ph.D. thesis, Wageningen University, Wageningen, Netherlands.
URL http://edepot.wur.nl/187198

Kempen, B., Brus, D. J., de Vries, F., 2015. Operationalizing digital soil mapping for nationwide updating of the 1:50,000 soil map of the Netherlands. Geoderma 241-242, 313–329.
URL http://www.sciencedirect.com/science/article/pii/S0016706114004285

Kempen, B., Brus, D. J., Heuvelink, G. B. M., Stoorvogel, J. J., 2009. Updating the 1:50,000 Dutch soil map using legacy soil data: A multinomial logistic regression approach. Geoderma 151 (3), 311–326.
URL http://www.sciencedirect.com/science/article/pii/S0016706109001475

Kempen, B., Brus, D. J., Stoorvogel, J. J., Heuvelink, G. B. M., de Vries, F., 2012. Efficiency comparison of conventional and digital soil mapping for updating soil maps. Soil Science Society of America Journal 76, 2097–2115.
URL http://dx.doi.org/10.2136/sssaj2011.0424

Kirkby, M. J., 1977. Soil development models as a component of slope models. Earth Surface Processes 2 (2-3), 203–230.
URL http://dx.doi.org/10.1002/esp.3290020212

Kline, R. B., 2015. Principles and practice of structural equation modeling. Guilford Publications, Inc., 370 Seventh Avenue, Suite 1200, New York, NY 10001.

Knotters, M., Brus, D. J., Voshaar, J. H. O., 1995. A comparison of kriging, co-kriging and kriging combined with regression for spatial interpolation of horizon depth with censored observations. Geoderma 67 (3), 227–246.
URL http://www.sciencedirect.com/science/article/pii/001670619500011C

Koopmans, T., 1945. Statistical estimation of simultaneous economic relations. Journal of the American Statistical Association 40 (232), 448–466.
URL http://amstat.tandfonline.com/doi/abs/10.1080/01621459.1945.10500746

Krause, P., Boyle, D. P., Bäse, F., Dec. 2005. Comparison of different efficiency criteria for hydrological model assessment. Advances in Geosciences 5, 89–97.
URL https://hal.archives-ouvertes.fr/hal-00296842

Kröhling, D. M., Iriondo, M. H., Dec. 2003. El loess de La Pampa norte en el bloque de San Guillermo. Revista de la Asociación Argentina de Sedimentología 10, 137–150.
URL http://www.scielo.org.ar/scielo.php?script=sci_arttext&pid=S1853-63602003000200004&nrm=iso

Kuenzer, C., Dech, S., Wagner, W., 2015. Remote Sensing Time Series Revealing Land Surface Dynamics: Status Quo and the Pathway Ahead. Springer International Publishing, Cham, Ch. 1, pp. 1–24.
URL https://doi.org/10.1007/978-3-319-15967-6_1

Kuhn, M., Johnson, K., 2013. Applied predictive modeling. Vol. 810. Springer.
URL https://link.springer.com/book/10.1007%2F978-1-4614-6849-3

Kuo, Y.-F., Wu, C.-M., Deng, W.-J., 2009. The relationships among service quality, perceived value, customer satisfaction, and post-purchase intention in mobile value-added services. Computers in Human Behavior 25 (4), 887–896.

URL http://www.sciencedirect.com/science/article/pii/S0747563209000363

Lacoste, M., Minasny, B., McBratney, A., Michot, D., Viaud, V., Walter, C., 2014. High resolution 3D mapping of soil organic carbon in a heterogeneous agricultural landscape. Geoderma 213, 296–311.
URL http://www.sciencedirect.com/science/article/pii/S0016706113002358

Lagacherie, P., Legros, J. P., Burrough, P. A., 1995. A soil survey procedure using the knowledge of soil pattern established on a previously mapped reference area. Geoderma 65 (3), 283–301.
URL http://www.sciencedirect.com/science/article/pii/001670619400040H

Lamb, E. G., Mengersen, K. L., Stewart, K. J., Attanayake, U., Siciliano, S. D., 2014. Spatially explicit structural equation modeling. Ecology 95 (9), 2434–2442.

Landrum, C., Castrignanò, A., Mueller, T., Zourarakis, D., Zhu, J., Benedetto, D. D., 2015. An approach for delineating homogeneous within-field zones using proximal sensing and multivariate geostatistics. Agricultural Water Management 147, 144–153.
URL http://www.sciencedirect.com/science/article/pii/S0378377414002121

Lark, R. M., 2000. A comparison of some robust estimators of the variogram for use in soil survey. European Journal of Soil Science 51 (1), 137–157.
URL http://dx.doi.org/10.1046/j.1365-2389.2000.00280.x

Laurencena, P., Varela, L. B., Kruse, E., Rojo, A., Deluchi, M., 2002. Características de las variaciones freáticas en un área del Noreste de la Provincia de Buenos Aires. In: Aguas Subterráneas y Desarrollo Humano (Mar del Plata, 2002).

Lawley, D. N., Jan. 1940. VI.–The estimation of factor loadings by the method of maximum likelihood. Proceedings of the Royal Society of Edinburgh 60, 64–82.
URL http://journals.cambridge.org/article_S037016460002006X

Lee, S.-Y., Song, X.-Y., 2004. Evaluation of the Bayesian and maximum likelihood approaches in analyzing structural equation models with small sample sizes. Multivariate Behavioral Research 39 (4), 653–686.
URL http://dx.doi.org/10.1207/s15327906mbr3904_4

Legendre, P., 1993. Spatial autocorrelation: Trouble or new paradigm? Ecology 74 (6), 1659–1673.
URL http://dx.doi.org/10.2307/1939924

Legendre, P., Fortin, M. J., 1989. Spatial pattern and ecological analysis. Vegetatio 80 (2), 107–138.

Lichstein, J. W., Simons, T. R., Shriner, S. A., Franzreb, K. E., 2002. Spatial autocorrelation and autoregressive models in ecology. Ecological Monographs 72 (3), 445–463.
URL http://dx.doi.org/10.1890/0012-9615(2002)072[0445:SAAAMI]2.0.CO;2

Mallavan, B. P., Minasny, B., McBratney, A. B., 2010. Homosoil, a Methodology for Quantitative Extrapolation of Soil Information Across the Globe. Springer Netherlands, Dordrecht, pp. 137–150.
URL https://doi.org/10.1007/978-90-481-8863-5_12

Malone, B. P., Jha, S. K., Minasny, B., McBratney, A. B., 2016. Comparing regression-based digital soil mapping and multiple-point geostatistics for the spatial extrapolation of soil data. Geoderma 262, 243–253.
URL http://www.sciencedirect.com/science/article/pii/S0016706115300665

Malone, B. P., Styc, Q., Minasny, B., McBratney, A. B., 2017. Digital soil mapping of soil carbon at the farm scale: A spatial downscaling approach in consideration of measured and uncertain data. Geoderma 290, 91–99.
URL http://www.sciencedirect.com/science/article/pii/S0016706116309922

Marsh, H. W., 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 (3), 320–341.
URL http://dx.doi.org/10.1207/s15328007sem1103_2

Matteson, K. C., Grace, J. B., Minor, E. S., 2013. Direct and indirect effects of land use on floral resources and flower-visiting insects across an urban landscape. Oikos 122 (5), 682–694.
URL http://dx.doi.org/10.1111/j.1600-0706.2012.20229.x

McBratney, A. B., Field, D. J., Jarrett, L. E., 2017. General Concepts of Valuing and Caring for Soil. Springer International Publishing, Cham, pp. 101–108.
URL https://doi.org/10.1007/978-3-319-43394-3_9

McBratney, A. B., Santos, M. L. M., Minasny, B., 2003. On digital soil mapping. Geoderma 117 (1), 3–52.
URL http://www.sciencedirect.com/science/article/pii/S0016706103002234

McKenzie, N. J., Gallant, J. C., 2006. Chapter 24 Digital soil mapping with improved environmental predictors and models of pedogenesis. In: Lagacherie, P., McBratney, A., Voltz, M. (Eds.), Digital Soil Mapping. Vol. 31 of Developments in Soil Science. Elsevier, pp. 327–349.
URL http://www.sciencedirect.com/science/article/pii/S0166248106310240

Minasny, B., Malone, B. P., McBratney, A. B., Angers, D. A., Arrouays, D., Chambers, A., Chaplot, V., Chen, Z.-S., Cheng, K., Das, B. S., Field, D. J., Gimona, A., Hedley, C. B., Hong, S. Y., Mandal, B., Marchant, B. P., Martin, M., McConkey, B. G., Mulder, V. L., O'Rourke, S., de Forges, A. C. R., Odeh, I., Padarian, J., Paustian, K., Pan, G., Poggio, L., Savin, I., Stolbovoy, V., Stockmann, U., Sulaeman, Y., Tsui, C.-C., Vågen, T.-G., van Wesemael, B., Winowiecki, L., 2017. Soil carbon 4 per mille. Geoderma 292, 59–86.
URL http://www.sciencedirect.com/science/article/pii/S0016706117300095

Minasny, B., McBratney, A. B., 2016. Digital soil mapping: A brief history and some lessons. Geoderma 264, 301–311.
URL http://www.sciencedirect.com/science/article/pii/S0016706115300276

Moore, I. D., Burch, G. J., 1986. Modelling erosion and deposition: Topographic effects. Transactions of the ASAE 29 (6), 1624–1630.
URL https://elibrary.asabe.org/abstract.asp?aid=30363

Moore, I. D., Grayson, R. B., Ladson, A. R., 1991. Digital terrain modelling: A review of hydrological, geomorphological, and biological applications. Hydrological Processes 5 (1), 3–30.
URL http://dx.doi.org/10.1002/hyp.3360050103

Moran, P. A. P., 1950. Notes on continuous stochastic phenomena. Biometrika 37 (1/2), 17–23.
URL http://www.jstor.org/stable/2332142

Morrás, H. J. M., 1999. Geochemical differentiation of Quaternary sediments from the Pampean Region based on soil phosphorus contents as detected in the early 20th century. Quaternary International 62 (1), 57–67.
URL http://www.sciencedirect.com/science/article/pii/S1040618299000233

Morrás, H. J. M., Altinier, M., Castiglioni, M., Grasticini, C., Ciari, G., Cruzate, G. A., 2002. Composición mineralógica y heterogeneidad espacial de sedimentos loéssicos superficiales en la Pampa Ondulada. In: Actas XVIII Reunión Argentina de la Ciencia del Suelo. AACS.

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. Springer International Publishing, Cham, pp. 193–209.
URL https://doi.org/10.1007/978-3-319-19159-1_11

Mulder, V. L., Lacoste, M., de Forges, A. C. R., Martin, M. P., Arrouays, D., 2016. National versus global modelling the 3D distribution of soil organic carbon in mainland France. Geoderma 263, 16–34.
URL http://www.sciencedirect.com/science/article/pii/S001670611530063X

Muthén, L. K., Muthén, B. O., 1998. Mplus: The comprehensive modeling program for applied researchers; user's guide;[Version 1.0].

O'Geen, A. T., 2012. Soil water dynamics. Nature Education Knowledge 3 (6), 12.
URL https://www.nature.com/scitable/knowledge/library/soil-water-dynamics-59718900

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.
URL http://www.sciencedirect.com/science/article/pii/S1040618214001074

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 (230), 119–130.
URL http://www.sciencedirect.com/science/article/pii/S0016706114001736

Padarian, J., Minasny, B., McBratney, A. B., 2017. Chile and the Chilean soil grid: A contribution to GlobalSoilMap. Geoderma Regional 9, 17–28.
URL http://www.sciencedirect.com/science/article/pii/S2352009416301390

Pearl, J., 1988. Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

Pearl, J., 1995. Causal diagrams for empirical research. Biometrika 82 (4), 669–688.
URL +http://dx.doi.org/10.1093/biomet/82.4.669

Pearl, J., 1998. Graphs, causality, and structural equation models. Sociological Methods & Research 27 (2), 226–284.
URL http://dx.doi.org/10.1177/0049124198027002004

Pearl, J., 2009. Causal inference in statistics: An overview. Statistics Surveys 3, 96–146.
URL https://doi.org/10.1214/09-SS057

Phillips, J. D., 1993. Progressive and regressive pedogenesis and complex soil evolution. Quaternary Research 40 (2), 169–176.
URL http://www.sciencedirect.com/science/article/pii/S0033589483710690

Poggio, L., Gimona, A., 2017. Assimilation of optical and radar remote sensing data in 3D mapping of soil properties over large areas. Science of The Total Environment 579, 1094–1110.
URL http://www.sciencedirect.com/science/article/pii/S0048969716325177

Poggio, L., Gimona, A., Brewer, M. J., Nov. 2013. Regional scale mapping of soil properties and their uncertainty with a large number of satellite-derived covariates. Geoderma 209-210, 1–14.
URL http://linkinghub.elsevier.com/retrieve/pii/S0016706113001997

Quirós, R., 2005. La ecología de las lagunas de las Pampas. Investigación y Ciencia 1, 1–13.
URL https://pdfs.semanticscholar.org/0bbf/d8b1517d70d672412f23c265486ea095631f.pdf

R Core Team, 2017. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.
URL https://www.R-project.org/

Rosseel, Y., 2012. lavaan: An R package for structural equation modeling. Journal of Statistical Software 48 (2), 1–36.
URL http://www.jstatsoft.org/v48/i02/

Rosseel, Y., Jul. 2013. The lavaan tutorial. Department of Data Analysis – Ghent University, Belgium.
URL http://dornsifecms.usc.edu/assets/sites/210/docs/GC3/lavaan_tutorial.pdf

Rosseel, Y., Oct. 2017. lavaan: latent variable analysis. Accessed on 23 Oct. 2017.
URL http://lavaan.ugent.be/

Rossiter, D. G., Liu, J., Carlisle, S., Zhu, X. A., 2015. Can citizen science assist digital soil mapping? Geoderma 259-260, 71–80.
URL http://www.sciencedirect.com/science/article/pii/S0016706115001548

Runge, E. C. A., 1973. Soil development sequences and energy models. Soil Science 115 (3), 183–193.
URL http://journals.lww.com/soilsci/Fulltext/1973/03000/SOIL_DEVELOPMENT_SEQUENCES_AND_ENERGY_MODELS_.3.aspx

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, 214–227.
URL http://www.sciencedirect.com/science/article/pii/S001670611400456X

Schaetzl, R. J., Anderson, S., 2005. Soils: Genesis and geomorphology. Cambridge University Press, Cambridge; New York.

Schaetzl, R. J., Frederick, W. E., Tornes, L., 1996. Secondary carbonates in three fine and fine-loamy Alfisols in Michigan 60, 1862–1870.
URL http://dx.doi.org/10.2136/sssaj1996.03615995006000060035x

Schoeneberger, P. J., Wysocki, D. A., Benham, E. C., Broderson, W. D., Soil Survey Staff, 2012. Field Book for Describing and Sampling Soils, Version 3.0. Natural Resources Conservation Service, National Soil Survey Center, Lincoln, NE.

Schoorl, J. M., Veldkamp, A., Bouma, J., 2002. Modeling water and soil redistribution in a dynamic landscape context. Soil Science Society of America Journal 66 (5), 1610–1619.
URL http://dx.doi.org/10.2136/sssaj2002.1610

Schumacker, R. E., Jun. 1998. Interaction and Nonlinear Effects in Structural Equation Modeling. Routledge.
URL http://dx.doi.org/10.4324/9781315092614

Scoppa, C., Dec. 1975. La mineralogía de los suelos de la llanura Pampeana en la interpretación de su génesis y distribución. In: VII Reunión Argentina de la Ciencia del Suelo. AACS.

Shipley, B., 2000. Cause and correlation in biology : A user's guide to path analysis, structural equations and causal inference, first edition Edition. Cambridge [etc.], GB: Cambridge University Press.

Silva, S. H. G., de Menezes, M. D., Owens, P. R., Curi, N., 2016. Retrieving pedologist's mental model from existing soil map and comparing data mining tools for refining a larger area map under similar environmental conditions in Southeastern Brazil. Geoderma 267, 65–77.
URL http://www.sciencedirect.com/science/article/pii/S0016706115301762

Simonson, R. W., 1959. Outline of a generalized theory of soil genesis. Soil Science Society of America Journal 23 (2), 152–156.

Simonson, R. W., 1978. A multiple-process model of soil genesis. Quaternary Soils, 1–25.

Sobel, M. E., 1982. Asymptotic confidence intervals for indirect effects in structural equation models. Sociological Methodology 13, 290–312.
URL http://www.jstor.org/stable/270723

Soil Science Division Staff, 2017. Soil Survey Manual. USDA Handbook 18. Government Printing Office, Washington, D.C.
URL https://www.nrcs.usda.gov/wps/portal/nrcs/detailfull/soils/ref/?cid=nrcs142p2_054262

Soil Survey Staff, 1951. Soil Survey Manual. USDA, Printing Office. Washington DC, handbook no. 18 Edition.

Soil Survey Staff, 1960. Soil Classification, a Comprehensive System—7th Approximation. USDA, Printing Office.

Washington DC.

Soil Survey Staff, 2014. Keys to Soil Taxonomy. USDA-Natural Resources Conservation Service, Washington, DC, twelfth Edition.
URL https://www.nrcs.usda.gov

Soil Survey Staff, 2016. Soil Survey Geographic (SSURGO) Database. Database, Natural Resources Conservation Service, United States Department of Agriculture.
URL https://www.nrcs.usda.gov/wps/portal/nrcs/detail/soils/survey/?cid=nrcs142p2_053627

Sommer, M., Gerke, H. H., Deumlich, D., 2008. Modelling soil landscape genesis — A “time split” approach for hummocky agricultural landscapes. Geoderma 145 (3), 480–493.
URL http://www.sciencedirect.com/science/article/pii/S0016706108000153

Sonneveld, M. P. W., Heuvelink, G. B. M., Moolenaar, S., 2014. Application of a visual soil examination and evaluation technique at site and farm level. Soil Use and Management 30 (2), 263–271.
URL http://dx.doi.org/10.1111/sum.12117

Spearman, C., 1904. "general intelligence," objectively determined and measured. The American Journal of Psychology 15 (2), 201–292.
URL http://www.jstor.org/stable/1412107

Steeb, W., 2012. Kronecker Product. World Scientific Publishing, University of Johannesburg, South Africa, pp. 152–173.
URL http://www.worldscientific.com/doi/abs/10.1142/9789812772930_0009

Stockmann, U., Minasny, B., McBratney, A. B., 2014. How fast does soil grow? Geoderma 216, 48–61.
URL http://www.sciencedirect.com/science/article/pii/S0016706113003601

Stoorvogel, J. J., Kempen, B., Heuvelink, G. B. M., de Bruin, S., 2009. Implementation and evaluation of existing knowledge for digital soil mapping in Senegal. Geoderma 149 (1), 161–170.
URL http://www.sciencedirect.com/science/article/pii/S0016706108003534

Taalab, K., Corstanje, R., Zawadzka, J., Mayr, T., Whelan, M. J., Hannam, J., Creamer, R., 2015. On the application of Bayesian networks in digital soil mapping. Geoderma 259, 134–148.
URL http://www.sciencedirect.com/science/article/pii/S0016706115001688

Taboada, M. A., Damiano, F., Lavado, R. S., 2009. Inundaciones en la Región Pampeana. Consecuencias sobre los suelos. Alteraciones de la fertilidad de los suelos: el halomorfismo, la acidez, el hidromorfismo y las inundaciones, 103–127.

Temme, A. J. A. M., Schoorl, J. M., Veldkamp, A., 2006. Algorithm for dealing with depressions in dynamic landscape evolution models. Computers & Geosciences 32 (4), 452–461.
URL http://www.sciencedirect.com/science/article/pii/S0098300405001810

Temme, A. J. A. M., Vanwalleghem, T., 2016. LORICA —A new model for linking landscape and soil profile evolution: Development and sensitivity analysis. Computers & Geosciences 90, 131–143.
URL http://www.sciencedirect.com/science/article/pii/S0098300415300297

Temme, A. J. A. M., Veldkamp, A., 2009. Multi-process Late Quaternary landscape evolution modelling reveals lags in climate response over small spatial scales. Earth Surface Processes and Landforms 34 (4), 573–589.
URL http://dx.doi.org/10.1002/esp.1758

Vanwalleghem, T., Poesen, J., McBratney, A. B., Deckers, J., 2010. Spatial variability of soil horizon depth in natural loess-derived soils. Geoderma 157 (1), 37–45.
URL http://www.sciencedirect.com/science/article/pii/S001670611000090X

Vanwalleghem, T., Stockmann, U., Minasny, B., McBratney, A. B., 2013. A quantitative model for integrating landscape evolution and soil formation. Journal of Geophysical Research: Earth Surface 118 (2), 331–347.
URL http://dx.doi.org/10.1029/2011JF002296

Verbesselt, J., Hyndman, R., Newnham, G., Culvenor, D., 2010. Detecting trend and seasonal changes in satellite image time series. Remote Sensing of Environment 114 (1), 106–115.
URL http://www.sciencedirect.com/science/article/pii/S003442570900265X

Viglizzo, E. F., Pordomingo, A. J., Castro, M. G., Lértora, F. A., Bernardos, J. N., 2004. Scale-dependent controls on ecological functions in agroecosystems of Argentina. Agriculture, Ecosystems & Environment 101 (1), 39–51.
URL http://www.sciencedirect.com/science/article/pii/S0167880903002299

Viscarra Rossel, R. A., Chen, C., Grundy, M. J., Searle, R., Clifford, D., Campbell, P. H., 2015. The Australian three-dimensional soil grid: Australians contribution to the *GlobalSoilMap* project. Soil Research 53 (8), 845–864.
URL https://doi.org/10.1071/SR14366

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 proper-

ties. Geoderma 131 (1), 59–75.
URL http://www.sciencedirect.com/science/article/pii/S0016706105000728

Wackernagel, H., 1995. Multivariate geostatistics: An introduction with applications, 1st Edition. Springer, Berlin, Heidelberg, Germany.

Wall, M. M., 2012. Spatial structural equation modeling with an application to U.S. behavioral risk factor surveillance survey data. In: Hoyle, R. H. (Ed.), Handbook of Structural Equation Modeling. The Guilford Press, 72 Spring Street, New York, NY 10012, pp. 636–649.

Webster, R., Oliver, M. A., 2007. Geostatistics for environmental scientists. John Wiley & Sons.
URL http://onlinelibrary.wiley.com/book/10.1002/9780470517277

WEPAL, 2015. Certificate of analysis. International soil-analitical exchange. Reference Material ISE Sample 900.
URL http://www.wepal.nl/website/download_files/consensus/ISE/ISE900.pdf

Were, K., Bui, D. T., Dick, Ø. B., Singh, B. R., 2015. A comparative assessment of support vector regression, artificial neural networks, and random forests for predicting and mapping soil organic carbon stocks across an Afromontane landscape. Ecological Indicators 52, 394–403.
URL http://www.sciencedirect.com/science/article/pii/S1470160X14006049

Wilkinson, M. T., Richards, P. J., Humphreys, G. S., 2009. Breaking ground: Pedological, geological, and ecological implications of soil bioturbation. Earth-Science Reviews 97 (1), 257–272.
URL http://www.sciencedirect.com/science/article/pii/S0012825209001470

Wright, S., 1921. Correlation and causation. J. Agric. Res. 20, 557–585.

Xu, S., An, X., Qiao, X., Zhu, L., Li, L., 2013. Multi-output least-squares support vector regression machines. Pattern Recognition Letters 34 (9), 1078–1084.
URL http://www.sciencedirect.com/science/article/pii/S0167865513000196

Yang, R.-M., Zhang, G.-L., Liu, F., Lu, Y.-Y., Yang, F., Yang, F., Yang, M., Zhao, Y.-G., Li, D.-C., 2016. Comparison of boosted regression tree and random forest models for mapping topsoil organic carbon concentration in an Alpine ecosystem. Ecological Indicators 60, 870–878.
URL http://www.sciencedirect.com/science/article/pii/S1470160X15004689

Zárate, M. A., 2003. Loess of southern South America. Quaternary Science Reviews 22 (18), 1987–2006.
URL http://www.sciencedirect.com/science/article/pii/S0277379103001653

Zhu, A. X., Hudson, B., Burt, J., Lubich, K., Simonson, D., 2001. Soil mapping using GIS, expert knowledge, and fuzzy logic. Soil Science Society of America Journal 65 (5), 1463–1472.
URL http://dx.doi.org/10.2136/sssaj2001.6551463x

Summary

Climate change and land degradation are of increasing societal and governmental concern. For this reason, several international programs have been initiated in the last decade, such as the *4 per 1000* initiative and the Sustainable Development Goals of United Nations. The soil science community is actively working under different national and international organizations to provide regional and global soil information to support these programmes. Digital Soil Mapping (DSM), a relatively new methodology to create soil maps based on (geo)statistical methods, has became operational during the last fifteen years and has now been adopted by several organizations. It is defined as computer-assisted production of digital maps of soil type and soil properties, by use of mathematical and statistical models that combine information from soil observations with information contained in correlated environmental variables.

Most studies in DSM spatially predict soil properties or classes from either new or legacy laboratory data and spatially exhaustive environmental covariates (GIS layers of biophysical land surface properties), typically using empirical statistical methods. These methods have shown to result in accurate maps at different scales, but do not provide knowledge about the interrelationships between the soil properties and the functioning of the soil and soil-landscape system. We not only need to properly describe or map soil spatial variation, but also to understand soil behaviour. This is needed to answer questions such as: which are the dominant soil processes in a certain region? How will the soil react under increased productivity pressure? How vulnerable is the soil to erosion or pollution? How much organic carbon can we store in the soil at a given location?

Mechanistic soil-landscape models do include process-knowledge but cannot be applied easily for soil mapping because of their high complexity and large uncertainty. A solution could be to use structural equation modelling (SEM), which is a hybrid approach that combines elements of empirical and mechanistic models. SEM can model continuous soil properties while taking soil property interrelationships into account. In SEM, we first create a conceptual model, similar to the mental model of

soil surveyors, which is converted into a graphical model, that represents the system interrelationships. This is the mechanistic side of SEM. The empirical side takes place after we translated the graphical model into a mathematical model, which is calibrated with observational data to estimate the model coefficients. Next, the calibrated model can be used to predict target variables, such as soil properties. These characteristics of SEM indicate that it could be a very useful technique to bridge the gap between empirical and mechanistic approaches for DSM. Thus, the objective of this thesis is to extend DSM with soil process information through the **development**, **calibration**, **application** and **validation** of a structural equation model.

After a general introduction to this thesis in Chapter 1, Chapter 2 describes how SEM can be implemented for DSM. In this chapter I argue that current DSM methods have limitations. For instance, it is difficult to predict a large number of soil properties simultaneously, while preserving the relationships between them. Furthermore, current widely applied prediction models use pedological knowledge in a very crude way only. To address these issues in DSM, I investigated the use of SEM. I introduced SEM theory and presented a case study in a 23 000-km^2 region in the Argentinian Pampas, where I applied SEM to map seven key soil properties for the A horizon. I started with identifying the main soil forming processes in the study area and determined for each process the main soil properties affected. Based on this analysis I defined a conceptual soil-landscape model, which was subsequently converted to a SEM graphical model. The graphical model was translated to a mathematical model in the statistical software R using the latent variable analysis (`lavaan`) package. The prediction accuracy was poor, which was caused by a large measurement error in combination with a homogeneous study area. Nevertheless, the outcomes demonstrated that SEM can be used to explicitly include pedological knowledge in prediction of soil properties and modelling of their interrelationships.

In Chapter 3 I explored the capabilities of SEM for three-dimensional soil mapping. Since many soil processes operate within the soil profile, SEM might be suitable for simultaneous prediction of soil properties for multiple soil layers. The objectives of this chapter therefore were to i) apply SEM to multi-layer and multivariate soil mapping, ii) test SEM functionality for improving model performance by using model suggestions and iii) assess whether SEM reproduced the soil property covariation better than a multiple linear regression (MLR) model. I applied SEM to model and predict the lateral and vertical distribution of the cation exchange capacity (CEC), organic carbon (OC) and clay content for the A, B and C horizons for the study area of Chapter 2. I found that SEM reproduces the interrelationships between soil properties more accurately than MLR and that the model suggestions helped to improve the fit of the model. I concluded that SEM can be used to predict several soil properties for multiple layers simultaneously while retaining soil property interrelationships.

Given that SEM is a hybrid between mechanistic and empirical models, I hypothesised in Chapter 4 that SEM should have better extrapolation properties than a purely empirical model. I therefore investigated the extrapolation of a SE model from one region to another region with similar conditions. Empirical models have been used in DSM for extrapolation with varying success. The objective of this chapter was to investigate the extrapolation capability of SEM by testing and comparing six different model settings for extrapolation. I applied the structural equation model from Chapter 3 to a similar soil-landscape in the Great Plains of the United States to predict clay, OC and CEC for the same three major horizons A, B, and C.

I evaluated the performance of the SE mathematical model extrapolation, as well as the graphical model extrapolation (without coefficients). I defined four SE models that differed in the degree to which these were tailored to the US case study. I started with extrapolating the Argentinian SE mathematical model and ended with an extrapolated graphical model that was fitted and adapted on the basis of model suggestions using the US data. I also evaluated two more models using MLR to assess if SEM was better than purely empirical models.

The Argentinian SE mathematical model gave the worst results in the US while the extrapolated graphical model that was fitted with US data and adapted based on model suggestions performed best. Interestingly, I found that a SE graphical model only based on the conceptual model performed better that a more precise Argentinian SE graphical model. For this reason, I concluded that the adaptation of the conceptual model to a specific study area can improve local prediction but harm the potential predictive power for extrapolation. The prediction performance of the SE mathematical model was not substantially better than MLR. However, system relationships that were well supported by pedological knowledge showed consistent and equal behaviour in both study areas. Contrary, differences in the sign and strength of the relationships between covariates and soil properties of both areas reduced the performance of the mathematical model extrapolation. Thus, I concluded that knowledge-based links between system variables are more effective than data-driven links for model extrapolation. In addition, a deeper understanding of indicators of soil-forming factors could strengthen conceptual models for DSM.

Spatial correlation is an important feature in spatial analysis, especially in DSM. So far, current implementations of SEM do not take spatial correlation in data into account. The objective of Chapter 5 therefore was to extend SEM by accounting for residual spatial correlation using a geostatistical approach. I presented how the SE model definition and parameter estimation can be generalised to the spatially correlated case. The spatial SE model was applied to map the same soil properties of

Chapter 4 in the Great Plains study area. The SE model residuals showed substantial spatial correlation, which suggests that including spatial correlation yields more accurate predictions. I also compared spatial SEM with standard SEM in terms of SEM model coefficients. There was significant differences, although none of the coefficients changed sign. Presence of residual spatial correlation suggests that some of the causal factors that explain soil variation were not captured by the set of covariates. In such case it is worthwhile to search for and include additional covariates leaving only unstructured residual noise, but as long as this is not achieved, it pays off to include residual spatial correlation in soil mapping using SEM.

Chapter 6 presents a synthesis of the main findings of this thesis and reflects on the possible role of SEM in DSM. It discusses some limitations of SEM, such as the assumed linearity of relationships (while in reality relationships between soil properties and environmental covariates are often non-linear) and its static nature, as well as several challenges. One of the main challenges I encountered in the case studies was the lack of good-quality proxies (covariates) to adequately represent the soil-forming processes. Since we cannot include hundred of covariates to predict soil properties in SEM, as done in machine-learning techniques, we need to develop proper covariates to achieve a good SE model. Nevertheless, I concluded that SEM is an appropriate technique to include pedological knowledge in DSM, and could potentially fill an important niche. Developing a SE model requires thorough work to translate a conceptual model to a graphical model. By doing so, one is able to test pedological hypotheses and learn from the data, because the model suggestions bring new considerations and research. SEM brings added value to DSM, because the graphical model unites pedological knowledge and statistical modelling in a single framework. What is more, every model coefficient can be analysed in terms of its sign and magnitude with respect to the other coefficients, and in terms of its pedological meaning. For that reason, I think that SEM can be a method to do *conscious* DSM since it helps one to become more aware of the processes behind the system interrelationships.

This thesis only introduced SEM for DSM. There are several features and possibilities of SEM that I did not research. The use of latent variables to represent conceptual variables, such as soil fertility, soil quality, etc., should be the next step in its adaptation for DSM purposes. Implementing categorical variables and non-linear relations will also bring more flexibility to the model, and can provide solutions for areas with different environmental conditions. In the face of the large demand for answers on global issues that are often addressed at national or regional scale, I am convinced that SEM is a suitable framework to meet these demands. Also, I am confident that SEM fills a vacant niche in DSM, since it does not compete with machine learning techniques and mechanistic modelling approaches, and provides a framework for *conscious* DSM.

Resumen

El cambio climático y la degradación del suelo son problemas cada vez más preocupantes para la sociedad y los gobiernos. Por este motivo se han iniciado numerosos programas internacionales en la última década, como por ejemplo la iniciativa *4 per 1000* y los Objetivos de Desarrollo Sustentable de las Naciones Unidas. La comunidad científica del suelo está trabajando activamente bajo diferentes organizaciones nacionales e internacionales para proporcionar información de este recurso a escala regional y mundial. Para ello, durante los últimos quince años se ha implementado el Mapeo Digital de Suelos (DSM, por sus siglas en Inglés *Digital Soil Mapping*), una metodología relativamente nueva para crear mapas de suelos, y que ahora está siendo adoptada por varias organizaciones, entre ellas la FAO. El DSM se define como la producción asistida por computadora de mapas digitales de tipos y propiedades de suelo, mediante el uso de modelos matemáticos y estadísticos que combinan información de observaciones del suelo con información contenida en variables ambientales correlacionadas.

La mayoría de los estudios que emplean DSM predicen las propiedades o clases de suelos espacialmente a partir de datos de laboratorio (nuevos o legados de previos relevamientos) y de covariables ambientales (capas GIS de propiedades biofísicas de la superficie terrestre). Generalmente se utilizan métodos estadísticos empíricos, que han demostrado producir mapas precisos a diferentes escalas. Sin embargo, estos métodos no generan conocimientos sobre las relaciones entre propiedades del suelo y sobre la relación suelo-paisaje. Los mapas de suelos no sólo necesitan describir apropiadamente la variación espacial del mismo, sino también proveer información de su comportamiento. Este tipo de información es necesaria para responder a preguntas tales como: ¿cuáles son los procesos dominantes del suelo en una determinada región? ¿Cómo reaccionará el suelo bajo una mayor presión de productividad? ¿Qué tan vulnerable es el suelo a la erosión o la contaminación? ¿Cuánto carbono orgánico podemos almacenar en el suelo en un lugar determinado?

Otros modelos utilizados en el estudio de las variaciones espaciales de los suelos son los modelos mecanísticos (también llamados modelos físicos) que incluyen pro-

cesos físico-químicos que ocurren en el sistema suelo-paisaje. Sin embargo, no se pueden aplicar fácilmente al mapeo de suelos debido a su alta complejidad y gran incertidumbre. Una solución intermedia entre los modelos empíricos y los modelos mecanísticos sería utilizar modelos de ecuaciones estructurales (SEM, por sus siglas en Inglés *Structural Equation Modelling*). Mediante SEM se puede modelar propiedades continuas del suelo teniendo en cuenta las interrelaciones existentes entre ellas. En SEM, primero se crea un modelo conceptual, similar al modelo mental que desarrollaban los clásicos reconocedores de suelos. Este modelo conceptual es posteriormente transformado en un modelo gráfico, el cual representa las interrelaciones del sistema. Luego, el modelo gráfico se traduce a un modelo matemático que se calibra con datos medidos de las propiedades involucradas dando como resultado los coeficientes del modelo. A continuación, el modelo calibrado se puede utilizar para predecir las propiedades de suelo. Estas características de SEM lo convierten en una herramienta útil para cerrar la brecha entre los enfoques empíricos y mecánisticos en DSM. Por lo tanto, el objetivo de esta tesis es expandir las metodologías de DSM mediante la incorporación de conocimiento pedológico a través del **desarrollo**, **calibración**, **aplicación** y **validación** de un modelo de ecuaciones estructurales.

Después de una introducción general a esta tesis en el Capítulo 1, el Capítulo 2 describe cómo se puede implementar SEM en DSM. Aquí argumento que los métodos actuales de DSM presentan limitaciones. Por ejemplo, es difícil predecir un gran número de propiedades del suelo simultáneamente preservando las relaciones entre ellas. Además, la mayoría de los modelos actuales en DSM utilizan el conocimiento pedológico de una manera muy simple. Para abordar estos problemas, se propone investigar el uso de SEM. Para ello, se hace una introducción a los aspectos teóricos de SEM y se presenta un caso de estudio en una área correspondiente a la Pampa Ondulada argentina donde se aplicó este método para mapear siete propiedades de suelo. Se identificaron los principales procesos de formación de suelo en el área de estudio y se determinó para cada proceso sus principales propiedades afectadas. De acuerdo a este análisis, se definió un modelo conceptual de suelo-paisaje, que posteriormente se convirtió a un modelo gráfico. El modelo gráfico se tradujo a un modelo matemático en el software estadístico R utilizando el paquete `lavaan`. La precisión de predicción fue pobre a causa de un alto error de medición en combinación con un área de estudio homogénea. Sin embargo, el caso de estudio demostró que el conocimiento pedológico se puede incluir explícitamente en SEM, permitiendo además predecir simultáneamente varias propiedades de suelo y modelar sus interrelaciones.

En el Capítulo 3 se exploraron las bondades de SEM para el mapeo del suelo en sus tres dimensiones. Dado que muchos procesos operan dentro del perfil de suelo, se planteó que SEM podría ser adecuado para la predicción simultánea de sus propiedades en sus diferentes capas. Por lo tanto, los objetivos de este capítulo fueron i) aplicar SEM al mapeo simultaneo de propiedades de suelos en múltiples capas, ii) evaluar las sugerencias de mejoras en la predicción que puede proveer SEM y iii) evaluar si SEM puede reproducir la covariación de la propiedad del suelo de mejor manera que un modelo de regresión lineal múltiple (MLR). Para ello se utilizó el área de estudio del Capítulo 2. Se modelaron y predijeron la capacidad de intercambio catiónico (CIC), el carbono orgánico (CO) y el contenido de arcilla de los horizontes A, B y C. De esta manera se comprobó que SEM reproduce las interrelaciones entre las propiedades del suelo con mayor precisión que MLR. Las sugerencias provistas por SEM ayudaron a mejorar la precisión de predicción del modelo. Como conclusión de este capítulo se pudo afirmar que SEM es una metodología viable para predecir varias propiedades del suelo en múltiples capas al mismo tiempo, conservando las interrelaciones de sus propiedades.

Dado que SEM puede ser considerado un modelo híbrido entre los modelos mecanísticos y los empíricos, en el Capítulo 4 se planteó como hipótesis que SEM debería ser más preciso que un modelo puramente empírico cuando se aplica en extrapolación (o predicción fuera del área utilizada para calibrar el modelo). En DSM, la extrapolación de modelos empíricos ha producido resultados variados que usualmente han desalentado su aplicación. En este capítulo, por lo tanto, investigué si SEM podía ser utilizado para extrapolar un modelo de una región a otra donde las condiciones de suelos eran similares. Para ello se diseñaron y compararon seis modelos. Las áreas incluidas en este estudio fueron la Pampa Ondulada en Argentina (misma área que los capítulos previos) y una región del Great Plains en Estados Unidos que abarca parte de los estados de Kansas y Nebraska. Al igual que en el Capítulo 3, para el estudio se tomaron las propiedades CIC, CO y contenido de arcilla de los horizontes A, B y C.

Para evaluar las aptitudes de extrapolación de los modelos, no sólo se tomó en cuenta el modelo calibrado en el área de estudio de origen (es decir, las ecuaciones con sus parámetros calibrados), sino también el modelo gráfico sin calibrar, es decir, su estructura. SEM se aplicó en cuatro modelos, diferenciados entre si por su adaptabilidad a las condiciones del área de Estados Unidos. Así, el primero de ellos fue calibrado en Argentina y la predicción fue hecha en Estados Unidos; en el segundo modelo se utilizó sólo la estructura del modelo de Argentina (modelo gráfico sin coeficientes) y se calibró con datos de Estados Unidos; el tercer modelo consistió en extrapolar el modelo gráfico original (Fig. 4.6), el cuál se calibró con datos de Estados Unidos; en el cuarto modelo se adaptó el modelo gráfico de Argentina a las condiciones de Estados Unidos (tomando las sugerencias que provee SEM) y se calibró en esa misma área. Todos los modelos fueron siempre aplicados sobre el área de estudio de Estados Unidos. Para los otros dos modelos se utilizó MLR a fin de evaluar si SEM presentaba mejores resultados que un modelo completamente empírico.

Las principales conclusiones de este capítulo fueron (1) que a pesar que SEM dio resultados mejores que MLR, las diferencias fueron pequeñas y la precisión del modelo muy baja; (2) que la adaptación del modelo gráfico a un área de estudio específica, a través de las sugerencias que brinda el software, mejoran la precisión de predicción en esa área, pero va en detrimento de su capacidad de extrapolación. Dicha conclusión surge de observar que el (tercer) modelo que representaba el modelo conceptual original tuvo mejor comportamiento que el (segundo) modelo que representaba el modelo mejor adaptado a las condiciones de Argentina. Se concluyó también (3) que las interrelaciones entre variables que están bien fundamentadas son más efectivas que aquellas producidas enteramente por modelos empíricos, dado que SEM fue siempre más preciso que MLR. En este capítulo, además, pude observar que las relaciones de las propiedades de suelos entre el horizonte A y el B en la Pampa Ondulada fueron más débiles que las de Estado Unidos, reforzando la hipótesis de otros autores que el horizonte A de la Pampa Ondulada podría tratarse de un material de diferente ciclo pedológico que el que se encuentra en los horizontes B y C.

La correlación espacial es una muy importante característica en DSM. Hasta aquí, la implementación de SEM al mapeo de suelos no toma en cuenta que las propiedades de suelos pueden estar espacialmente correlacionadas, lo que implica que las observaciones no son independientes. Por lo tanto, el objetivo del Capítulo 5 fue tomar en cuenta la correlación espacial en los residuales del modelo mediante la incorporación del modelo geoestadístico en SEM. Este capítulo muestra cómo se pueden estimar los parámetros del modelo cuando las variables están espacialmente correlacionadas. Para ejemplificar su uso se utilizó la misma área de estudio del Capítulo 4 y el modelo mejor adaptado a las condiciones de Estados Unidos. El modelo presentaba alta correlación espacial en sus residuales, lo cual sugería que incluyendo esta característica en el modelado podría incrementar significativamente su precisión. A fin de evaluar las virtudes del modelo espacial se consideró (1) la precisión de la predicción y (2) en que grado se vieron afectados los coeficientes del modelo. Los resultados mostraron que el modelo espacial fue significativamente más preciso, y que si bien los parámetros del modelo cambiaron, ninguno de los coeficientes cambió su signo. En conclusión, vale la pena incluir dicha característica en el modelo a fin de lograr una predicción más precisa. Sin embargo, hay dos aspectos a tener en cuenta: uno es que la presencia de correlación espacial en los residuos sugiere que los factores causales de la distribución de suelos no han sido captados por las variables predictoras; y el otro que el método estadístico no permite dilucidar cuáles son esos factores de variación de los suelos. Por ello, sería recomendable buscar e incluir aquellas covariables que sí representan los verdaderos factores de variación de los suelos.

El Capítulo 6 presenta una síntesis de los principales descubrimientos de esta tesis y una reflexión del posible rol de SEM en DSM. Aquí se discuten algunas limitaciones de SEM, tales como el supuesto de linealidad de las relaciones entre variables, así como también su característica estática. También, en este capítulo se discuten algunos desafíos que encontré al implementar SEM. Uno de ellos fue la falta de covariables de calidad que representen de mejor manera los procesos formadores de suelos. Dado que en SEM no es posible (ni tampoco tendría sentido) incluir cientos de covariables, así como puede hacerse con técnicas como *machine-learning*, aquí se requiere identificar los procesos formadores de suelos y desarrollar las covariables que los representan. Así mismo, mi conclusión es que SEM es una técnica apropiada para incluir el conocimiento pedológico en DSM, que puede cubrir un nicho que hasta ahora se mantiene vacío. El modelado con SEM requiere un trabajo minucioso para transformar conceptos a modelos gráficos. A través de este proceso, uno es capaz de probar estadísticamente hipótesis pedológicas y desarrollar conocimiento a partir de las observaciones, ya que el modelo puede sugerir conexiones que el investigador no había considerado. Además, SEM aporta un elemento de análisis nuevo en DSM a través del modelo gráfico, el cuál resume las características estadísticas del modelo y los conceptos pedológicos en un solo gráfico. Lo que es más, dado que en SEM cada coeficiente debe ser analizado en detalle para corroborar su significado, considero que SEM es un método para el mapeo digital *conciente* de las propiedades de suelos.

En esta tesis tan sólo he introducido como implementar SEM en DSM. Hay muchas características y desarrollos en la literatura que no han sido incluidas. Quizás, una de las más relevantes es el uso de variables latentes para representar propiedades conceptuales, como podría ser la fertilidad o la salud del suelo, variables que no pueden ser medidas directamente sino a través de indicadores. También, el uso de variables categóricas y de relaciones no lineales está desarrollado en la literatura de SEM, lo cual sería de gran utilidad en DSM. Frente a la gran demanda de respuestas acerca de problemas ambientales globales y/o regionales, estoy convencido de que SEM es una herramienta apta para suplir dicha demanda. Y para finalizar, me gustaría destacar que SEM puede cubrir un nicho vacío en DSM, ya que no compite con las técnicas de *machine-learning* ni tampoco con los modelos mecanísticos, y requiere que tanto especialistas en pedología y en estadística trabajen juntos para producir un mapeo digital de suelos conciente.

Agradecimientos
Acknowledgements

Este doctorado significó emprender un largo viaje lleno de desafíos e incertidumbres, el cual no podría haber realizado si no fuera gracias a la ayuda de muchas personas. Por ello en este apartado quiero agradecer a quienes, directa o indirectamente, han hecho posible esta experiencia.

Me gustaría comenzar agradeciendo a mi mamá y a mi papá, Susana y Hugo, que siempre me incentivaron a hacer lo que me gusta con pasión y dedicación, que me dieron todo cuanto pudieron para apoyarme en este viaje, y que soportaron la distancia y el tiempo que nos separó.

Gracias a mi gran compañera de aventuras, Indira, y a nuestr*x*s tres hij*x*s, Sofía, Bruno y Amelie, a quienes considero el motor de este viaje, y por lo tanto autores morales de este trabajo. Sin dudas que este ha sido un logro de nosotros cinco.

Comenzar el doctorado no hubiera sido posible sin el apoyo de mis directores del INTA, Miguel A. Taboada y Pablo Mercuri, a quienes les estoy muy agradecido por creer en mi. Agradezco también a Héctor J.M. Morras por sus sabios consejos y por incentivarme a lanzarme a esta experiencia.

Pienso que la formación del personal en una institución no sólo se da gracias a las políticas definidas por las autoridades, sino también gracias al esfuerzo mancomunado de sus trabajadores. Por eso considero que este logro personal en realidad es gracias a las labores de cada uno de mis compañeros del Instituto de Suelos. Quiero agradecer especialmente a Darío, quien me ha *bancado* y me ha dado soporte durante todo este período. También agradecer el trabajo de Alicia, Horacio, Beatriz y Santiago por la carga de datos que fueron utilizados en esta tesis, y a Lucas y Emiliano por facilitarme todos los medios a su disposición para realizar el trabajo de campo.

Durante el trabajo de campo que realicé entre agosto y octubre de 2014, tuve la suerte de compartir jornadas con muchas personas que dieron su apoyo desinteresado. Quiero agradecer a quienes me han acompañado en las labores de toma de muestras, como ser a mi *viejo*, (gracias Pa por estar siempre al pié del cañón), mi hermano Hernán, mis cuñados Fernando y Enrique, mis amigos y colegas Federico y Darío, a Ignacio Tuja y Matías Rau (estudiantes), y a Héctor (profesor y coautor de uno de los capítulos). Quiero también expresar mi total gratitud a todos los productores que con mucha amabilidad me permitieron ingresar a sus lotes, así como a todos aquellos que me ayudaron a localizar los dueños o encargados de los sitios que yo debía visitar, como ser Lito Suse de Mariano Alfonso, Anahí Cortese, Jorge Marchen y Fernando Moreno de Agrícola Conesa, a Mario García de La Violeta, a los vecinos de Los Indios, entre muchos otros que no me es posible recordar sus nombres.

I am deeply thankful to Budiman Minasny for having given me the opportunity to take my first steps in Digital Soil Mapping in Sydney, when my English was terribly bad. This experience was essential to start my PhD in this subject. Also, I am thankful for his friendship and kindness which I will never forget.

In 2012 I was at a conference in Rome where I met Hannes Reuter. Thanks to him and his advice I started this PhD in the Netherlands. He put me in contact with Gerard. He was part of the project committee of this thesis and participated in the project proposal development. Since he left ISRIC, it was difficult for me to continue the collaboration. However, I fully appreciate all his comments and advice.

When we (my family and I) arrived in the Netherlands, two people were extremely valuable for our adaptation to the new environment. Gerard found a house for us to live in and paid the rent in advance without having met me before. On my first day in Wageningen, he also carried me at the back rack of his bike to buy some food, and brought us many goods, such as TV, DVD movies, DVD player, game boards, a toolbox, etc. The other person was Margaret, who is actually called Margaretha (Meijer) but simplified her name for us. She was our first neighbour. After welcoming us with a cake, she introduced us to the Dutch culture in the most practical way, just living it. She introduced us to all our neighbours, shops, activities, school, sportive centres, etc. That was simply impressive. I thank them both infinitely.

Also, I would like to thank our cordial neighbours from Kamperfoelielaan and from Meidoornplantsoen who integrated us in their community and shared their friendship. My family and I have enjoyed all the fantastic parties, meetings and nice events organised together. Special thanks to our great friends Ruth and Ronald and the Canela band (Nico, Tere, Alejandro, Saskia, Marcel and Kees) who helped me to develop my creativity through art and music. When I first came to the Netherlands, I expected to complete a PhD, but I had never thought I would be able to paint and to play in a band, which were some of my best "*cables a tierra*".

I want to express my sincere gratitude to all my colleagues (PhDs, guest researchers, students and staff members) of ISRIC and the Soil Geography and Landscape group. Thanks for all the great moments that we lived together, such as barbecues, outings, excursions, lectures, lunches, presentations, cakes, beers, we-days, running, soil augering championships, worm charming championships, PE&RC days and weekends, Christmas lunch and dinners, courses, discussion groups, fruitful meetings, conferences, as well as many other pleasant moments. Thank you all for your advice, support and patience with me. Special words of thanks to Eloi, Rafael and Thomas for your friendship, to my roommates Chantal, Luc, Maricke, Kasia, Rocky, Alessandro, Francis, Shangguan Wei, and Eliana, and to the rest of my PhD colleagues, Simona, Alex, Marijn, Jasper, Selçuk, and Cindy. I will miss your friendship. Thanks Jelmer! for making music together. Also I wish to thank Ad for so many enthusiastic talks and plans, and for helping me to understand the lab procedures, to Ype and Arnaud for your help, especially in my first steps of my research, and to Jan and Stephan for teaching me all about monoliths. David, I wish to extend my gratitude to you not only for providing me valuable comments and support for my research, but also for sharing your music and your great experiences around the world. Thanks also to Titia, Luc, Alex and Cathelijne for helping me with my last step, the thesis defence. I might have forgotten some people. My apologies!

I would like to acknowledge the support of Lennart and Claudius from the PE&RC graduate school, as well as María Pereira, my buddy when I arrived in the Netherlands. You guys have been brave! Helping so many PhD candidates, especially those foreign people like me, that not only need help with the common issues of a PhD, but also need to overcome the cultural differences and meet the expectations of our Dutch colleagues and supervisors. I have enjoyed very much the activities organised by you.

Special thanks to my teacher of English, Enid Tomkinson, who helped me to improve my English. I found learning English a huge challenge. With you Enid, I did not only learn, but also enjoyed your lessons and your friendship.

Finally, I would like to remark how proud I am for having been supervised by Gerard and Bas, to whom I will be grateful forever. Thanks Gerard and Bas! for your patience, for always devoting a lot of time to our meetings, for understanding my limitations, for being sincere and direct to me, for supporting me in many aspects of my life, and for being available every time I needed. I will be greatly indebted with you.

About the author

Marcos E. Angelini was born in Luján, Buenos Aires, Argentina on the 7th May, 1975. In 1993 he began a five-year programme on Agricultural Engineering at the National University of Luján (UNLu). In 1996 he joined the remote sensing laboratory in the same university, where he developed skills on processing remote sensing data. Marcos served as an assistant teacher in several undergraduate and graduate courses. He achieved his degree of Agronomical Engineer in September 2002. In September 2003 he obtained a fellowship from the CONAE (National Commission for Space Activities) for an internship at the Consiglio Nazionale delle Ricerche in Italy for "Early Detection Of Fire Through Remote Sensing." In 2004, he obtained a permanent position at the Soils Institute of the National Institute of Agricultural Technology (INTA). Between 2004 and 2007 Marcos taught in the "Specialization In Remote Sensing And Geographic Information Systems" programme at UNLu. He also was hired by UNEP (United Nations Environment Programme) as a trainer on remote sensing and to estimate cropping area in several provinces of Argentina using remote sensing data. In 2004 he participated in a Workshop on Global Land Cover of FAO in Ecuador. From 2006 to 2009, Marcos participated in different INTA projects related to soil mapping, carbon sequestration, soil quality indicators, and land cover and land use change. In September 2009, Marcos obtained a grant from INTA, for a short stay (3 months) as a guest researcher at The University of Sydney, to work on Digital Soil Mapping. After his return, Marcos started to coordinate a project called "Digital Soil Mapping: developing, testing and validating new soil survey methodologies." Then, in 2010, he participated in the "1st Latinoamerican Workshop of GlobalSoilMap.Net and Atlas of Soil" in EMBRAPA Solos, Brazil. In the same year, Marcos obtained a grant from the Australian Leadership Awards (ALA) Fellowships (AusAID) for participation in the program "Adapting Agriculture And Natural Resource Management to Climate Variability And Change; What

Skills Will We Need?” with the supervision of Prof. Willem Vervoort and Dr Budiman Minasny. In August 2011 he became the chair of the Cartography Department of the Soil Institute. Marcos also participated in one of the Global Soil Partnership meetings in Italy in 2012, participated in the SISLAC program in the Latin American soil congress (Mar del Plata, Argentina) and in the SISLAC-I meeting in Cali, Colombia. In March 2013, Marcos started this PhD at Wageningen University, the Netherlands, while he maintained his post at the Soil Institute (INTA). By the end of his PhD, Marcos will return to Argentina to continue working at INTA.

PE&RC Training and Education Statement

With the training and education activities listed below the PhD candidate has complied with the requirements set by the C.T. de Wit Graduate School for Production Ecology and Resource Conservation (PE&RC) which comprises of a minimum total of 32 ECTS (= 22 weeks of activities)

Review of literature (4.5 ECTS)

- Including pedological knowledge in statistical modelling of soil properties

Writing of project proposal (4 ECTS)

- Integrating soil forming process information with digital soil mapping in Phaeozems of Argentinian Pampas (2014)

Post-graduate courses (3.4 ECTS)

- Hands-on global soil information facilities; ISRIC (2013)
- Structural Equation Modelling (SEM); WIAS/PE&RC (2015)
- The Science of conservation: managing biodiversity in a changing world; WASS/PE&RC/SENSE and University of Venda (2016)

Invited review of (unpublished) journal manuscript (4 ECTS)

- Revista Argentina de la Ciencia del Suelo: soil mapping (2016)
- Revista Argentina de la Ciencia del Suelo: land evaluation (2016)
- Congreso Internacional SELPER: remote sensing; two manuscripts (2016)

Deficiency, refresh, brush-up courses (13.5 ECTS)

- Geology and landscapes of the world; WUR (2012-2013)
- Basic statistics; PE&RC/SENSE (2013)
- Environmental data collection and analysis; WUR (2013-2014)

Competence strengthening / skills courses (2.7 ECTS)

- The essentials of scientific writing & presenting; Wageningen in'to language (2014)
- Project and time management; WGS (2014)

PE&RC Annual meetings, seminars and the PE&RC weekend (1.8 ECTS)

- PE&RC Day (2013)
- PE&RC Weekend first year (2013)
- PE&RC Weekend middle term (2015)

Discussion groups / local seminars / other scientific meetings (5.6 ECTS)

- R User discussion group and organizer of R-User group (2014-2017)
- Seminars at ISRIC (2014-2017)
- Modelling and Statistics Network (MSN) (2016)

International symposia, workshops and conferences (7.5 ECTS)

- Wageningen Soil Conference; oral presentation (2015)
- Pedometrics; oral presentation (2015)
- Pedometrics; oral presentation (2017)

The research described in this thesis was financially supported by the National Institute of Agricultural Technology approved under resolution 744/12 (*Instituto Nacional de Tecnología Agropecuaria*, INTA, Argentina).

Propositions

1. Unlike machine learning, structural equation modelling encourages *conscious* digital soil mapping.
   (this thesis)

2. It is not possible to achieve accurate soil maps in homogeneous areas.
   (this thesis)

3. The shopping behaviour of most environmental scientists does not agree with what they proclaim at work.

4. Patents and copyright regulations obstruct scientific progress.

5. Rather than pursue comfort, we must seek challenges.

6. When you live abroad, the larger the cultural differences between your country and the residence country, the more you learn about your own culture.

Propositions belonging to the thesis entitled:

Structural equation modelling for digital soil mapping

Marcos Esteban Angelini
Wageningen, 6 March 2018